UNIVERSITÉ DU QUÉBEC

MÉMOIRE PRÉSENTÉ À

L'UNIVERSITÉ DU QUÉBEC À CHICOUTIMI

COMME EXIGENCE PARTIELLE

DE LA MAÎTRISE EN INFORMATIQUE

PAR

FANGJUN SHEN

A COMPUTATIONAL MODEL OF MUTUAL TRUST BETWEEN THE USER AND HIS AGENT ACTING ON HIS BEHALF

1 JUNE 2004

Bibliothèque
Paul-Émile-Boulet

UQAC
Université du Québec
à Chicoutimi

Mise en garde/Advice

Afin de rendre accessible au plus grand nombre le résultat des travaux de recherche menés par ses étudiants gradués et dans l'esprit des règles qui régissent le dépôt et la diffusion des mémoires et thèses produits dans cette Institution, **l'Université du Québec à Chicoutimi (UQAC)** est fière de rendre accessible une version complète et gratuite de cette œuvre.

Motivated by a desire to make the results of its graduate students' research accessible to all, and in accordance with the rules governing the acceptation and diffusion of dissertations and theses in this Institution, the **Université du Québec à Chicoutimi (UQAC)** is proud to make a complete version of this work available at no cost to the reader.

L'auteur conserve néanmoins la propriété du droit d'auteur qui protège ce mémoire ou cette thèse. Ni le mémoire ou la thèse ni des extraits substantiels de ceux-ci ne peuvent être imprimés ou autrement reproduits sans son autorisation.

The author retains ownership of the copyright of this dissertation or thesis. Neither the dissertation or thesis, nor substantial extracts from it, may be printed or otherwise reproduced without the author's permission.

RÉSUMÉ

Le sujet de ce mémoire s'inscrit dans le cadre de la mesure de la confiance d'un agent logiciel. Les systèmes multi-agents actuels ne prennent pas en considération les problèmes de cohabitation entre un utilisateur et son agent virtuel agissant en son nom, telles que la confiance mutuelle et la délégation sécuritaire de tâches. Ainsi, la cohabitation utilisateur-agent nécessite une compréhension mutuelle, ce qui signifie que les deux entités devraient être aptes à comparer leurs points de vue ou opinions respectifs avant la délégation d'une tâche commune. Pour répondre à cette problématique, le modèle théorique proposé utilise la logique terminologique comme une approche ontologique pour garantir l'égalité sémantique de leurs opinions avant la mesure du degré de confiance mutuelle. Ce modèle non seulement réduit au minimum la communication entre l'usager et son agent mais aussi constitue une solution pour la délégation sécuritaire de tâches. Le cas concret servant de validation concerne le domaine du commerce électronique.

CATALOGUE OF THESIS

RÉSUMÉ ii

CATALOGUE OF THESIS iii

LIST OF TABLES v

LIST OF FIGURES vi

LIST OF FIGURES vi

CHAPTER 1 INTRODUCTION 1

CHAPTER 2 TRUST IN CYBER-SYSTEM 8

2.1 What is trust? 9

2.1.1 Definition of trust 9

2.1.2 Why trust 10

2.1.3 Kinds of trust 11

2.1.4 Risk and trust 12

2.1.5 Utility and importance 13

2.2 What is security? 14

2.2.1 Security goals 14

2.2.2 Encryption technology 15

2.2.2.1 Encryption mechanism *16*

2.2.3 RSA algorithm 17

2.2.4 PKI (Public-Key Infrastructure) technology 19

2.3 Related work 20

2.3.1 Marsh's work 21

2.3.2 The work of Castelfranchi and Falcone 25

2.3.3 Pavlou's work 29

2.3.4 Chandrasekharan's work 33

2.3.5 The work of Griffiths and Luck 38

2.3.6 The work of Gefen, Srinivasan and Tractinsky 41

2.4 Conclusion 45

CHAPTER 3 AGENT'S PARADIGM 48

3.1 Introduction 49

3.1.1 Agent's notion 49

3.1.2 Multi-Agent System (MAS) 51

3.1.3 Software agent 53

3.1.4 Mobile agent 54

3.1.5 Interface agent 54

3.1.6 Reactive agent 55

3.1.7 Cognitive agent 55
3.2 BDI Model 56
3.2.1 Belief 56
3.2.2 Desire 57
3.2.3 Intention 58
3.2.4 State 58
3.2.5 BDI Architecture 58
3.2.6 BDI Semantic 61
3.3 Interaction in multi-agent system 64
3.3.1 Interaction's situations 65
3.3.2 Cooperation 68
3.3.2.1 Collaboration 70
3.3.2.2 Coordination 70
3.3.2.3 Conflict resolution 71
3.3.2.4 Delegation 71
3.4 Conclusion 73

CHAPTER 4 MUTUAL TRUST MODEL FOR USER IN THE LOOP 75
4.1 Introduction 76
4.2 Terminological logic 77
4.3 Mutual trust approach 77
4.3.1 Similarity notion 79
4.3.2 Viewpoint concept 80
4.3.3 Trust of the agent 82
4.3.4 Trust of the user 84
4.3.5 Mutual trust 84
4.3.6 Trust and risk 86
4.3.7 Mutual trust and security 87
4.3.7.1 Empirical test 88
4.4 Validation 89
4.4.1 Application's architecture 92
4.4.2 Evaluation scenario 95
4.5 Conclusion 102

CHAPTER 5 GENERAL CONCLUSION 104

REFERENCE 109

APPENDIX 1 THE RESULT OF INVESTIGATION 118

LIST OF TABLES

Table 3.1 : Classification of interaction situations 65

Table 4.1 : Regression analysis for the factors of trust 90

LIST OF FIGURES

Figure 2.1 : Basic PKI architecture 20
Figure 2.2 : Conceptual framework 32
Figure 2.3 : Agent design 36
Figure 2.4 : Mediating relationship 43
Figure 2.5 : Moderating relationship 44
Figure 3.1 : A multi-agent system representation according to Ferber 52
Figure 3.2 : BDI state 57
Figure 3.3 : BDI architecture 59
Figure 3.4 : The model of delegation 73
Figure 4.1 : An example of viewpoint 78
Figure 4.2 : Opinions based on terminological logic 81
Figure 4.3 : Electronic market's architecture 91
Figure 4.4 : Interface of the client application 93
Figure 4.5 : Window of agent's viewpoint 94
Figure 4.6 : Window of user's viewpoint 94
Figure 4.7 : Comparison from viewpoints 95
Figure 4.8 : Result of viewpoints' comparison 101

CHAPTER 1
INTRODUCTION

In our lives, trust is a common phenomenon. Every day, we make trusting decisions many times [Luhmann, 1979]. For example, every morning we go to work and we believe we can get there, after work we trust we can come back home, and so on. Thus, we can see trust as an important concept, at the same time it is also a confusing concept, which is difficult to define. This thesis is concerned with the introduction of a formalism to give a definition and way to measure the degree of trust. It approaches the concept from the point of view of an artificial intelligent agent domain. An agent is a program able to make its own decisions (autonomous), to move on the network, to exchange the information with other agents, learning, etc. The Multi-Agents Systems (MAS) is concerned by the design of agents of organizing collectively to achieve the functionalities which are required [Chaib-Draa *et al.*, 1992] [Ferber, 1995] [Wooldridge, 2000]. Trust has been studied extensively in multi-agent systems [Marsh, 1994] [Elfoson, 1998] [Jonker and Treur, 1999] [Schillo *et al.*, 2000], where it means an agent has with respect to the dependability/capabilities of some other agent (maybe itself). We are also primarily concerned with how an agent can use the degree of trust in reasoning about cooperation.

Today, more and more application areas require systems that are able to interact with its users. Socially Intelligent Agents (SIA) are agent systems that are able to connect and interface to humans, for instance, they can find themselves in the role of observer, assistant, collaborator, competitor, customer, etc. The importance of such work is demonstrated in application areas such as e-commerce, agents for training, learning and therapy environments. Therefore, in delegation of task to the agent, the user needs an action

of the agent and includes it in his own plan in order to achieve the goal through the agent. In all these application areas the user's attitudes towards the agent, in terms of believability, credibility, trust, etc., are important factors that determine the acceptance and success of such a system and its utility in real-world applications [Kerstin, 2000]. One of the major problems of this cohabitation relates to the measure of mutual trust between the user and the agent acting on his behalf.

The general objective of our research is to find out a computational model of trust based on comparing from the viewpoints of the user and those of the agent, before the takeover of a task. A viewpoint is an opinion resulted from a terminological conceptualization [Nebel, 1995] of beliefs, goals and action plans to carry out a task. In other words, we want to find those factors, which influence trust. Trust is multi-dimensioned which concerns many different attributes such as reliability, dependability, security, honesty, etc [Grandison and Sloman, 2000]. However, we want to judge the trust depending on the similarity of viewpoints of the user and agent in order to take into account the competence factor of the agent and to translate the problem of measuring trust to the classification of opinions based on terminological logic [Nebel, 1995] [Baader *et al.*, 2003]. Therefore, our prime objective consists in giving our hypothesis and formulae according to some literature theories and practical investigation. The second objective is about the comparison between the similarity in agents' viewpoints. This comparison will help us to achieve our goal to use the formula to calculate trust. Finally, our last objective consists of

validating the proposed approach on a concrete case, which relates to the electronic commerce.

It takes several steps to achieve these goals. At first, we presented a detailed review related to the existing methods of trust, such as the Marsh's work [Marsh, 1994], which proposes an analytical approach to trust. This formalism provides a tool for measuring trust and its applicability in the MAS domain. This work is a step in the direction of a proper understanding and definition of human trust. The work of Castelfranchi and Falcone [Castelfranchi and Falcone, 2001] is about social trust. It focuses on a cognitive approach to trust. The work has shown how trust is the mental background of delegation in their relationship, and it implies trust has been derived from its reference to a goal from the action of delegating and from the uncertainty of trust-beliefs. A distributed cognition approach comes from Chandrasekharan's work [Chandrasekharan, 2002], which considers trust as a distributed cognition problem. It also suggests an agent design framework inspired by distributed cognition as well as a programming language that can act as an institution to partially solve the trust problem. We have also made an investigation into the relationship between trust and security by analyzing the work of Pavlou [Pavlou and Ramnath, 2002], which supposes trust is influenced by perceived information security and distinguishes it from the objective assessment of security threats. Finally, we look at the study of the relationship between trust and risk by the work of [Griffiths and Luck, 1999] [Gefen *et al.*, 2003], which emphasizes that cooperation inherently involves an element of

risk, due to the unpredictable nature of other's behavior, and they suppose some formula to show the relationship between trust and risk.

These models propose a general framework for automated trust. Nevertheless, they have not taken into account the specificity of the user in cohabitation with the agent acting on his behalf, because the divergence of the opinions between them is not necessarily conflicting. It can be seen like two different interpretations of the same reality and it becomes impossible to trust entirely one or the other [Bouzouane, 2002]. For instance, in e-commerce, these two entities (agent and user) have common goal which consists of purchasing product. The plan of the agent may be partial. We suppose that the agent decides to supply the product but it did not yet choose the provider. It plans to find a set of wholesalers, negotiate a price and a method for the payment, checks the product and finally determines the date of supplying a product. On the other hand, the user who maybe wishes to intervene in order to take control of the transaction by proposing a similar plan but with other constraints such as the date of supply of the selected item and the supplier of the product are known. These two opinions are similar even though they use different vocabularies and constrains. For example, the actions of «contact supplier» and «find wholesalers» are semantically similar. Therefore, rather than communicating directly with the agent or losing the control to benefit of the user, we propose to measure the mutual trust the commercial transaction takes place. In our second step, we thus put forward our formula based on the comparison of the viewpoint concepts between user and agent by using the

terminological logic to guarantee their semantic equality between the viewpoints and then measure the trust in order to ensure a safe delegation of common task.

Finally, we realize an application by using as a validation relates to the electronic commerce.

The following chapter 2 presents a deeper discussion about our subject and introduces the concept of trust. We also present a comprehensive survey of much of the literature on trust. We want to provide an understanding of what work has been done with trust. Formalism is a major contribution to these works, which provides a tool for the discussion and clarification of trust; using this formalism, we are able to ascertain whether what we are representing is "trust" and modify it accordingly.

Chapter 3 is about the agent. It constitutes a state on the cooperation in an agent's society. Also, it introduces the basic notions of the field and presents a review of the various methods of cooperation previously quoted. This chapter also make it possible to position our work in this field of cooperation in multi-agent systems. In addition, we introduce the BDI model [Rao and Georgeff, 1992], which has inspired our work.

Chapter 4 consists of our contribution in the trust domain and makes it possible to present a formal model of mutual trust between user and his agent acting on his behalf. Initially, we present the basic concepts of the proposed approach, followed by a formal

formula of the model. Finally, a concrete case is used to validate our model, which relates to the electronic commerce is presented. Also, we present a conclusion of some of the results obtained from implementations of formalism. Let us mention that an article, presented in the Seventh international conference on International Association of Science and Technology for Development (IASTED'03) [Bouzouane, Bouchard and Shen, 2003].

The general conclusion of this thesis make it possible to present the assessment of our contribution to the advance of the field as well as the new ways of interesting research, which could result from this.

CHAPTER 2

TRUST IN CYBER-SYSTEM

2.1 What is trust?

Trust is undoubtedly an important feature of our everyday lives. We often say "I trust you" but what does that mean? Does it mean, for example, that I trust you more than 60% of what I consider to be complete trust, or some other arbitrary figure, which means that my trust in you is greater than some threshold value? Or is it a general statement of fact, which requires analysis on any action that should be taken? People talked about it from different points of view: biology, sociology, social psychology, economics, history and philosophy [Marsh, 1994].

2.1.1 Definition of trust

Trust (or symmetrically, distrust) is a particular level of the subjective probability with which an agent assesses that another agent or group of agents perform a particular action, both before it can monitor such an action (or independently or his capacity ever to be able to monitor it) and in a context in which affects his own action [Gambetta, 2000]. For instance, in the case of the Internet relationship, the consumer would believe that a benevolent vendor, who has promised to provide goods and services in a proper and convenient way, would do so.

The term "subjective probability" is important in the above definition, because it points to a certain amount of arbitrariness in the trust metric. Trust is not something that can be fully obtained using objective measures, but a subjective degree of belief about others' competence and disposition. That means that a person depends on another person

with a feeling of relative security, even though negative consequences are possible [McKnight *et al.*, 1998]. An example is the famous prisoners' dilemma [Deutsch, 1958] [Solomon, 1960].

Therefore, when we say we trust someone or that someone is trustworthy, we implicitly mean that the probability that he will perform an action that is beneficial, or at least not detrimental to us, is high enough for us to consider engaging in some form of cooperation with him. Correspondingly, when we say that someone is untrustworthy, we imply that probability is low enough for us to refrain from doing so.

In general, trust supposes a situation of risk, which integrates various views [Marsh, 1994]:

- A means of understanding and adapting to the complexity of the environment
- A means of providing added robustness to independent agents
- A useful judgment in the light of experience of the behavior of others
- Applicable to inanimate others (including artificial agents)

2.1.2 Why trust

Many techniques, such as contracts [Ingersoll, 2000] and signatures [Atreya *et al.*, 2002], have been developed to assume secure and reliable communication between agents, but they are not yet satisfied. For instance, in the electronic commerce domain trust has been recognized as one of the key factors for successful electronic commerce adoption. In

that field, trust is magnified, because agents reach out far beyond their familiar trade environments and communicate with someone they will never meet.

The concept of trust provides us with an ideal measure of risk since we cannot assume to know what the behavior of the agent may be at any given time. Therefore, the notion of trust relies on a judgment based on experience, coupled with past knowledge of the agent to be trusted and their behavior [Shen *et al.*, 2001].

2.1.3 Kinds of trust

Trust is applied to a range of phenomena, involving objects, processes, and people. Three general types of trust have been identified [Chandrasekharan, 2002]:

a) **Dispositional trust** describes an internal state of the trustor, a basic trusting attitude. This is a sense of basic trust, which is a pervasive attitude towards oneself and the world [Abdul-Rahman *et al.*, 2000]. Suppose that we asked an employee if she trusted her newly hired manager, whom she had never before met, she said that she did trust her, because she always trusts new people until they gave her some reason not to trust them.

b) **Impersonal trust** refers to trust on perceived properties or reliance on the system or institution within which the trust exists. For example, people believe in the efficiency of a bank to take care of their money because of laws and

institutions like the Federal Deposit Insurance Corporation (FDIC) that assure against loss.

c) **Interpersonal trust** refers to the trust one agent has on another agent directly. This can be seen as dispositional trust directed towards an animate system. This trust is agent and context specific. For instance, person *A* might trust person *B* in the context of fixing a furnace, but not for fixing a car. This trust relates to the behavior of the agent. The behavior is directly caused by trusting intentions and trusting beliefs of the agent because it tends to translate their beliefs and intentions into actions, which are influenced by dispositional trust [Nelson and Cooprider, 1996].

2.1.4 Risk and trust

Risk was broadly defined as an attribute of a decision alternative that reflects the uncertain and variance of its possible outcomes [Gefen *et al.*, 2003]. From much literature [Griffiths and Luck, 1999] [Marsh, 1992], risk is said to be intimately related to the amount of perceived costs and benefits of a situation: the higher the potential costs, the higher is the risk, with the amount of benefits having a more or less equal effect on risk, so that, in a simple estimate:

$$Risk = \frac{Costs}{Benefits}$$

Trust reflects a willingness to assume the risk of a situation and trust can be represented a function of the degree of these risks [Mayer *et al.*, 1995]. For example, risk in electronic commercial transactions is primarily created by threats of information security.

2.1.5 Utility and importance

Utility means that property in any object, whereby it tends to produce benefit or to prevent the happening of risk [Marsh, 1994]. Sometimes, the utility can measure the total benefit attaching to each of a set of alternative courses of action. While this may well be feasible in situations with few outcomes, which are known with some certainty, there will be situations where some outcomes are unknown, or have a probability of unknown occurrence. An agent can thus rely on, such as a weighing up of the costs and benefits that it estimates that the situation to hold.

It may seem that the importance and the utility are one in the same, although the distinction is not obvious. In particular, utility is generally measurable, or at least relatively straightforward to find an estimate for, whereas importance is a subjective judgment of a situation on the part of the agent concerned. The agent itself may change, it may receive specific orders to carry out some actions, which make those actions much more important at one time than another. In addition, trust cannot be based on rationality alone [Hertzberg, 1988] [Lagenspetz, 1992]. The subjective concept of importance allows something additional to rationality to be considered. Importance gives the formalism of trust added prescriptive and descriptive power.

2.2 What is security?

Security and trust are easily confused. Just as what we talked about above, while trust is a subjective idea, the security is a technological concept [Pavlou and Ramnath, 2002], which includes encryption technology [Lamsal, 2001], Public-Key Infrastructure (PKI) technology [He *et al.*, 1998], and so on.

The objective of Information Technology (IT) security is to ensure an implementation of a system with the security goals.

2.2.1 Security goals

Security goals can be achieved by considering the security requirements of information technology [Stoneburner, 2001]:

- **Availability:** assures that the system works properly and the services are available or not denied to authorized users. It should also ensure that the system is available to the intended users and for intended use only.

- **Integrity:** means that the data is free from unauthorized manipulation of data, which can happen either in storage, during processing or during transmission. Like data integrity, system integrity means that the system has not been manipulated or even accessed in an unauthorized manner. To protect data against

this sort of attack, cryptographic techniques are required, for more details the section 2.2.2 will develop this aspect.

- **Confidentiality:** means that only the intended user receives the information and that the information is not disclosed to any unauthorized individual. The confidentiality principle applies to data in storage, processing, and in transmission. One of the methods to realize is RSA algorithm [Lenstra and Verheul, 1999], we will talk about it in section 2.2.3.

- **Accountability:** is a requirement that actions of an entity must be traced uniquely to that entity. It becomes significant for issues like non-repudiation, fault isolation, intrusion detection and prevention, after-action recovery and legal action. For example, in the economic commerce, the vendor should be responsible for his actions.

- **Assurance:** is required to show that the security measures have been properly implemented and they work as intended.

2.2.2 Encryption technology

Cryptography helps us to achieve the security goals. Usually, encryption is defined as the process of translating information from its original form (called plaintext) into an encoded, incomprehensible form (called ciphertext), in order to ensure privacy and

authentication of messages and between agents [Diffie and Hellman, 1976] [McKnight *et al.*, 1998]. A system provides privacy or confidentiality when the message it sends is exclusively accessible to only the specified receiver, and the sender is assure this will happen. Similarly, a system provides authentication if the message is exclusively accessible only to the sender and the receiver is assured this will happen. To illustrate, let us consider an example. An entity Alice sends a message to another entity Bob through a communication channel. The communication channel provides privacy or confidentiality if only Bob can receive the message and Alice knows this fact that only Bob has access to the message. The system provides authentication if only Alice has access to the message and Bob knows this fact that it is only Alice who has access and therefore she must have originated it.

2.2.2.1 Encryption mechanism

The encryption mechanisms are a combination of complex mathematical algorithms and keys. The process of encryption on the Web is implemented through the use of Web servers and browsers that are built with this technology referred to as secure socket layer (SSL) [Mitchell *et al.*, 1998]. A cryptographic system (S_K) can be as follows:

$$S_K: \{P\} \rightarrow \{C\}$$

This is an invertible transformation from $\{P\}$ to $\{C\}$, where: P is a space of plaintext message; C is a space of ciphertext message; K is the key and is selected from a finite set $\{K\}$ called keyspace.

The key is used to both encrypt and decrypt messages to guarantee privacy and authentication. Public key cryptography involves a pair of keys for each of the two communicating parties, one of which is made public and the second one is private to the entity. The public key is used to encrypt a message and the private key is used to decrypt a message. For instance, when Alice wants to send a message to Bob, she takes Bob's public key, which is available in public, encrypts the message with this public key and sends the encrypted message to Bob. Bob then receives the encrypted message and decrypts it with his private key. Also, Bob can use Alice's public key to verify a digital signature signed by Alice using her private key. One of the techniques used to support the encryption is RSA algorithm.

2.2.3 RSA algorithm

The RSA algorithm was invented in 1978 by Ron Rivest, Adi Shamir and Leonard Adleman [Lenstra and Verheul, 1999]. Here is the general idea of the encryption algorithm:

1) *Find two large prime numbers P and Q (e.g., 1024-bit);*

2) *Choose* E *such that E is greater than 1, E is less than P×Q, and* E *and (P-1) × (Q-1) are* relatively prime, *which means they have no prime factors in common.* E *does not have to be a prime, but it must be odd. (P-1) × (Q-1) can't be prime because it's an even number.*

3) *Compute D such that (D×E - 1) is evenly divisible by (P-1) × (Q-1). Mathematicians write this as D×E = 1 (mod (P-1) × (Q-1)), and they call D the* multiplicative inverse *of E. This is easy to do -- simply find an integer X which causes D = (X × (P-1) × (Q-1) + 1)/E to be an integer, then use that value of D.*

4) *The encryption function is C = (T^E) mod P×Q, where C is the ciphertext (a positive integer), T is the plaintext (a positive integer), and ^ indicates exponentiation. The message being encrypted, T, must be less than the modulus, P×Q.*

5) *The decryption function is T = (C^D) mod P×Q, where C is the ciphertext (a positive integer), T is the plaintext (a positive integer), and ^ indicates exponentiation.*

The "public key" is the pair of numbers ($P \times Q$, E). This can be published freely. Your "private key" is the number D, and must be kept secret. This means that anyone can encrypt messages to me using my public key, but only I can read them using my private key.

This works because there is no known way to work out D, P or Q given ($P \times Q$, E), except to factorize $P \times Q$. If each of P and Q has around 1024 digits, in binary, this factorization would take billions of years using present-day computers.

2.2.4 PKI (Public-Key Infrastructure) technology

A big question of public key cryptography is how does an entity trust that the publicly available key is the genuine key belonging to the other entity it is trying to communicate with? For instance, how can Alice be perfectly sure that the public key she obtains from a public domain really belongs to Bob? For this reason it is critical to have a proper key management system before the public key cryptography can be used securely.

PKI defines a framework for obtaining and trusting a public key of an entity in order to encrypt information to be decrypted by that entity or in order to verify the digital signature [Atreya *et al.*, 2002] of that entity.

In order to establish the relationship between the entity and its public key, PKI uses a data structure called a certificate, which binds the entity's identity with its public key and also contains information on how to use the public key. The certificate resolves only one part of the trust issue but the issue that still needs resolving is how to trust a certificate. The accepted solution within PKI to resolve this issue is to use a trusted entity called a Certificate Authority (CA) [Lamsal, 2001] to issue the certificates. The CA can digitally sign a certificate and send it to a requesting entity. This entity, which trusts the CA, can verify the CA's digital signature. If this verification is successful, the entity can believe that the certificate it has received is genuine and then can use the public key inside the certificate. Figure 2.1 illustrates this concept.

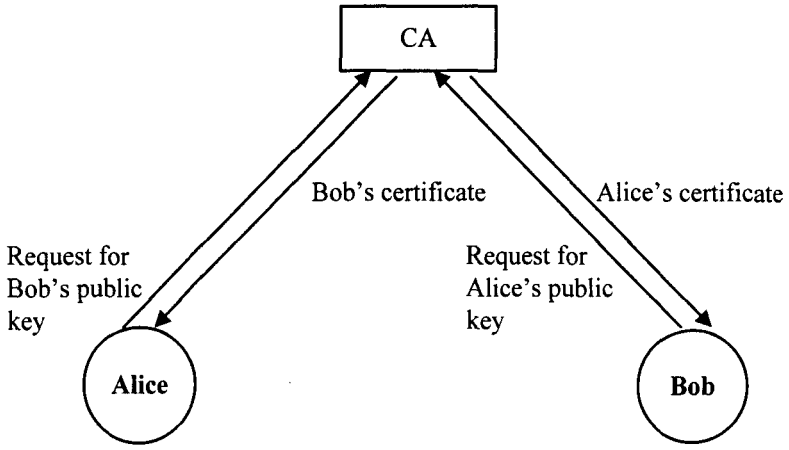


Figure 2.1 : Basic PKI architecture

Here, both Alice and Bob trust the certificate authority CA. Both of them can request the CA for each other's public key. CA digitally signs the certificate, which both Alice and Bob must be able to verify using CA's public key. Once the verification is successful, they can use the public key inside the certificate to communicate with each other.

2.3 Related work

The available literature on trust is substantial, considering several people have worked in this field, and their contributions have been significant. Within our area, the following work is inspiring.

2.3.1 Marsh's work

Marsh's work [Marsh, (1992, 1994)] is often referenced because his analysis is thorough. It suggests that trust includes:

- **Basic trust** is a disposition to trust. The basic trust of agent x is represented by T_x Basic trust is not directed to any particular agent or situation, but is a state derived from past experiences in all situations.

- **General trust** is not relative to any specific situation α, it simply represents an interpersonal trust on general trust in another agent. $T_x(y)$ represents the general trust between agent x and agent y. The estimate of general trust is notated $\widehat{T_x(y)}$ for the amount x trusts y and given by the formula 1:

$$\widehat{T_x(y)} = \frac{1}{|A|}\sum_{a \in A} T_x(y) \qquad (1)$$

A: is the set of situations that is similar to the present situation α, which x has experienced with y.

- **Situational trust** is of most importance when considering trust in cooperative situations. The notation for "x trusts y in situation α" is $T_x(y,a)$ given by the following formula 2:

$$T_x(y,a) = U_x(a) \times I_x(a) \times \widehat{T_x(y)} \qquad (2)$$

$U_x(a)$: *is the utility agent x gains from situation a;*

$I_x(a)$: *is the importance of situation a for agent x.*

In order to estimate situational trust, *x* will need to consider several aspects of the situation. It tries to increase utility as far as possible. While this is an accepted definition of economic rationality, it is felt to lack elements in the decision to make specifically situational agent-subjective measures. It is for this reason that Marsh adds a consideration of the importance of the situation to that agent.

For this reason, the agent will decide to cooperate, Marsh suggests an important formalism of a cooperation threshold which is given by formula 3:

$$Cooperation\_Threshold_x(y, \alpha) = \frac{Perceived\_Risk_x(y, \alpha)}{Perceived\_Competence_x(y, \alpha) + \widehat{T_x(y)}} \times I_x(\alpha) \qquad (3)$$

Where:

The *cooperation_threshold* is considered to be a "subjective measure, tempered by objective beliefs". Marsh considers different variations of the formula, including ones where the competence of the trustee is not known in advance.

The general trust $\widehat{T_x(y)}$ plays a role in the mediation of the cooperation threshold: a very low value of this trust will ensure cooperation is less likely to occur than if it is high.

The *Perceived_Competence* measure is based on experiences in similar situations, experiences of the same agent in similar situations, and knowledge of that agent's capabilities in similar situations, evaluated as follows:

$$Perceived\_Competence_x(y, \alpha) = \frac{1}{|A|}\sum_{b \in B} Experienced\_Competence(y, b) \times \widehat{T_x(y)} \qquad (4)$$

The *Perceived_Risk* involves a weighing up of the costs and benefits of the situation, whether it is worth risking the costs in order to obtain the benefits of the situation being resolved.

$$Perceived\_Risk_x(y, \alpha) = \frac{C_x(y, \alpha)}{B_x(y, \alpha)} \times I_x(\alpha) \qquad (5)$$

Using the concept of trust to decide whether an agent will work with another agent, Marsh gives us several possibilities:

- The most intuitive method is to select the most trusted agent to cooperate with, i.e., if $T_x(y) > T_x(z)$, here z is another agent, then x will choose to cooperate with y.

- It is possible for x to choose the agent who is trusted more in this situation, taking the maximum of the situational trust. For different agents, the utility to be gained from the same situation may be different for the trustor.

- Without taking trust into consideration, x could always decide to cooperate with the agent whose cooperation threshold was the lowest.

- x may choose to cooperate with the agent who has the largest gap between cooperation threshold and situational trust, as if:

$$T_x(y, \alpha) - Cooperation\_Threshold_x(y, \alpha) > T_x(z, \alpha) - Cooperation\_Threshold_x(z, \alpha) \quad (6)$$

For all of these choice methods, agent x's final consideration may be to take all of these choice methods, and choose to cooperate with the agent who 'wins' the most.

Marsh's work is interesting in the sense that it demonstrates the computational aspect of the trust but it has some limitations:

- It considers the trustee to be a passive entity. However, in the real world, the trustee is either positive or negative.

- The role of the environment is not captured, though trust is considered as "situated". A situation is considered as something like a "box" or a "framework" within which the trusting decision is made. However, in the real world, a situation is not a slice of time and space, but a broader intermingling of contexts.

- Trust is closely connected to reputation and social institutions. The role of these institutions is assumed, particularly in the Perceived_Competence variable, but the roles are not captured formally by the model.

- The crucial role played by communication in trust is not captured.

- Trust is considered to be a distinguishable and independent state of mind. However, there is a set of mental states (like beliefs) that contribute to trust [Castelfranchi and Falcone, 1998a].

2.3.2 The work of Castelfranchi and Falcone

The work of Castelfranchi and Falcone [Castelfranchi and Falcone, 1998a] is more cognitively oriented. It considers trust to be a "cluster" mental state, which consists of:

- **Competence belief:** an agent x should believe that an agent y can do action α. In other word, a positive evaluation of y is necessary, x should believe that y is useful

for this goal of its, that *y* can produce/provide the expected result, that *y* can play such a role in *x*'s plan/action, that y has some function.

- **Disposition belief:** *x* should believe that *y* is willing to do α. With cognitive agents this will be a belief relative to their willingness, this makes them predictable [Miceli *et al.*, 1995].

- **Dependence belief:** *x* believes it must rely on *y* (strong dependence) or *x* believes it is good to rely on *y* (weak dependence).

- **Fulfillment belief:** *x* believes that goal *g* will be achieved (thanks to *y* in this case).

- **Willingness belief:** *x* has to believe that *y* has decided and intends to do action α.

- **Persistence belief:** *x* believes that *y* is stable in his intentions, and will persist with α.

- **Self – confidence belief:** *x* believes that *y* knows that *y* can do α.

- **Motivation Belief:** *x* believes that *y* has some motives to help it.

The authors consider the action/goal pair $\tau = (\alpha, g)$ as the real object of cooperation, and they calls it "task". Then by means of τ, the author will refer to the action (α), to its resulting world state (g), or to both. Thus, the authors simplify and formalize social trust based on mental state of the agent as follows:

$$\begin{aligned} Trust\ (x, y, \tau) = {} & Goal_x\ \tau \wedge B_x PracPoss_y(a, g) \ \wedge \\ & B_x Prefer_x(Done_y(a, g), Done_x(a, g)) \ \wedge \\ & (B_x(Intend_y(a, g) \wedge Persist_y(a, g)) \ \wedge \\ & (Goal_x\ (Intend_y(a, g), Persist_y(a, g))) \end{aligned}$$

Where:

- $PracPoss_y(a, g) = g \wedge Ability_y\ (a)$
- $Goal_x\ \tau$: the goal of agent x is to do task τ
- $B_x PracPoss_y(a, g)$: agent x believes y has the same goal and y has the ability to do task τ
- $B_x Prefer_x(Done_y(a, g), Done_x(a, g))$: agent x will compare who is the best to do it.
- $B_x(Intend_y(a, g) \wedge Persist_y(a, g)$: agent x believes y has intentions and y will do it for a long time
- $Goal_x\ (Intend_y(a, g), Persist_y(a, g))$: the goal of agent y intentions and y will do it for a long time.

Therefore, the degree of trust is based on the "strength" of its component beliefs. The notion of the degree of trust of x in y about τ is $DoT_{xy\tau}$ $(0 \leq T_{xy\tau} \leq 1)$ and is given by the following formulas:

$$DoT_{xy\tau} = DoC_x\ [Opp_y(a, g)]\ \times\ DoC_x\ [Ability_y(a)]\ \times\ DoC_x\ [WillDo_y(a, g)] \quad (1)$$

Where:

- $DoC_x\ [Opp_y(a, g)]$, is the degree of credibility of x's beliefs about y's opportunity of performing α to realize g;
- $DoC_x\ [Ability_y(a)]$, is the degree of credibility of x's beliefs about y's ability/competence to perform α;
- $DoC_x\ [WillDo_y(a, g)]$, the degree of credibility of x's beliefs about y's actual perform;

Where, $DoC_x\ [WillDo_y(a, g)]$ is defined as follows:

$$DoC_x\ [WillDo_y(a, g)] = DoC_x\ [Intend_y(a, g)] \times DoC_x\ [Persist_y(a, g)] \quad (2)$$

Where, $Intend_y(a, g)$: means that y intends to do α in order to achieve g

$Persist_y(a, g)$: means that y persists to do α in order to achieve g

The decision to trust is based on this rule:

$$T_x(y, a) > Cooperation\_Threshold_x\ (y, a) \Rightarrow Will\ Delegation\ (x, y, a) \quad (3)$$

Here the authors talk about the concept of delegation, which is a decision of transferring the realization of the task. We will discuss "delegation" in more detail in Chapter 3.

This work has some limitations. One of the problems is the broad scope of the competence belief. For instance, it is not clear why a "fulfillment belief" is needed. The other problem is this work proposes very complex formulas, which are difficult to compute in some domains. Also, as Marsh's work, it depends on the modeling of the trustee's mental state to get to a trust metric, focuses on the internal state of the trustor and ignores communication, which is a crucial component of any trusting decision.

2.3.3 Pavlou's work

We are interested in Pavlou's work [Pavlou and Ramnath, 2002] because of his contribution to the definition of security and the method on how to combine security and trust.

In this work, security is defined as the subjective probability with which consumers believe that their personal information will not be viewed, stored or manipulated during

transit or storage by inappropriate parties. As a recent phenomenon, the Internet customer perceptions of information security have been influenced by certain factors such as encryption, authentication, protection and verification. In these factors, the concept of verification is hard to define. It means the most important difference between electronic and traditional transactions is the lack of implicit identity verification associated with the transaction. In some domains, it is not only easy for someone to create a phony web page, but it is also equally possible for a malicious operator to create an entirely spurious web site. All consumers need to know that Citibank is housed at "www.citibank.com" and not at "www.citibank.net" or even that Citibank is spelt with an "*i*" and not a "*y*" as in Citybank?

Pavlou gives the hypothesis formula of Security as follows:

$$Security = \alpha_0 + \alpha_1 Encryption + \alpha_2 Protection + \alpha_3 Verification + \alpha_4 Authentication$$

He also ran an investigation based on regression analysis to obtain the values of the coefficients α_i (α_i =*0, 1*). An empirical study with 179 participants was performed. The participants were asked to assess the degree of perceived security with which they have little experience. The analysis indicates that the effect of encryption, protection and authentication on the perceived security were significant; while verification had an influence it was non-significant. Thus he obtained the formula as follows:

$$Security = 0.19Encryption + 0.29Protection + 0.11Verification + 0.16Authentication$$

Also, Pavlou advances the concept of trust as follows:

$$Trust = \beta_0 + \beta_1 Security + \beta_2 liability + \beta_3 Reputation$$

Where:

- *Financial liability*: Assuming that consumers are mostly concerned with the monetary aspect of electronic commercial (EC) transactions. There may be a reason why credit cards are the most common means for financial instruments in EC transactions. For instance, now Visa (www.visa.com) offers zero liability on such transactions, thus completely removing monetary risks from consumers.

- *Reputation*: The proposed definition of trust in EC transactions is essentially twofold; trust in Internet retailers and also trust in the security of the underlying medium. To show that perceived security indeed engenders trust in EC transactions, the author controls for the effect of store reputation in EC transactions.

The complete conceptual framework is shown in Figure 2.2.

Pavlou used the same method, ran an investigation and used least-squares regression analysis to analyze data and to obtain these results: the effect of perceived security and reputation on trust was significant; in contrast, the effect of financial liability on trust was non-significant. Therefore the author gets:

$$Trust = 0.53Security + 0.07Liability + 0.20Reputation$$

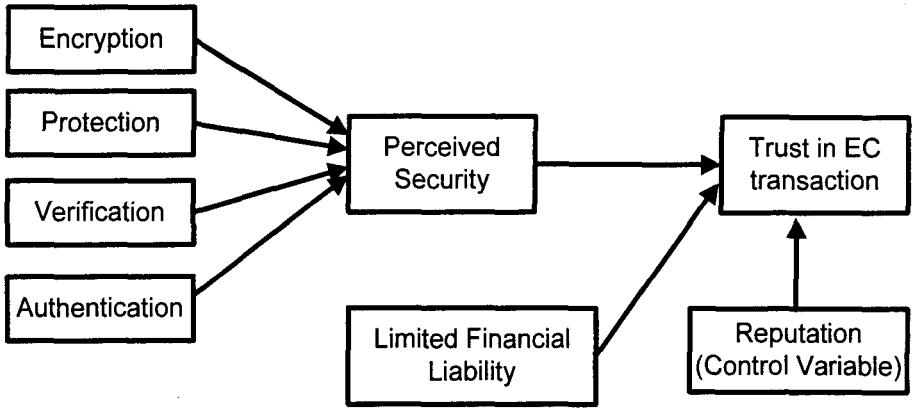


Figure 2.2 : Conceptual framework

In his work, Pavlou proposes a set of antecedents to perceived security from technological mechanisms that are visible and therefore perceptible to the consumer. Also, he proposes a subjective approach to combine a trust and security. However, this approach is limited to Electronic-commerce domain, for example, some customers do not choose the Internet retailers, because they are in different fields, although they believe in the Internet retailers' security. For the trust between agents, the security is not the only important factor. For instant, an agent wants to cooperate with other agents, at first, what it wants to know are not their intentions but their security. He also suggests the subjective attitude of the method, but he does not discuss the manner to obtain the estimation of reputation and those variables in the formula of security.

2.3.4 Chandrasekharan's work

The model of trust proposed by Chandrasekharan [Chandrasekharan, 2002] is based on communication, which is a crucial component of any trusting decision. When an agent is faced with a situation that it cannot compare with an internal structure on which to base a decision, the agent has to do a large amount of querying or communication with the world, and use that information to make a decision.

Chandrasekharan believes that this representational querying process is fundamentally the understanding of the trust problem. Human beings have the following unique capacity: they can have one internal state, and express a very different one. In other words, the internal state and its representation need not be directly correlated. It is called a gap between the representations and the internal states. Then, the aim of this work is; inter-agent trust involves making sure that there is no representational gap. That is, making sure that the external expressions of people correlate with their postulated internal states.

However, if human being's language and behavior were a direct one-to-one correspondence between internal state and external expression, human beings do not have to run elaborate inference procedures to trust people. For example, the dog language and behavior is like a protocol, where the expressions have specific and particular meanings. For systems that work using protocols, there is no need for extensive querying or trust calculation. All we need to do is to compare representational patterns. For trust situations involving agents with such a link between internal structure and external representation, the

calculation of the trust metric is essentially a process of verifying the extent of the link between an agent's external expressions and its internal state, using queries.

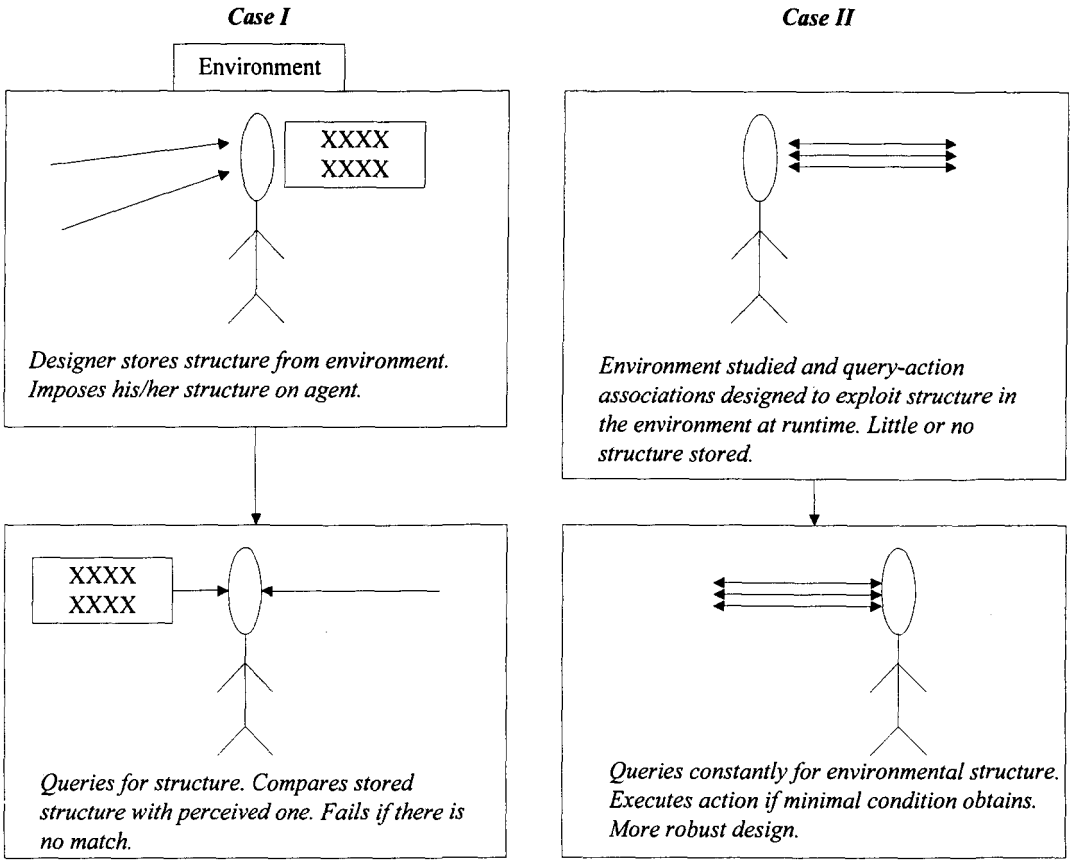
Then, Chandrasekharan considers the Distributed Cognition (DC) model is a good method to deal with the problem of the gap between the representations and the internal states. DC model considers intelligence to be spread out among other agents and functional contexts, and it emphasizes both representations and the role of the environment. Cognitive processes are considered as distributed across the members of a social group, and the functioning of the cognitive system involves coordination between internal and external structure. Processes are also distributed over time; so earlier events can influence later events.

The distributed cognition approach assumes that cognitive systems consisting of more than one individual have different cognitive properties from the cognitive properties of individuals that participate in such systems. If the task is collaborative, as in most trust situations, individuals working together will possess different kinds of knowledge. The individuals will therefore engage in interactions that will allow them to pool the various resources to accomplish the task. Since the knowledge is shared by the participants, communicative practices that exploit this shared knowledge can be used, like having a shared information structure such as a speed bug in a cockpit [Hutchins, 1995].

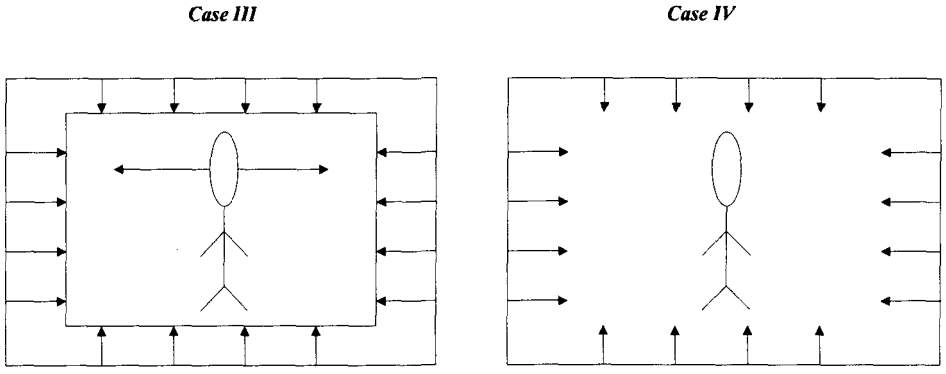
This instance of an extended mind and the focus on both internal and external representational structures makes distributed cognition an ideal framework to explore the trust problem. This is because the environment involved in an inter-agent trust decision is one of representations, and more than one agent and (possibly) artifacts is involved. Also, as Khare and Rifkin [Khare and Rifkin, 1997] point out, any trust decision involving artificial agents bottoms out to a trust decision involving humans. So the problem-space of trust is a distributed socio-technical system, consisting of people and artifacts (agents).

The author suggests an agent design framework, inspired by a distributed cognition, as illustrated in figure 2.3. In the first two cases the environment is considered as a given, and the designer makes no changes to the environment. This is passive design. In the third case, the designer actively intervenes in the environment and gives structure to it, so that the agent can function better in it. The agent only needs to query for the structure provided by the designer. This is active design, where the knowledge is split equally between the agent and the environment. The agent and the environment evolve together. In the fourth case, it is the environment that is designed, and the agent is assumed to have minimal capabilities.

Chandrasekharan applies the DC model to agent systems to suggest a programming language that can act as an institution to partially solve the trust problem. Because he tries to create an institutional structure, which guarantees that the representations generated to



Passive design



Active design

Figure 2.3 : Agent design

accurately reflect the internal state of an agent. In agent systems, this means creating Agent Communication Languages (ACLs) that reflect accurately the internal state of the agent. For this, Chandrasekharan hopes to combine the ACL with the programming language used to build the agent. This is true of interface agents as well. A human will trust an interface agent only if the agent provides the user with ways to reduce the representational gap, either by induction through extensive interaction, or through secondary sources like experts and reviews, or by being able to know, and change, the internal. This method is against the spirit of the agent paradigm, because the paradigm seeks to allow any agent, created using any language, to work with each other.

The DC approach has quite a few problems associated with the language. The major one being the restrictions the language places on the roles of agents. For instance, the notion of an agent being called a "free – agent", which can do any task that comes up, is sacrificed. Another problem is that the approach is not easily applicable to learning agents. This is because learning involves change of competence, and hence a change of roles. Thus, learning would result in the agent having a different competence to communicate. Therefore, the production language will have to allow for changes to be made to the claims. Changes in claims will weaken the guarantee provided by the language and bring up the trust problem again.

2.3.5 The work of Griffiths and Luck

The trust's model proposed by Griffiths [Griffiths and Luck, 1999] is based on an extension of BDI model (for more details see Chapter 3) by introducing the concept of trust in planning decisions.

The authors define a plan of making decision as sequence of steps, where a step either is an individual action, a joint action, a set of concurrent actions, or a sub-goal. Their aim is to choose the best plan - the plan that is most likely to be successful, with least cost in terms of time and resources, and the least risk. Therefore, they identify four primary factors relevant in comparing plans in respect of risk:

- **Agent Capabilities**: Knowledge of others' capabilities helps to determine which agents might perform the required actions.

- **Risk from Others:** Once potential cooperating agents are identified, they may be evaluated in terms of the risk involved in interacting with them. Plans involving interaction more likely to be successful should be rated higher than those involving interactions less likely to be successful.

- **Risk from view of self:** Knowledge of the view of oneself in the eyes of others in terms of risk of interaction may also be useful in assessing plans. It can provide a measure of the likelihood that another agent will agree to cooperate, since an

agent is more likely to cooperate with another if it has confidence in the success of that interaction.

- **Agent Preferences:** It might also be possible to assess plans in relation to the higher-level motivations of the agents involved in them, and whether cooperation would be likely.

In the problem of plan-selection, trust is a main factor. As recognized by several researchers [Castelfranchi and Falcone, 1998a] [Deutsch, 1962] [Gambetta, 2000] [Luhmann, 1990] [Marsh, 1994], trust implies some form of risk, and that entering into a trusting relationship is choosing to take an uncertain path that can lead to either benefit or cost depending on the behavior of others. The authors suggest an inverse relationship between trust, T, and risk, R, as follows:

$$R = \frac{1}{T} \qquad (1)$$

In assessing the merit of a plan, an agent must make a judgment about the risk attached to each action in the plan requiring cooperation, by examining the trust value in its model of each of the possible cooperating agents. Suppose that an agent knows of n others, x_1; x_2; .....; x_n, with the required capabilities for performing a given action, and ordered such that $T_{x_{n-1}} \geq T_{x_n}$, where T_{x_n} denotes the trust in agent x_n. Then, the agent would first try to cooperate with x_1 and, if unsuccessful, would then try x_2, and so on, but for each

successive agent, the likelihood of success decreases. To address this, authors give the formula as follows:

$$R_{action} = \frac{1}{\sum_{i=1}^{n} \frac{T_{x_i}}{i}} \qquad (2)$$

R_{action} : denotes the risk of the action.

Trust in all relevant agents is considered in relation to the likelihood of cooperation with them. Using this measure of risk, the author can determine the cooperative rating of a plan by summing the risk associated with each action in it. Thus a plan with few high-risk actions may be rated better (or less risky) than a plan with many low risk actions. For a plan with m actions, α_1; α_2;.....; α_m, the cooperative rating C for that plan is given by the following equation.

$$C = \sum_{i=1}^{m} R_{a_i} \qquad (3)$$

Because of an inverse relationship between trust and risk, in fact, the authors give us the relationship between trust and cooperation:

$$C = \frac{1}{\sum_{i=1}^{m} T_{a_i}} \qquad (4)$$

In this work, the authors think the limitation of BDI architecture is that it is typically focused on execution for individual agents. Then they extend a BDI-like architecture to include those higher-level control strategies. However, several questions are described: firstly, the authors consider how to incorporate the notion of an agent's rights to perform actions, not only when in terms of an agent not having the right to perform an action and so needing to cooperate, and also when assessing the risk involved in a plan in relation to the rights of other. Secondly, as Marsh [Marsh, 1994] points out, an agent's trust in another is dependent on the action being considered. This would provide a richer basis for plan selection if incorporated into the assessment of plans, but at a cost of increasing the overhead of modeling others.

2.3.6 The work of Gefen, Srinivasan and Tractinsky

In recent years, there are some troublesome trends about trust and risk [Mayer *et al.*, 1995]. This confusion in the relationship between risk and trust is expressed as follows: it is unclear whether risk is an antecedent to trust, or an outcome of trust. Gefen [Gefen *et al.*, 2003] offers the relationship between trust and risk, they discuss the concepts of trust and risk.

- **Concept of trust**

  Comparing with several studies, such as the distinction between trusting beliefs, attitudes, intentions and behavior, and the distinction between initial and ongoing trust, the authors will limit the discussions of trust to a discussion of the need for distinction between trust and trustworthiness.

Mayer [Mayer *et al.*, 1995] indicated that perceived trustworthiness is the trustor's perception of how trustworthy the trustee is, while trust, is the trustor's willingness to engage in a risky behavior that stems from the trustor's vulnerability to the trustee's behavior. Trustworthiness is a characteristic of the trustee, and may stem from several perceptions of the trustor about the trustee. For example, some people say "I believe that Amazon.com is trustworthy". According to Mayer *et al.*, the perceptions of the trustor, that affect his/her perception of the trustee, are the trustee's ability, integrity and benevolence. Trust on the other hand refers to the trustor's intentions or behavior with respect to the transaction.

- **Concept of risk**

Risk was broadly defined as an attribute of a decision alternative that reflects the variance of its possible outcomes. The more recent terminology is originated in the fields of risk assessment and risk management, which relate risk with the costs of those outcomes. Managers tend to define risk more by the magnitude of the value of the outcome, rather than by taking its likelihood into account. This may be because risk is present in a situation where the possible damage may be greater than the advantage that is sought [Luhmann, 1990]. Now let's look at the three models about the relationship between trust and risk:

a) **Mediating relationship:** It argues that the existence of trust reduces the perception of risk, which in turn increases the willingness to engage in a transaction. In other word, trust affects perceived risk, which affects behavior [Jarvenpaa *et al.*, (1999, 2000)], as shown by the following figure:

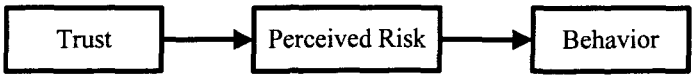


Figure 2.4 : Mediating relationship

For example the perceived risk of lending money to a trusted friend will be lower than that of lending money to a stranger. This work goes on to add that trust is one of the key factors for reducing the perceived risk of a negative outcome in a given situation. Thus, among ecommerce researchers, there appears to be an overwhelming subscription to the mediating role of risk in the relationship between trust and behavior.

b) **Moderating relationship:** It is believed that the effect of trust on behavior is different when the level of risk is low versus when the level of risk is high as shown by the following figure:

The primary belief is that when risk is high, trust is relevant; when risk is low, trust is not relevant. This means that perceived risk moderates the relations between behavior and trust. It demonstrates the moderating effect of risk on the

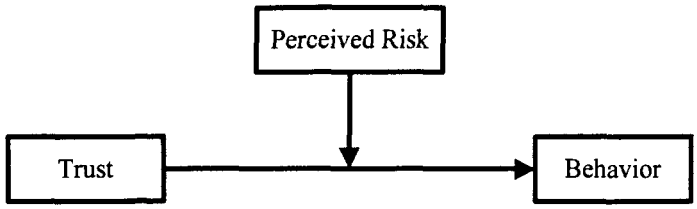


Figure 2.5 : Moderating relationship

trust-behavior relationship in an experimental study, in which subject trade with each other under conditions of high risk and low risk. This shows that trust between trading partners is higher in the high-risk condition than trust between trading partners in the low-risk condition.

c) **Threshold model:** It assumes that trust is formed independent of risk perceptions. If the level of trust surpasses the threshold of perceived risk, then the trustor will engage in a risk taking relationship. Our work is connected with these cases. For example, lending money is a risky action that requires the lender to trust the receiver. For a given level of trust, less risk (e.g., smaller amount of money) would increase the tendency to lend.

The primary goal of this work is to alert researchers in e-commerce that we are in imminent danger of expending a great deal of effort to produce a chaotic mess of empirical evidence, without the means to integrate all the evidence to get a defensible view of the role

of risk and trust in e-commerce. However the authors only give theoretical advice, but do not provide us with a practical method to deal with these questions.

2.4 Conclusion

In cyber-systems, trust is an important concept. In this chapter, we mainly talked about the definition of trust. Though it is hard to accurately define it, we can come to a conclusion that trust is a subjective project and not something that can be fully obtained using objective measures. Therefore, through analysis of the kinds of trust and the relationships between trust and risk, we found it is a subjective degree of belief about others' competence and disposition. We also hope to measure the probability of cooperation, so we raise some concepts, such as risk, utility, importance, benefit, cost, and so on, to reach our goal. Using them, we obtain some formulas, or formalism, of trust from different researchers. These formalisms are of unquestionable worth for the understanding of the trust's concept. To date, discussions of trust have suffered from vagueness and the lack of an agreed definition. Thus these formalisms present a means of establishing a clear, precise, and easily understood language for that discussion.

Another confusion is the relationship between trust and security. The reason to talk about trust is to ensure the security of communication. Thus, we discussed security from abstract definitions to quantifiable formulas. In fact, security is more technological. It includes: encryption, protection, verification and authentication. Moreover, according to Pavlou and Ramnath [Pavlou and Ramnath, 2002], security is one of those factors that

influence trust and the effect of perceived security on trust is significant. Therefore, we think these two definitions are different yet also interactive.

In our area, we study the trust between agents. Therefore we are more concerned with trust in social trust and cognitive agents' trust. We introduce a typology conceptual definition of cognitive agents' trust and the conjunction with each trust type.

At the end of this chapter, we presented the view of trust of several researchers whose work has greatly inspired us. Marsh's work is thorough. Not only did he give us the definition and formula of trust and the degree of trust, but he also raises his method about how to judge cooperation. The work of Castelfranchi and Falcone is more cognitively oriented. It considers the role of beliefs and makes the basic assumption that only an agent with goals and beliefs can trust. As well as Marsh's work, the work of Castelfranchi and Falcone will become the basis of our work. Compared with Marsh and Castelfranchi, Pavlou talks about the relationship between trust and security. He gives us some details about security to show it is a technological conception. Pavlou offers a useful method to test his hypothesis. This method will also be helpful for our work. Chandrasekharan thinks the trust models of Marsh and Castelfranchi ignore a crucial role in trust, the communication, so he tries to find a better model of trust. He introduces the concept of representation and the distributed cognition model to suggest a programming language that can act as an institution to partially solve the trust problem; Griffiths and Luck discuss and extend BDI agent architectures. In order to choose the best plan, they try to measure the

trust using the degree of risk. They raise the inverse relationship between trust and risk and give their formula. The primary goal of the work of Gefen *et al.* is to alert researchers that we are in imminent danger if we cannot integrate all the evidence to get a defensible view of the role of risk and trust in ecommerce. For this reason, the authors address three important areas: the distinction between trust and trustworthiness, the conceptualization of risk, and the relationships between risk as well as trust and behavior, and give conceptualizations and models in this area.

CHAPTER 3

AGENT'S PARADIGM

3.1 Introduction

In chapter 2, we introduced trust. Trust is a multi-dimension concept and concerns many different attributes such as reliability, dependability, security, honesty, competence, etc, which may have to be addressed based upon the environment where it is specified [Grandison and Sloman, 2000]. However, we are concerned with how the concept of trust can be used in relation to cooperation between artificial agents. In this chapter, we will introduce the common knowledge about agents system and specially will discuss BDI model [Rao and Georgeff, 1992] that can help us to understand the chapter below and how to realize trust in agents.

The evolution of artificial intelligence systems has quickly developped over recent years. More and more, we are becoming interested not only in the creation of intelligent systems, but also with the concepts of autonomy, mobility, representation of knowledge, distributed problem resolution, communication, etc. In order to answer these problems, over the past few years, the community working in the field of distributed artificial intelligence [Jennings, 1993] [Ferber, 1995] [Tambe, 1997] [Grosz and Kraus, 1998] [Wooldridge, 2000] has been trying to elaborate theories on the concepts of agent and multi-agent system.

3.1.1 Agent's notion

For the last twenty-five years, the community working in the field of artificial intelligence for distributed systems has tried to formally define the concept of the agent. So

far, there is no general consensus standard that has been obtained. However, we can note that certain definitions are usually used in literatures. In particular, the case of the definition of Wooldridge and Jennings [Wooldridge and Jennings, 1995], which is one of the most recognized and has been adopted in this research project. According to them, an agent can be defined as being a software entity (virtual, data-processing module) or material (robot), which can support the four following properties:

- **Autonomy**: this property defines that an agent has to function without the direct intervention of a human operator or other entity. It also defines that this agent will have a certain control on its actions and internal state.

- **Social Skills**: this property defines that an agent will have to interact with other agents (or possibly humans) via a communication language. A famous example in the used languages is KQML (Knowledge Query and Manipulation Language), which is a high level protocol based on the acts of language [Finin *et al.*, 1994].

- **Reaction**: this property states that an agent will have to perceive its environment and react according to the changes produced by the environment.

- **Pro-action**: this property states that an agent will not only be able to act according to what occurs in its environment, but also to act and take initiatives according to the goals it wishes to reach.

In a more formal way, we can define an agent A in a minimal manner by the following triplet:

$$A = < \Sigma, O_p, MS >$$

Σ : *represents the set of the possible states of the environnement in which agent A moves*

O_p : *represents the set of operations able to be carried out by A*

MS : *represents the set of the possible configurations for the mental states of agent A*

In a general manner, we can then see an agent A as being defined function f_A as the following:

$$f_A : \Sigma \times MS \rightarrow O_p$$

The agents are able to take action and not only to reason as in traditional systems. The action, which is a fundamental concept for the multi-agent systems, rests on the fact that the agents achieve actions, which will modify their environments, including their mental states.

3.1.2 Multi-Agent System (MAS)

A multi-agent system, such as illustrated in figure 3.1, is composed of the following elements [Ferber, 1995]:

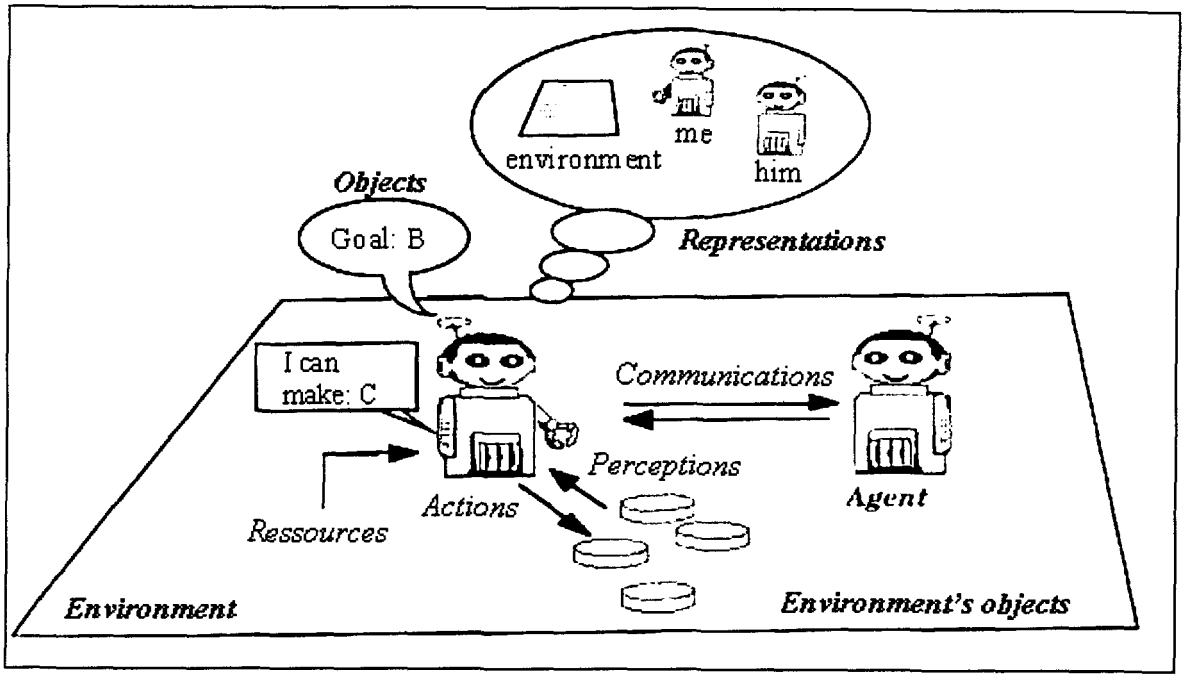


Figure 3.1 : A multi-agent system representation according to Ferber [Ferber, 1995]

a) Environment E, meaning a metric space generally layed out.

b) A unit of objects O. These objects are placed, meaning that for any object, it is possible, at any given time, to assign a position in E. These objects are passive, meaning it can be perceived, created, destroyed and modified by the agents.

c) A unit of agents A, particular objects ($A \subseteq O$), represent the active entities of the system.

d) The overall relations R link objects (and thus agents) between them.

e) A unit of operations O_p allows agents A to perceive, produce, consume, transform and handle objects O.

f) An operators unit allows to carry out various operations O_p and to produce world reactions to these operations. This is what we call the laws of the universe.

There are certain particular system cases where $A = O$ and $E = \varnothing$. In this case, relations R define a network: each agent is directly dependent on a unit of other agents. These systems, called purely communicating MAS, are very often used in the field of distributed artificial intelligence. Their field of predilection is the cooperation of software modules whose function is to solve a problem or elaborate on an expertise (such as the interpretation of signals), starting with specialized modules, as in the case of distributed system control, where E is defined by the structure of the subjacent network. These systems are characterized by the fact that the interactions are primarily intentional communications and that their working method resembles social welfare (working group, company, administration, etc).

3.1.3 Software agent

For Brenner [Brenner *et al.*, 1998] three categories of agents can be distinguished: human agents, hardware agents and software agents. Intelligent software agents are defined as being a software program that can perform specific tasks for a user and possess a degree

of intelligence that permits it to perform parts of its tasks autonomously and to interact with its environment in a useful manner.

3.1.4 Mobile agent

Mobile agents are software agents that are capable of moving from one machine to another automatically. One advantage compared to a static agent residing on a particular machine is that this can decrease the network communication load. Indeed, if we suppose to consider an agent involved in e-commerce trying to buy videos on specific topics, then the agent will first have to select which sites to visit, then for each interesting site it will request samples of videos in order to select those of interest. After having repeated the process for a number of sites, it will be able to place an order.

3.1.5 Interface agent

Interface agents facilitate the interaction between a user and a computer system. They are intended to improve interaction, such as accessing information, assisting with current work, learning, or just providing entertainment [Maes, 1997]. The kind of the agent already developed range from easy-to-program simple services for users [Malone *et al.*, 1997] to quite complex reasoning assistants [Ball *et al.*, 1997]. A major problem with interface agents is their acceptance by users. They should ideally behave like English builds [Negroponte, 1997] knowing their master's or mistress's preferences and offering services in a tactful fashion. To be able to do so, however, they require a sophisticated model of their owner, and need to be able to acquire and maintain such a model.

3.1.6 Reactive agent

In the mid eighties, Minsky introduced the concept of the Society of Mind [Minsky, 1985], a scheme in which each mind is made of smaller processes. These we shall call agents. Each mental agent by itself can only do simple things that need no mind or thought at all. Yet, when we join these agents in societies – in certain ways this leads to true intelligence. Along the same line, Brooks [Brooks, 1990] argued against founding artificial intelligence upon complex symbol systems, which are too difficult to construct and to manipulate. He offers, on the contrary, a new methodology bases its decomposition of intelligence into individual behavior generating module, whose coexistence and cooperation let more complex behavior emerge. Brooks demonstrated the approach by building mobile robots as interconnected simple modules involving sensors reacting to external conditions and driving actuators.

3.1.7 Cognitive agent

In contrast to the reactive agents described above, cognitive agents possess an explicit knowledge of their environment. The heritage off cognitive agents is clearly artificial intelligence. They use sophisticated knowledge representations, have expertise, goals, and plans, and so on. In addition to their traditional mechanisms they must interact and engage in cooperation with other agents. Many types of cognitive agents have been proposed. One of the best known agent architectures for reasoning agents is the belief-desire-intention (BDI) model [Rao and Georgeff, 1992].

3.2 BDI Model

The BDI model considers an agent as a human being, has a certain state spirit, a point of view on a situation. This mental state can be defined as the dynamic configuration of an agent's internal state in one moment. It consists of information coming from perceptions of the environment, so called the percepts, of the objectives of the agent, of the beliefs of the agent compared to the percepts and of its intentions. The model is a reference of the mental state of the agent, which has a number of applications ranging from air traffic control, telephone call centers to the handling of malfunctions on NASA's Space Shuttle [Busetta and Ramamohanarao, 1998].

3.2.1 Belief

The beliefs are the result of the perceived information of the environment and the knowledge infered by an agent starting from the information. The figure 3.2.a shows the representation of the agent's beliefs. An agent A could obtain a knowledge P_1 of an existence state s_1, coming from a perception $Percept_A(s_1)$[1]. Then, it could infer a knowledge P_2, by conducting a $Prediction_A(P_1, R, C)$[2], to leave of a series of rules of an inference R and current beliefs C, which would be showed as follows:

$$((P_1 = Percept_A(s_1)) \wedge (P_2 = Prediction_A(P_1, R, C))) \Rightarrow (Bel_A P_1 \wedge Bel_A P_2)$$

---

[1] $Percept_A(s_1)$, is an opertator (method) of perception indicating that agent A perceives a situation s.

[2] $Prediction_A(P_1, R, C)$, is an operator (method) of forecast indicating that agent A envisages a future situation to leave the knowledge P, according to a certain number of rules R and according to its current beliefs C.

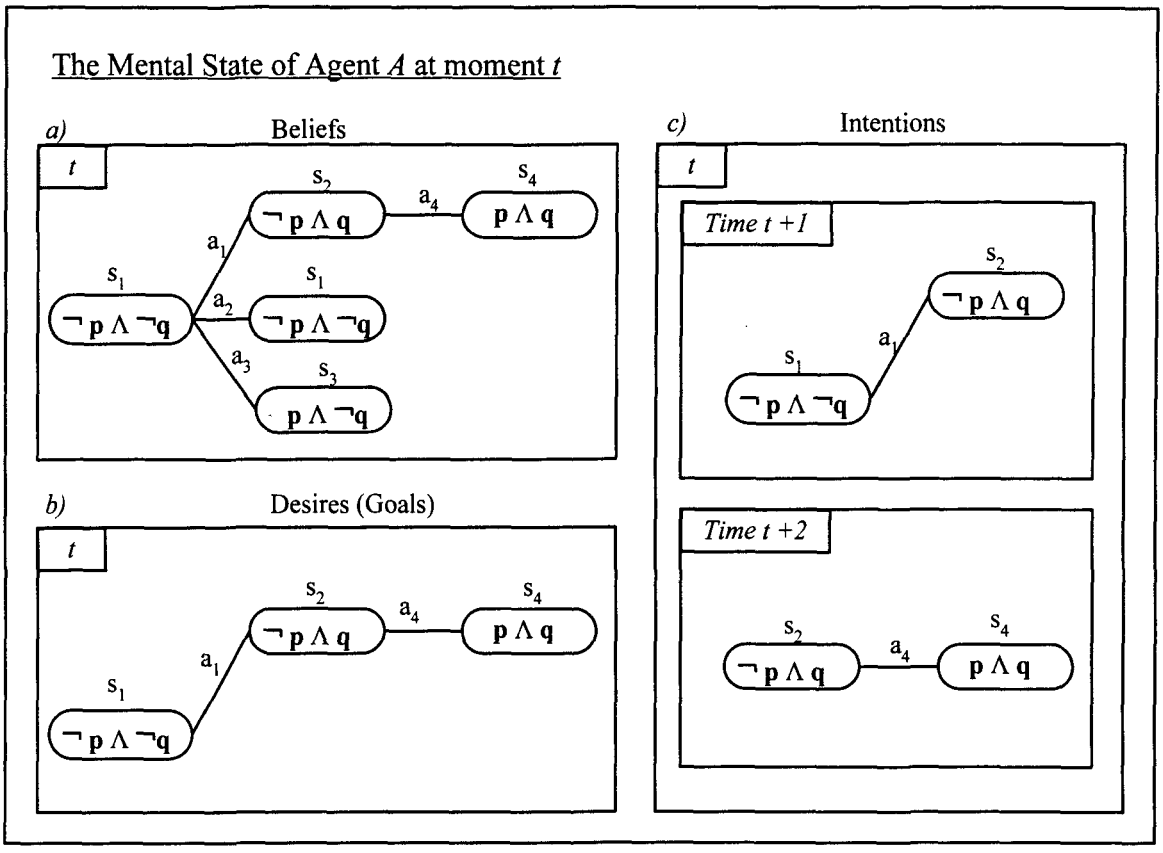


Figure 3.2 : BDI state

3.2.2 Desire

The desires or the goals of an agent constitute its motivation to act [Wooldridge and Jennings, 1995]. These goals are represented by a series of states, which the agent wishes to reach, such as shown in figure 3.2.b. Certains goals can be established during the creation of the agent. Suppose that an ore extracting robot could have the basic objective to bring back ore. Other desires can be starting from the beliefs of the agent, for example a robot extractor believes the fuel lack, it could infer a new consistent objective.

3.2.3 Intention

The intentions of an agent represent the actions that one intends to carry out in the future so that it will reach some or all its objectives (desires). Figure 3.2.c shows a representation of the intentions of an agent A at time $t+1$. In fact, these constituted intentions are starting from the beliefs and the goals of the agent's plan. Suppose that a robot extractor, which aims to bring back ore, could start to dig the rock in 10 minutes, if it believes that it will arrive close to the layer only in 10 minutes.

3.2.4 State

Figure 3.2 clearly illustrates the operation of the BDI model. We can see there is one moment t where agent A believes in a current state s_1 of the environment. It also believes that it is possible to carry out the actions a_1, a_2 and a_3 with the instant $t+1$ thus allowing to respectively reach states s_1, s_2, s_3. We can see that agent A aims to reach state s_4, and so beforehand it is necessary for him to reach state s_2. Finally, we see that the intentions of agent A with the moment T are to reach the action a_1 with the instant $t+1$ so as to reach state s_2, and then to carry out the action a_4 with the instant $t+2$, which will allow it to reach its goal, state s_4.

3.2.5 BDI Architecture

A representative BDI architecture is illustrated in Figure 3.3. As this figure shows, the BDI architecture typically contains four key data structures [Wooldridge, 1996]: beliefs, desires, intentions and plan library. A plan library is a set of plans, which specify courses of

action that may be followed by an agent in order to achieve its intentions. A data input from sensors agent's plan library represents its procedural knowledge, or know-how. A plan contains two parts: a body, or program, which defines a course of action; and a descriptor, which states both the circumstances under which the plan can be used (i.e., its pre-condition), and what intentions the plan may be used in order to achieve (i.e., its post-condition).

In Figure 3.3, we can see the agent reacts to events, which are generated by modifications to its beliefs, additions of new goals or messages arriving from the external world. An event may invoke (trigger) one or more plans; the agent then commits to execute one or more of them, that is, they become intentions. This desire is then used as a trigger for a corresponding plan from a plan library. Intentions are executed one step at the time. A

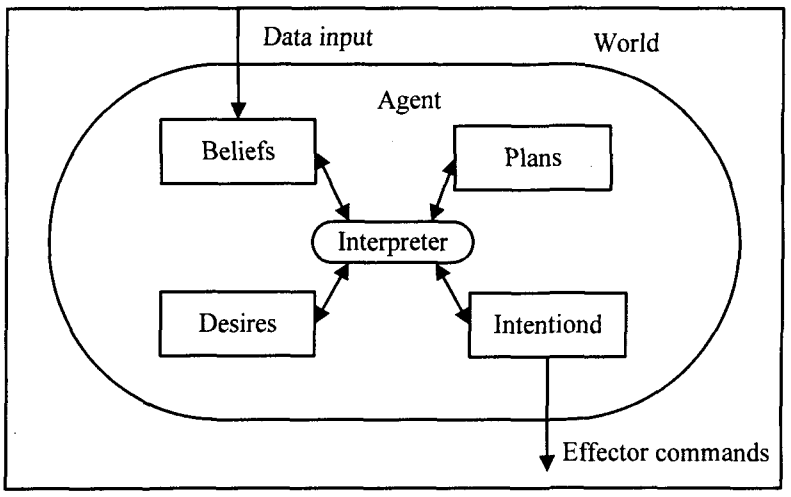


Figure 3.3 : BDI architecture

step can query or change the beliefs, perform actions on the external world, suspend the execution until a certain condition is met, and submit new goals.

The operations performed by a step may generate new events, which, in turn, may start new intentions. An intention succeeds when all its steps have been completed; it fails when certain conditions (either guarding its execution or being tested by a step) are not met, or actions being performed report errors, etc. An agent applies a set of default policies when selecting which plans become intentions, how to schedule the active intentions, etc. These can be overridden by policies defined by the user, usually invoked via the same event/plan/ intention mechanism described above.

Also, we can use an interpreter loop, as follows, is given below [Rao and Georgeff, 1995]. They assume that the event queue, belief, desire, and intention structures are global.

BDI-interpreter

initializestate();

repeat

options := option-generator(event-queue);

selected-options := deliberate(options);

update-intentions(selected-options);

execute();

get-new-external-events();

drop-successful-attitudes();

drop-impossible-attitudes();

end repeat

At the beginning of every cycle, the option generator reads the event queue and returns a list of options. Next, the deliberator selects a subset of options to be adopted and adds these to the intention structure. If there is an intention to perform an atomic action at this point in time, the agent then executes it. Any external events that have occurred during the interpreter cycle are then added to the event queue. Internal events are added as they occur. Next, the agent modifies the intention and desire structures by dropping all successful desires and satisfied intentions, as well as impossible desires and unrealizable intentions.

This abstract architecture is an idealization, including the various components of practical reasoning [Bratman, 1987]; namely, option generation, deliberation, execution, and intention handling. However, it is not a practical system for rational reasoning.

3.2.6 BDI Semantic

In order to give formal semantics to BDI architectures, a range of BDI logics have been developed by Rao and Georgeff [Wooldridge, 1996]. Most work on BDI logics has focused on possible relationships between the three "mental states" [Rao and Georgeff, 1991], and more recently, on developing proof methods for restricted forms of the logics

[Rao and Georgeff, 1995]. The traditional possible semantics worlds [Halpern and Moses, 1985] of beliefs consider each world w to be a collection of propositions and models belief by a belief-accessibility relation B linking these worlds [Rao and George, 1991]. A formula is believed in a world if and only if it is true in all its belief-accessible worlds.

We can consider each possible world to be a time tree. Each time tree denotes the optional courses of events chosen by an agent in a particular world. The belief relation maps a possible world at a time point to other possible worlds. We say that an agent has a belief ϕ, denoted $Bel(\phi)$, at time point t if and only if ϕ is true in all the belief-accessible worlds of the agent at time t. The semantics for belief can be defined formally as follows:

$M, v, w_t \mid = Bel(\phi)$ *iff* $\forall w' \in B^w{}_t \; M, v, w_t' \mid = \phi$

M : is an interpretation,

V : is a variable assignment

W : is the world, w_t' is a sub-world of the world

$B^w{}_t$: is the Belief at the time t in the world w

As the belief relation is time-dependent, the mapping of B at some other time point, say t_2, may be different from the one at t_1. Thus the agent can change its beliefs according to the available options.

The semantic of the modal operator goal is given in terms of a goal-accessible relation *G*, which is similar to that of the *B* relation. The goal-accessibility relation specifies situations that the agent desires to be in. Thus, in the same way that we treat belief, we say that the agent has a goal ϕ at time t if and only if ϕ is true in all the goal-accessible worlds of the agent at time *t*. The semantic for goal can be defined formally as follows:

$M, v, w_t \mid = Goal(\emptyset)$ *iff* $\forall w' \in G^w{}_t\ M, v, w_t' \mid = \emptyset$

$G^w{}_t$: is the Goal at the time t in the world *w*

One can view intentions as future paths that the agent chooses to follow. The intention-accessibility relation *I* will be used to map the agent's current situation to all her intention-accessible worlds. We shall say that the agent intends a formula at a certain time if and only if it is true in all the agent's intention-accessible worlds of that time.

We saw above that the goal-accessible worlds of the agent can be viewed as the sub-worlds of the belief-accessible worlds in which the agent desires to be. Similarly, one can view intention-accessible worlds as sub-worlds of the goal-accessible worlds that the agent chooses to follow (i.e., to act upon). Thus, one moves from a belief-accessible world to a goal-accessible world by desiring future paths, and from a goal-accessible world to an intention-accessible world by committing to certain desired future paths. The semantics for intention can be defined formally as follows:

$M, v, w_t \mid = Intend(\phi)$ *iff* $\forall w' \in I^w_t\ M, v, w_t' \mid = \phi$

I^w_t : is the Intention at the time t in the world w

We allow intentions over any well-formed formula, which means that agent can have intentions about intentions, intentions about goals, intentions about beliefs, and intentions to do certain actions. Some one might consider only the last type of intention to correspond with natural usage. While this is arguable, in our formalism the agent might have any type of intention but will only act on the last type of intention.

It is clear that at t_0, one of the goal formulas true in all goal accessible worlds is *inevitable*($\Diamond f$) ($\Diamond$ means eventually). This also implies that the agent believes that this goal is achievable; in other words, *Bel (optional* ($\Diamond f$)). From the beliefs, goals, and intentions of the agent, one can say that the agent believes that, if she succeeds in doing d_1, she will achieve the goal f.

3.3 Interaction in multi-agent system

The concept of interaction is in the center of problems surrounding multi-agent systems [Demazeau and Müller, 1991]. Interaction is a dynamic comparison of two or several agents by a reciprocal action. This interaction allows agents to increase their power, reach their goals, communicate their knowledge, increase the speed at which they carry out

a task, etc. Interaction between agents is also at the origin of conflicts caused by divergences from objectives and opinions, in short, by the fact that they interact.

3.3.1 Interaction's situations

There are multitudes of situations able to give an opportunity for the agents to interact. It is possible to classify these various situations compared to three main criteria: the objectives of the agents, the relations which the agents maintain towards the resources they have and the means (or competences) they have to achieve their goals [Ferber, 1995]. These criteria enable us to make a typology of the interaction situations, as shown in table 3.1. We can see that when the goals of the agents are compatible, it is possible, and in several cases even desirable, for the agents to cooperate with one another. However, when their goals are incompatible, a multitude of conflicts arise.

Table 3.1 : Classification of interaction situations

Goals	**Ressources**	**Competences**	**Types of situation**	**Category**
Compatible	Sufficient	Sufficient	Independance	Indifference
Compatible	Sufficient	Insufficient	Simple Collaboration	Cooperation
Compatible	Insufficient	Sufficient	Obstruction	
Compatible	Insufficient	Insufficient	Co-ordinate Collaboration	
Incompatible	Sufficient	Sufficient	Individual Competition	Antagonism
Incompatible	Sufficient	Insufficient	Collective Competition	
Incompatible	Insufficient	Sufficient	Individual Conflits	
Incompatible	Insufficient	Insufficient	Collectives Conflits	

- **Simple Collaboration (delegation):** this type of situation consists of a simple addition of competences not requiring any coordination between speakers. For instance, an agent not being competant to carry out a task could quite simply carry it out through another agent [Castelfranchi and Falcone, 1998b].

- **Obstruction**: this type of situation involves agents mutually obstructing the accomplissement of their tasks when they are not interdependant. For example, problems with network obstruction can come about when distant agents carry out too much communication.

- **Coordinated Collaboration**: this type of situation implies that agents have to coordinate their actions to achieve their goals [Ciancarini *et al.*, 1999]. Not having either resources, or necessary competencies, agents must cooperate. Suppose that a robot extractor of ore could cooperate with a specialized robot in the search of ore. The first robot could coordinate its actions and extract the ore only once the second has identified a layer.

- **Individual Competition**: this type of situation occurs when the agents have incompatible objectives. In other words, objective achievement implies that the others will not be able to carry out their own objectives. For example, if the objective of two agents consists of winning a part of failures that they have

disputed together, the realization of the goal will necessarily invoke failure for the other.

- **Collective Competition**: this type of situation occurs when agents having incompatible goals do not have the necessary competencies to carry out their objectives. In this case, the agents must gather within coalitions in order to acheive their goals. A characteristic example is a team competition, as in a match of soccer [Spaan, 2002].

- **Individual Conflicts for Resources**: this type of situation implies that the agents' resources are insufficient and cannot be divided. Access to resources thus becomes a source of conflict, where each agent tries to acquire sufficient resources to achieve its goals; such as several programs in competition for the use of a peripheral or printer at the same time.

- **Collective Conflicts for Resources**: this type of situation combines the collective competition with the individual conflict for resources. For example we can think of several robots playing a game of soccer being in a collective competition to obtain the one and only resource: the ball (Robocup'97) [Tambe *et al.*, 1997].

3.3.2 Cooperation

Cooperation is defined as being the pooling of resources and competence of a group of agents in order to achieve a common goal, and/or with the augmentation of their mutual performance [Galliers, 1991] [Doran *et al.*, 1997]. For example, in the case of robot extractors of ore, we could think that their regrouping would result in a simple summation of their individual performances. In other words, if a robot brings back 10 kilos of ore per hour, 5 robots working in concert will bring back 50 kilos of ore per hour. However, the improvement of their performances could be much more than linear; suppose that if certain robots concentrate on digging the ore and that the others bring the ore back, the robots' individual efficiency could go from 10 to 12 kilos per hour. In order to observe this increase, we would need to quantify performance improvement with the help of an index. In this case, the index improvement could be calculated in the following way:

$$\sigma_{improvement} = P_{total} - ((P_{ind} \times n) + a)$$

P_{total} : *the rough total performance of the robots*

P_{ind} : *the individual performance of an agent*

n : *the number of the agents*

a : *the lost performance in the management of the cooperation*

There are two large classes concerning cooperation:

a) **Cooperation as an intentional attitude**: this school of thoughts states that cooperation is the result of an attitude the agents have of wanting to cooperate after having identified a common goal [Galliers, 1991]. For example, two agents needing the mathematical result of a complex calculation could decide to associate with one another, so that each one carries out half of the calculation task then sharing the result. This intentional decision on behalf of the agents would thus constitute the proof of a true will to cooperate.

b) **Cooperation from the viewpoint of the observator**: the kind of thought considers cooperation as a qualification of the activity of a group of agents by an external observer not having mental access to the agents' state. For instance, if we can qualify the behavior of cooperative ants, it is because, as an observer, we note a certain number of phenomena we use as indices cooperative activity. It is thus possible to define units of observable indices (resource sharing, parallelization actions, etc.) allowing us to qualify a situation as being co-operative [Durfee *et al.*, 1989] [Bouron, 1992].

The enigma of cooperation is summarized by the following formula [Ferber, 1995]:

Cooperation = collaboration + coordination + resolving conflicts

3.3.2.1 Collaboration

This technique consists of making several agents work on a project by distributing tasks, information and resources so as to work toward a common goal - task sharing technique - [Castelfranchi and Falcone, 1998b]. The solution, which is used for the distributed task, is brought by using a coordinating agent [Tambe, 1997] [Tuomas, 1997]. Thus, this agent will be seen giving out roles to distribute the tasks to the available agents, according to their competence and needs. This type of system uses negotiable approaches based on contract protocol [Davis and Smith, 1987]. It consists of distributing the tasks according to the agents' competence and availabilities. With this intention, the centralized agent uses a mutual representation model of the capacities of each one, called a contract net protocol [Gasser *et al.*, 1987].

3.3.2.2 Coordination

This technique supposes that management of all agents generates a certain number of additional tasks, called coordination tasks, which are not directly productive, but are used so that the productive actions can be achieved under the best conditions [Durfee and Lesser, 1991]. For example, a conveying robot cannot bring back ore if it is not extracted beforehand. A formal model for the implantation coordination system was proposed by Ciancarini [Ciancarini *et al.*, 1999]. The coordination of the various agents' tasks can thank to a coordinating agent (it is the case for the distribution of tasks) or these tasks can be organised directly between two or more agents. The other recent methods were also

suggested, like shared plans [Grosz and Kraus, 1998], joint intentions [Jennings, 1992], delegation and adoption [Haddadi, 1996].

3.3.2.3 Conflict resolution

Life in society, a human society or agents society, implies the appearance of conflicts. For example, conflicts appear when several agents want to use the same resources at the same time or there is a contradiction between the various beliefs, intentions or goals of the agents [Chaudron *et al.*, 2000]. Physical conflit occurring on the resource level are called extrinsic conflict or non-analytical conflict [Castelfranchi, 2000]. The objectives of the agents are not logically contradictory or divergent, but they become incompatible because they do not have sufficient resources enabling agents to be able to achieve their goals. The conflict of knowledge is at the level of knowledge. It occurs when the beliefs, knowledge or the plans of the agents' actions are contradictory or divergent. It can also be called intrinsic conflict or analytical conflict. Suppose that if an agent A believes in a reality p and the agent B denies this same reality, they are in contradiction. There is not the ambiguity in situation.

3.3.2.4 Delegation

The delegation applies mainly with the tasks. An agent wants to delegate one or more tasks to another agent, in a gesture of cooperation [Haddadi, 1996]. For example, an agent asks another agent to carry out a complicated mathematical calculation. This form of delegation is widespread so far. It is also possible to discuss delegation of roles [Werner,

1990]. An agent can have a hierarchical statute in regrouping of agents, which constitutes its social role. Thus, this role can also be delegated, in a temporary or permanent way, with another agent. Finally, it is possible to carry out a delegation of control. When an agent is in control of an unfolding situation, it can delegate this control to another agent for various reasons, without yielding its role to him.

In this field, a lot of work has been done by Castelfranchi and Falcone [Castelfranchi and Falcone, (1997, 1998b, 2000)], which are with the original model of the delegation and the adoption of goals based on the plans [Castelfranchi and Falcone, 1998b]. This model suggests that a delegation is taken into account in mental state of the delegant agent. The given action is always used to reach a certain state, which can be perceived as a goal, the delegation of an action implies the adoption by another agent a pair action/goal noted $\tau = (a, g)$. By the symbol τ, we refers to an action a, which produces a result, a goal g. Ultimately, that means an agent A_1 which delegates an action or which needs the result of an action that carried out by an agent A_2, will include this action in its plan as if it executes itself, such as we can see it on figure 3.4.

The agent A_1 delegates a part of its plan of actions, that means the actions a_3 and a_5 will be realized by the agent A_2. However, it is noted that the agent A_1 keeps the execution of those interior plan. We can also note that the agent A_2 adopts a certain part of the object g of the agent just like state s_6. Then we will be able to say that the agent A_2 adopted a certain part of the plan of A_1's actions.

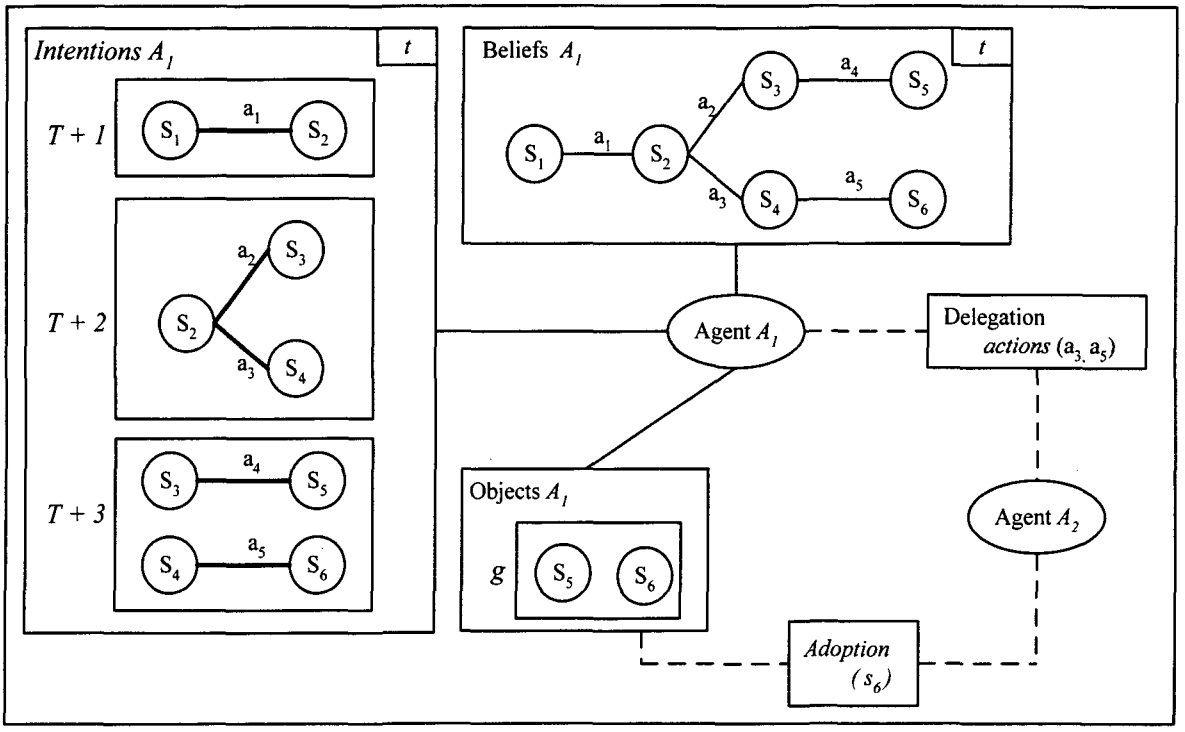


Figure 3.4 : The model of delegation

3.4 Conclusion

Because we are interested in how trust operates in social agents, in this chapter, we introduced the concept of the agent. There are at least two reasons why it is too difficult to define precisely what agents are. Firstly, agent researchers do not "own" this term in the same way as fuzzy logicians/AI researchers, for example, own the term "fuzzy logic" - it is one that is used widely in everyday parlance as in travel agents, estate agents, etc. Secondly, even within the software fraternity, the word "agent" is really an umbrella term for a heterogeneous body of research and development. For this reason, we analyze agent properties and try to define them using the formula with three triplets. Then, in order to

really understand the definition of the agent, we represent the existing Multi-agent system definition, which could conceivably be used in distributed domains to introduce different types of agents.

In our study, we are particularly interested in the interaction among agents. In those approaches to the study of rational agents, the BDI model has been received a great deal of attention. This model is not only the abstract architecture of a family of parallel and distributed systems working alone or in a team in dynamic environments, but also it has a number of applications. It adds a high degree of sophistication and sensitivity to the context when deciding how to react to changed conditions. Depending on design and implementation choices, agents in the BDI model can show very different levels of reactive (event driven) and planned (goal driven) behavior. The description of this behavior is done in cognitive terms, i.e., by attributing mental attitudes. Research is being conducted in a number of areas of relevance both to computer and cognitive sciences, such as cooperative work and social commitments and recognition of intentions [Rao, 1994].

In order to build formally verifiable and practical systems, we introduced a formalization of intentions based on a branching-time possible-worlds model and BDI semantic. We consider them to lay a foundation for our study, in the next chapter, in which we will represent our opinion and hypothesis about how to ensure the trust between the user and its agent acting on the user's behalf.

CHAPTER 4

MUTUAL TRUST MODEL FOR USER IN THE LOOP

4.1 Introduction

The goal of our work is to dig out principles and rules of trust between user and agent acting on his behalf. The cohabitation between the user and the agent differs from the agent-agent cohabitation because the specificity appears in the difficulty of modeling the user, and there is no model providing a kind of user-agent bijection because the proposed theories rest on hypotheses based on reduction [Grislin-Le, 1998] [Tessier *et al.*, 2001]. The user-agent cohabitation is classified according to the balance between the exerted control by the user and that by the agent. In our case, the agent is regarded as a teammate who interacts with the user to achieve jointly a common goal. In other words, the user and the agent have the same possibilities of proposing, refusing, stopping the other, etc. that is, what we call a user in the loop.

Therefore the most difficult question is however not how to share the task, but how to measure trust on the user or the agent, which is particularly significant when the task is critical. Our proposal is to consider mutual trust between the pair user-agent according to the similarity of the viewpoints, or opinions based on beliefs, goals and plans of the task by using the terminological logic [Nebel, 1995] [Baader *et al.*, 2003]. This metric not only makes it possible to make safe task delegation, but also to answer the tiresome task of trying to mediate between two diverging concepts like control and autonomy based on trust for designing human computer system [Castelfranchi and Falcone, 2000].

4.2 Terminological logic

The measure of mutual trust is based on the terminological logic, which seems to suit better the cohabitation context by relocating the problems of measuring the trust into a terminological classification problem [Bouzouane, Bouchard and Shen, 2003]. We can define terminological logic as being a formalism representation of knowledge, which is divided into two quite distinct parts: the terminological part, which makes it possible to define a whole of concepts like their roles, and the assessment part, which allows the individual creation pertaining to the various concepts of the terminological part. The assessment part of formalism allows the operations of classification and inference. The classification, which constitutes the base of the reasoning in terminological logic, supports the relation of subsumption, which allows the concepts' organization and the roles in a hierarchy of terms.

4.3 Mutual trust approach

Our theoretical approach is based on the concept of the viewpoint. A viewpoint is an opinion defined as a terminological structure based on the concepts of beliefs, goals and plans. Figure 4.1 shows us an example of a user and an agent that express their viewpoints in a controlled context of an enterprise's stock. In this example, the user and the agent have exactly the same beliefs, meaning that both of them believe in a lack of product, and also have the same objective, to purchase products. However, the agent actions plan is only partial, because it plans to order products but it does not select the supplier. Its plans consist

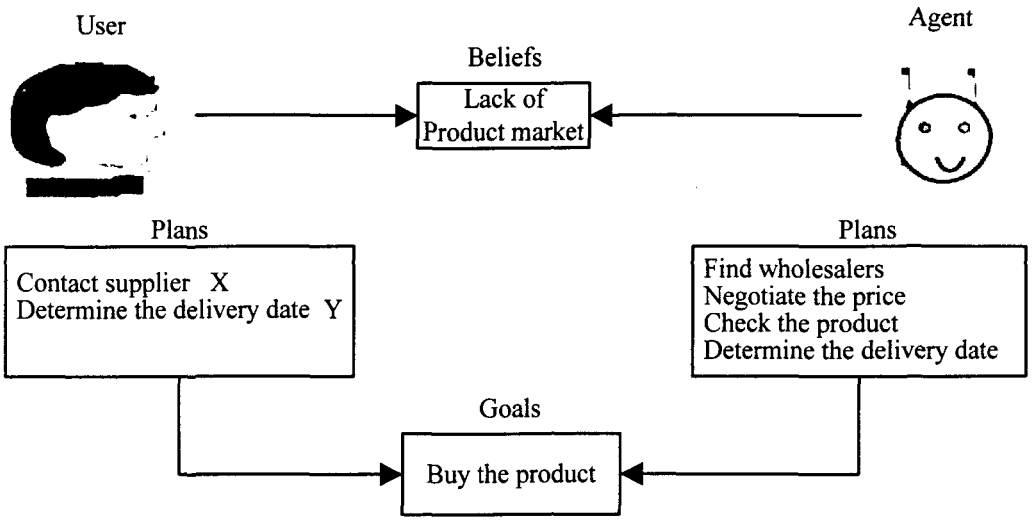


Figure 4.1 : An example of viewpoint

in finding a wholesaler, negotiating the price of the product, checking the quality of the product and finally determining a delivery date. On the other side, the user has a plan of action that is not packed, but complete. It wishes to contact a supplier «X» which has known and fixed with a date «Y». In other words, this example shows us an entity (the user) able to specify its needs without being ready to specify all the details of its actions plan and his/her colleague (the agent), which define his actions plan in detail by taking his information in his database.

These two opinions are similar even though they use different vocabularies and constraints. For example, the actions of «contact supplier» and «find wholesalers» share subsumption relation. A subsumption is an inference method in the terminological logic, to classify an object within taxonomy of objects. Therefore, rather than communicating

directly with the agent or losing the control to the benefit of the user, we first propose to compare their viewpoints by using the terminological logic and then measure the mutual trust based on a classification before delegation of the transaction takes place.

4.3.1 Similarity notion

The concept of similarity generally indicates a relation of resemblance between two or several objects [Larousse, 1980]. It can be defined as the sum of the points or the common characteristics between two objects. Thus, to measure the similarity it can be reduced to determine the number of common characteristics between two objects, comparatively with the total number of their characteristics [Diday *et al.*, 2000]. Suppose that the two states p and q imply one number of the other states, which can represent their characteristics:

$$p \Rightarrow (a \wedge b \wedge c \wedge d)$$

$$q \Rightarrow (a \wedge b \wedge e \wedge f)$$

Moreover, both of them include the two states a and b, which constitute their common characteristics. We could represent the similarity between the states p and q in the following way:

$$SM_{pq} \equiv (a \wedge b)$$

On the other hand, in order to be able to determine if two objects are similar, it is also necessary that they have comparable characteristics or components. It is difficult to

affirm that a television set is similar with a washing machine, because we cannot really compare their respective characteristics. In our case, before measuring the similarity of the viewpoints concepts, we guarantee their semantic equality by checking the existence of a subsumption relation between each concept. Ultimately, we define the similarity as a quantitative measurement of the resemblances, which exist between two comparable objects.

4.3.2 Viewpoint concept

A viewpoint can be defined as being an unspecified representation of mental states of an entity (user or agent) by the given situation. In fact, it means the opinion of the entity about the resolution of a problem. In our case, we will define the viewpoint of an entity by the help of the terminological structure. Thus, we will use the concepts of beliefs, goals and plans of action.

In order to be able to handle, analyze and compare the points of the user's and of agent's view, we propose to use terminological structure for the representation that is shown in figure 4.2.

The concept of supplier subsumes (subsumption relation is represented by the symbol $\preceq$) the concept of the wholesaler, which means that all the characteristics or the roles of the supplier subsume the characteristics of the wholesalers. The concept of the user viewpoint represented by VP^H is defined as a conjunction of the viewpoint and the beliefs

$VP^H \equiv (\cap$ *VIEWPOINT*
(∀ beliefs (∀ missing –product PRODUCT))
(∀ plans (($\exists^{\leq 1}$ *contact-supplier SUPPLIER)*
($\exists^{\leq 1}$ *determine-date-supply DATE))*
(∀ goals (∀ purchasing- product PRODUCT)))
WHOLESALER $\leq$ *SUPPLIER.*
SUPPLIER $\equiv (\cap$ *E-MARKET(∀ selling PRODUCT)).*
$VP^A \equiv (\cap$ *VIEWPOINT*
(∀ beliefs (∀ missing –product PRODUCT))
(∀ plans (∀ find – wholesalers WHOLESALER)
($\exists^{\geq 1}$ *find – wholesalers)*
(∀ check – item CHECKING)
(∀ determine-price PAYMENT)
($\exists^{\geq 1}$ *determine-date-supply DATE))*
(∀ goals (∀ purchasing- product PRODUCT)))

Figure 4.2 : Opinions based on terminological logic (Concept is in capital letter and a role in lower letter)

concerning all entities filling the «missing product» roles. The plan of VP^H consists of at most one specific supplier (represented by $\exists^{\leq 1}$ as a constraint), at most one date of supply and all goals to purchase a product. Notice that the concept of the supplier subsumes the concept of the wholesaler. The supplier is defined as a concept of an e-market of entities with selling roles. On the other hand, the viewpoint of the agent represented by VP^A is defined as a conjunction of the viewpoint and the beliefs concerning all products having «missing – product» roles. The plan of VP^A consists of at least one product wholesaler (represented by $\exists^{\geq 1}$ as a constraint). The found product has a checking and a price role, and at least one date of supply for each item. All goals concern the concepts of having the purchasing product roles.

By using the classification based on subsumption procedure between concepts, we deduce that the role of «contact supplier» with its constraints subsumes «find wholesalers» role. Consequently, the viewpoint of the user subsumes that of the agent.

4.3.3 Trust of the agent

Two similar concepts are not necessarily cloned, but they share a subsumption relation. From there, the user trusts an agent acting on his behalf if the viewpoint of the user subsumes that of the agent. Let us simplify and formalize this:

$$(VP^A \leq VP^H \Rightarrow \textit{User trusts the agent}) \Leftrightarrow$$

$$(\forall C^A \in VP^A \; \exists C^H \in VP^H \mid C^A \leq C^H \wedge SFC(C^H, C^A) \in [0, 1])$$

Where: C^A represents the concept inferred by the agent according to its opinion and C^H represents the concept proposed by the user. The Similarity Factor of Concepts represented by $SFC(C^H, C^A)$ is evaluated when the semantic equality between the two concepts (C^A, C^H) given by the subsumption relation $C^A \leq C^H$. This factor is used to estimate the degree of the credibility of trust, and it is defined like an average of similarity factor of each role making up the concepts. Thus this factor is not a subjective weight fixed arbitrarily but it is evaluated by a generic formula such as the following:

$$SFC\,(C^H, C^A) = \Big[\sum_{j=1}^{Min(C^H, C^A)} SFD\,(role_j^H, role_j^A) \times [(SFN\,(role_j^H, role_j^A) + SFV(role_j^H, role_j^A) / 2]] / Max(C^H, C^A)$$

- *Min* (C^H, C^A): is the minimal number of roles between the compared concepts (C^H, C^A) of each viewpoint.

- *Max* (C^H, C^A): is the maximal number of roles between the compared concepts (C^H, C^A) of each viewpoint.

- *SFD*: indicates the Similarity Factor of the Domain which models the semantic equality between the *j-th* role $«role_j^h»$ from the concept of the user's viewpoint, and its correspondence *j-th* role $«role_j^a»$ from the agent. We denote by D_j^h the domain of $«role_j^h»$ and D_j^a the domain of $«role_j^a»$. This factor *SFD* is equal to $Min(D_j^h, D_j^a) / Max(D_j^h, D_j^a)$, *if* $D_j^h \leq D_j^a$ or $Dj^a \leq D_j^h$ and $j \in [1, Min(D_j^h, D_j^a)]$. In the contrary case the factor is reduced to 0.

- *SFN* indicates the Similarity Factor of the Name of the roles. This factor is equal to 1 if the two roles are identical or synonymous. In the contrary case the factor is reduced to 0. At each case of subsumption between the roles, the synonymous database is updated.

- *SFV* indicates the Similarity Factor of the Values of the atomic roles. This factor is equal to 1 when the two values of roles are atomic and identical. If the role is not atomic, the above formula is used in a recursive way.

4.3.4 Trust of the user

The agent make credible in the user if the viewpoint of the agent subsumes that of the user. This formalize as below:

$$(VP^H \leq VP^A \Rightarrow \textit{Agent trusts the user}) \Leftrightarrow$$

$$(\forall C^H \in VP^H \; \exists C^A \in VP^A \mid C^H \leq C^A \wedge SFC(C^H, C^A) \in [0,1])$$

The similarity factor represented by $SFC(C^H, C^A)$ is evaluated when the semantic equality between the two concepts (C^A, C^H) given by the subsumption relation $C^H \leq C^A$ is guaranteed.

4.3.5 Mutual trust

There is a mutual trust between the user and his agent if only if the semantic equality of their viewpoints is guaranteed. This is formalized as follows:

$$(VP^A \equiv VP^H \Rightarrow \textit{Mutual trust between user and agent}) \Leftrightarrow$$

$$(VP^A \leq VP^H \wedge VP^H \leq VP^A \wedge MTrust(VP^H, VP^A) \in [0,1])$$

The mutual trust is a computational metric related to the similarity of concepts making up each viewpoint of the user and his agent. It is defined by the following formula:

$$MTrust(VP^H, VP^A) = \sum_{j=1}^{Min(VP^H, VP^A)} SFC(C_j^H, C_j^A) \times Utility(VP^H, VP^A) / Max(VP^H, VP^A)$$

- The mutual trust is represented by $MTrust(VP^H, VP^A)$ and belongs to [0,1] interval.

- $Min(VP^H, VP^A)$ is the minimal number of concepts compared between (VP^H, VP^A)

- $Max(VP^H, VP^A)$ is the maximal number of concepts compared between (VP^H, VP^A)

- $Utility(VP^H, VP^A)$ is the utility function that expresses the importance of the mutual trust. This utility value normalized over [0, 1], is measure of how much net benefit if the entity (user or agent) that is taking the trust. It can be estimated arbitrarily or in terms of payoff function [Rosenschein, 1985].

In our case, we use the formula from Marsh's work [Marsh, 1994]: *Utility = (Benefit – Cost) / Benefit,* to realize the *utility*. Here, *benefit* is the possible gains from their cooperation. Suppose that in electronic commercial, the price difference of products between Internet purchasing and other medium purchasing can be as the benefit. In other side, *cost* means the value measured by what must be given or done or undergone to obtain something. The purchasing price of the products, the price of Internet, etc, can be thought as cost.

- If $MTrust(VP^H, VP^A) = 1 \Rightarrow$ These two viewpoints are equivalent and the mutual trust is total.

- If $MTrust(VP^H, VP^A) < 1 \Rightarrow$ These two viewpoints are divergent and the mutual trust is partial.

A mutual trust is not an arbitrary computational metric without a real content, but it is based on the similarity of mental attitudes that support the decision of the common task delegation.

4.3.6 Trust and risk

The delegation by a user or an agent is the subject to the mutual trust according to the level of risk associated with the task. In our case, we use the formula of Marsh [Marsh, 1994] to calculate risk:

$$Risk = \frac{Cost}{Benefit} * Importance$$

Here, *Importance* is an agent-centered or subjective judgment of a situation on the part of the agent concerned. It is a subjective measure of the expected benefits to be gained from a situation under consideration.

If $MTrust(VP^H, VP^A) < Risk$, then the takeover is refused. In this case, the system will seek to integrate the viewpoint until $MTrust(VP^H, VP^A) \geq Risk$ is satisfied. Therefore the measurement of the mutual trust enables us to decide at what stage we can make a safe delegation of tasks to a teammate.

4.3.7 Mutual trust and security

Our study attempts to find a quantitative formula to trust. The above formula shows trust $MTrust(VP^H, VP^A)$ is influenced by utility and the similarity factor of the concepts of the agent's opinions. However, we think it exists other important factors. According to a study by *Business Week* (2000) [Pavlou and Ramnath, 2002], 61% of the survey respondents indicate that they would transact on the Internet if the security and privacy of their personal information could be adequately protected. Therefore, we should consider *security* as one of the factors influencing trust and we extend our formula as follows:

$$MTrust(VP^H, VP^A) =$$

$$\sum_{j=1}^{Min(VP^H, VP^A)} Security \times SFC\,(C_j^H, C_j^A) \times Utility\,(VP^H, VP^A)$$

$$/\,Max\,(VP^H, VP^A)$$

Here, we use the formula from Pavlou's work [Pavlou and Ramnath, 2002] in order to estimate the security factor and it is as follows:

$$Security = 0.19 \times Encryption + 0.29 \times Protection + 0.15 \times Authentication$$
$$+0.11 \times Verification$$

4.3.7.1 Empirical test

From the above formula, we know the factors, which influence trust are security, utility and similarity factors of concept of different viewpoints. Now we want to know if these factors are significant for trust. Therefore we will use regression analysis [Kleinbaum *et al.*, 1998] to measure these factors. The goal of regression analysis is to determine the values of parameters for a function that cause the function to best fit a set of data observations provided. In order to obtain a satisfied result, at first, an empirical study was performed to evaluate the proposed scales and validate the proposed set of inter-relationships related to trust. Thus, participants were asked to assess the degree of trust they would expect from security, utility and SFC with whom they have little experience. The

anchors for all items were "1=strongly disagree" to "3=neither agree nor disagree" and "5=strongly agree".

The final result shows (see the appendix, figure 1): the effect of security *($\beta_1 = 0.36$, prob(t) < 0.01)*, utility *($\beta_2 = 0.29$, prob(t) < 0.01)* and SFC *($\beta_3 = 0.33$, prob(t) < 0.01)* on trust are significant. The model coefficient was extremely significant *($F=331.20$, prob(F)< 0.001)*, and the data explained a substantial degree of the variation *($R^2 = 0.97$)*. Table 4.1 shows the result.

4.4 Validation

The concrete case being used as a validation relates to the electronic commerce. In recent years, many researchers as well as commercial companies have attempted to create intelligent agent based market on the Web. Within the user-agent cohabitation framework, an agent or a user can intervene at any time to take part in the realization of the transaction.

We use Java language and PowerLoom [PowerLoom] as a terminological system for knowledge representation of viewpoints. When a purchasing agent or a selling agent, is started, it creates a view point of the situation using several rules which forms its knowledge base relating to the market and the profile of the other agents. This point of view is transmitted via Internet to its user, as shown by the following figure 4.3. The viewpoints of user and its agent are represented with Java objects. Suppose that in order to take the control of the transaction, we make a decision the two viewpoints are converted in

Table 4.1 : Regression analysis for the factors of trust

Variables	Trust	T-value
Security	0.36	24.19
Utility	0.29	18.92
SFC	0.33	21.99
R^2	0.97	0.97 (adjusted)
F ratio	331.20	
	prob(t) < 0.01	

- *The **T statistic** tests the hypothesis that a population regression coefficient is 0 when the other predictors are in the model.*

- ***Prob(t)** labels the **P values** or the **observed significance levels** for the t statistics. The P values tell us whether a variable has statistically significant predictive capability in the presence of the other variables, that is, whether it adds something to the equation.*

- *The **F Value** is the test statistic used to decide whether the model as a whole has statistically significant predictive capability, that is, whether the regression is big enough, considering the number of variables needed to achieve it.*

- ***R^2** is the squared multiple correlation coefficient. It is the proportion of the variability in the response that is accounted for by the model. If a model has perfect predictability, $R^2=1$. If a model has no predictive capability, $R^2=0$.*

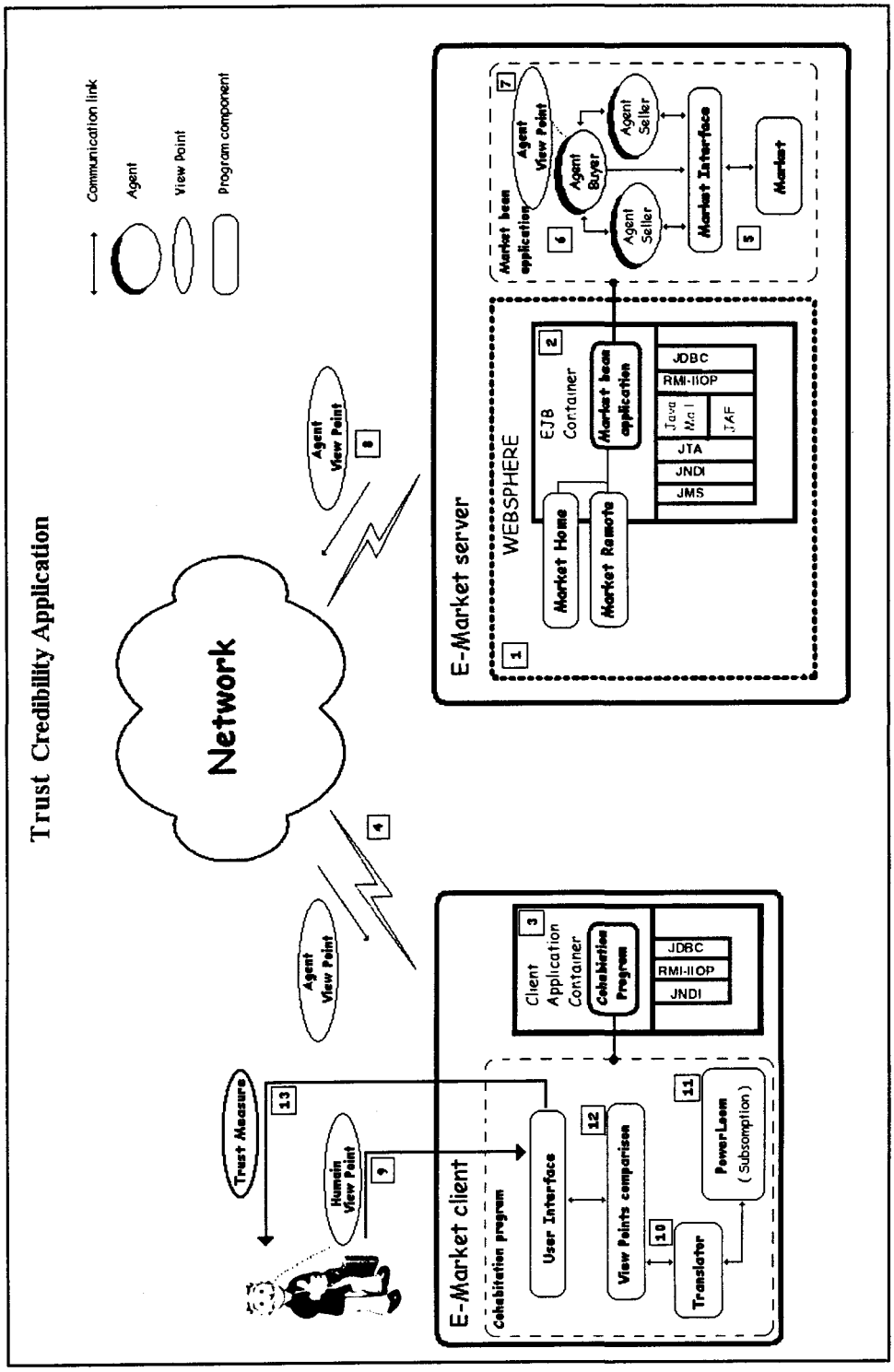


Figure 4.3 : Electronic market's architecture

the PowerLoom terminology by a translator, therefore we use the inference engine of PowerLoom to determine if there is a relation of subsumption between these two points of view. Finally, when PowerLoom deduce a subsumption relation between the two points of view, the evaluation of the mutual trust is engaged by using the previous formula.

4.4.1 Application's architecture

Here, we will use Bouchard's work [Bouchard, 2003] to achieve our goal. Figure 4.3 shows a completed application of electronic market's system. The two squares on the left and on the right of the image correspond respectively to the client and server part, which make the system.

The server part constitutes the electronic market as a whole. It includes the agents, the interface of communication between the agents and the market itself. This part is built in the interior of the component EJB (Undertaken Java Bean) [Deitel and Deitel, 2000] who executes via a server of E-trade IBM WebAphere [WebAphere]. This server of E-trade manages all the communications between the client part and the various components and the safety. For these intentions, this server uses RMI technology [Deitel and Deitel, 2000]. The agents are created directly by this server and they do not move. Therefore, the application must make a request to the server part so that the server creates an agent on the market A certain number of parameters, such as the type of the agent (salesman or purchaser), the type of product compromised, the various scales of price and the information concerning the product, will have to be sent to the market so that the server

creates an agent answering the criteria. Market will return then a number of the agent's identification to the application client, thus enabling the client to contact the new agent directly.

The client part constitutes the tool of the user. This part uses the interface of the application client to connect to the market. It requires visualizing what occurs on the market. We can see the interface of the client application on figure 4.4, also makes it possible to the user to contact his agent, so obtain its point of view on the market situation.

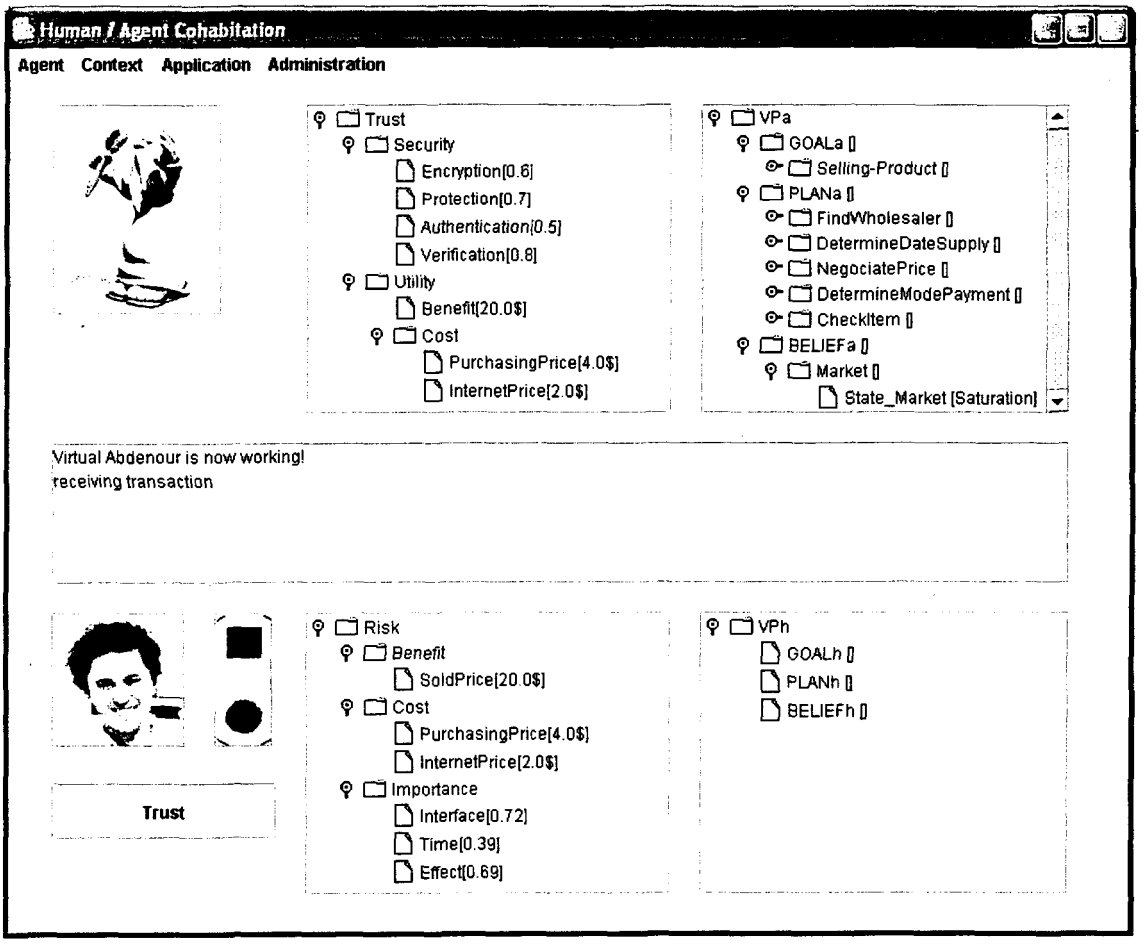


Figure 4.4 : Interface of the client application

Agent's viewpoint will be shown in figure 4.5. It will be able to then compare this one with its own point of view.

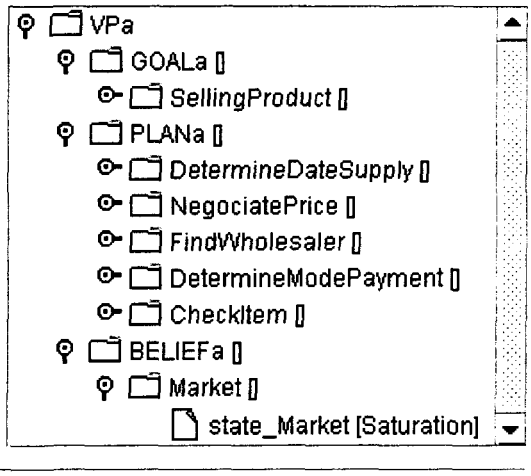


Figure 4.5 : Window of agent's viewpoint

When the user starts of the activities, he will choose his viewpoint as figure 4.6 showing. The viewpoint of the agent and the user are represented respectively in the different windows of the interface. The user can intervene constantly to modify his viewpoint.

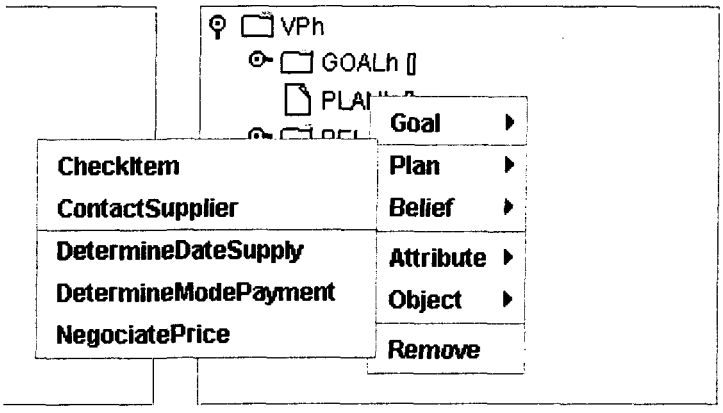


Figure 4.6 : Window of user's viewpoint

4.4.2 Evaluation scenario

Here we give an example to illustrate how to proceed by using our formula in order to estimate the mutual trust. Suppose that an agent (virtual manager) proposes to sell one kind of product in its stock, because it believes that the market of the product tends to saturation. Thus it wants to liquidate its stock, and its plan consists of finding a wholesaler, determining the delivery date, checking items, determining the payment mode and negotiating the price of product. On the contrary, a manager intends to purchase the product with a certain quantity because he believes that the market is in progress. His restocking plan consists in contacting suppliers and determining the delivery date according to potential suppliers. According to the example, in figure 4.7, we will show the similarity factors of the viewpoints between the user and the agent acting on his behalf.

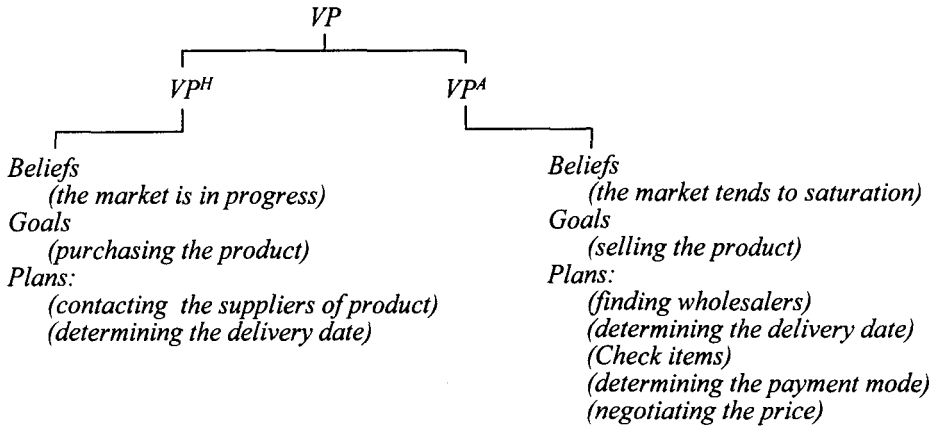


Figure 4.7 : Comparison from viewpoints

Using our formula, a value of viewpoints is represented as follows:

$$SFC\,(VP^H, VP^A) = \Big[\sum_{j=1}^{Min(VP^H, VP^A)} SFD(role_j^H, role_j^A) \times [(SFN(role_j^H, role_j^A) + SFV(role_j^H, role_j^A)] / 2 \Big] / Max(VP^H, VP^A)$$

Here, we have 3 roles: belief, goal and plan, so $Min\,(VP^H, VP^A) = 3$ and $Max\,(VP^H, VP^A) = 3$. For each role, firstly, «Belief», we can calculate it as follows:

$$SFC\,(Belief^H, Belief^A) = \Big[\sum_{j=1}^{Min(Belief^H, Belief^A)} SFD(theMarket^H, theMarket^A) \times [(SFN(theMarket^H, theMarket^A) + SFV(theMarket^H, theMarket^A)] / 2 \Big] / Max(Belief^H, Belief^A)$$

The object $Belief^H$ and $Belief^A$ are the root objects of the beliefs from each viewpoint. $Min\,(Belief^H, Belief^A) = 1$ and $Max\,(Belief^H, Belief^A) = 1$. Each object «Belief» from each viewpoint has only one attribute, which is «theMarket».

- *$SFD(theMarket^H, theMarket^A) = 1$. Because the names of the two attributes refer to the same domain: Market.*
- *$SFN(theMarket^H, theMarket^A) = 1$. Because the two attributes are identical.*

- $SFV(theMarket^H, theMarket^A) = \left[\sum_{j=1}^{Min(theMarket^H,\ theMarket^A)} SFD(state\_Market^H, state\_Market^A) \times [(SFN(state\_Market^H, state\_Market^A) + SFV(state\_Market^H, state\_Market^A)]/2 \right] / Max(theMarket^H, theMarket^A)$

Where:

Min $(theMarket^H, theMarket^A) = 1$ *and Max* $(theMarket^H theMarket^A) = 1$. *It is because the class «theMarket» in the two viewpoints is characterized by one attribute, which is «state_Market».*

- $SFD(state\_Market^H, state\_Market^A) = 1$. *Because the names of the two attributes refer to the same domain: «String».*
- $SFN(state\_Market^H, state\_Market^A) = 1$. *Because the attributes are identical.*
- $SFV(state\_Market^H, state\_Market^A) = SFV(«progress»,«saturation») = 0$.

Thus,

$SFC\ (theMarket^H, theMarket^A) = 1 \times (1 + 0) / 2 = 0.5,$

$SFC\ (Belief^H, Belief^A) = 1 \times (1 + 0.5) / 2 = 0.75.$

The user and the agent do not completely have the same beliefs completely, but they relate to the same object, which is «theMarket». They have similar beliefs 75% of which can be integrated. In other words, they are divergent to 25% because the user believes the state of market is «Progression» and the agent believes that is «Saturation».

According to the same step, the remaining roles can be calculated. Secondly, we calculate «Goal» as follows:

$$SFC\ (Goal^H, Goal^A) = \left[\sum_{j=1}^{Min(Goal^H, Goal^A)} SFD(PurchasingProduct^H, SellingProduct^A) \times [(SFN(PurchasingProduct^H, SellingProduct^A) + SFV(PurchasingProduct^H, SellingProduct^A)]/2 \right] / Max(Goal^H, Goal^A)$$

The object $Goal^H$ and $Goal^A$ are the root objects of the goals from each viewpoint.

$Min\ (Goal^H, Goal^A) = 1$ and $Max\ (Goal^H, Goal^A) = 1$.

- *$SFD(PurchasingProduct^H, SellingProduct^A) = 1$. Because the domains of the two attributes refer to the same domain.*
- *$SFN(PurchasingProduct^H, SellingProduct^A) = 0$.*
- *$SFV(PurchasingProduct^H, SellingProduct^A) = 0$.*

Thus, $SFC(Goal^H, Goal^A) = 1 \times (0+0) / 2 = 0$.

This result means that the goal of the agent is «selling the product», and, on the other hand, the goal of the user is «purchasing the product». These goals are semantically different.

Finally, for «Plan» we can calculate as follows:

$$SFC\ (Plan^H, Plan^A) = \left[\sum_{j=1}^{Min(Plan^H, Plan^A)} SFD(role_j^H, role_j^A) \times [(SFN(role_j^H, role_j^A) + SFV(role_j^H, role_j^A)] / 2 \right] / Max(Plan^H, Plan^A)$$

The object $Plan^H$ and $Plan^A$ are the root objects of the plans from each viewpoint.

$Min\ (Plan^H, Plan^A) = 5$ and $Max\ (Plan^H, Plan^A) = 2$.

- *$SFC(ContectSupplier^H, FindWholesaler^A) = 1 \times (1 + 1) / 2 = 1$, because there is a subsumption relation between the «ContectSupplier» and «FindWholesaler».*
- *$SFC(DetermineDateSupply^H, DetermineDateSupply^A) = 1 \times (1 + 1) / 2 = 1$.*

Thus, $SFC(Plan^H, Plan^A) = 2 / 5 = 0.4$.

The user and the agent do not completely have the same plans completely but the agent's plans subsume the user's plans. They have similar plans, 40% of which can be integrated. In other words, they are divergent to 60%, because the user's plans are partial.

Finally, $SFC(VP^H, VP^A) = (0.75+0+0.4) / 3 = 0.383 < 1$. We can get the conclusion: these two points of view are divergent to 0.617 (1 - $SFC(VP^H, VP^A)$) and they have similar beliefs, 38.3% of which can be integrated. From there, we can say that there is a possibility of cohabitation between the user and his agent because there is possibility of integration of the viewpoints.

Meanwhile, we have to subjectively measure other factors such as: *utility* and *security*. The formula of *utility* given in Section 4.3.5 and the arbitrary fixed values of its factors are: *benefit* and *cost*. Here we suppose the price difference of products between Internet store and physical store can be as the *benefit*, is twenty dollars. The purchasing price of the products, the price of Internet, etc, can be thought as the *cost*, and they are respectively four and two dollars. The formula of *security* given in Section 4.3.7 and the arbitrary fixed values of its factors are: Encryption, Protection, Authentication and Verification. Here we suppose they are respectively 0.6, 0.7, 0.5 and 0.8. Therefore, using our formula, we can calculate the value of mutual trust. Finally, using the result to compare the value of *risk,* we will decide if the delegation will be happened. The formula of *risk* given in Section 4.3.6 and the arbitrary fixed values of its factors are *benefit*, *cost* and *importance*. Here we suppose *importance* is influenced by user friendly Interface, Time and Effect, and these values are respectively 0.72, 0.39 and 0.69. If *risk* >*Mutual trust*, the trust of application will be refused, otherwise, it will be accepted. Therefore, the final result of our case is shown in the following table:

SFC	0.38
Security	0.48
Utility	0.70
MTrust	0.12
Risk	0.18

The implementation of the evaluation scenario is shown in figure 4.8.

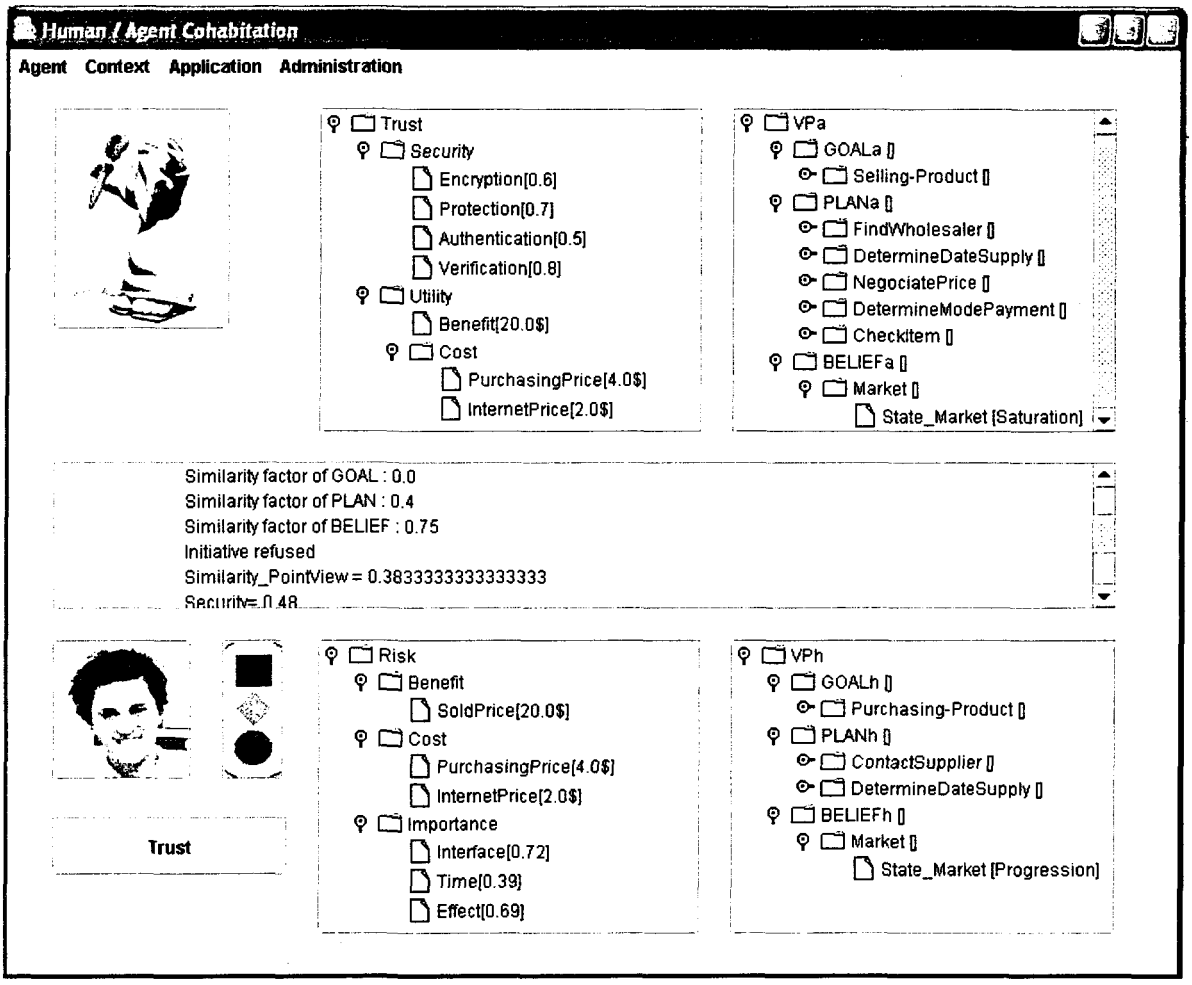


Figure 4.8 : Result of viewpoints' comparison

The viewpoint of the agent is represented in the upper part of the window, and the user in the lower part. The user can intervene constantly to modify his viewpoint. In the

"*trust*" window of the interface, we can know the value of *security* of the agent and *utility* of the trade. In the "*risk*" window, we can know the threshold of *risk* associated to the task. During the development from the calculation, the «yellow» button is lit. If the trust is allowed to continue the realization of the task in progress, the «green» button will be lit. Otherwise, the «red» button is lit, and the trust will be refused to the applicant. The final result of trust can be found in the middle window. For our case, the delegation was refused.

4.5 Conclusion

We described the challenge of mutual trust measurement as a computational aspect between the opinions of the user operator and his agent acting on his behalf in order to ensure a safe delegation of tasks in both sides. This measurement of this mutual trust, not only minimizes the mutual reasoning, but also mediates between two diverging concepts such as autonomy and control [Tambe *et al.*, 2002].

In this chapter, we have supposed and analyzed theoretically our formula. We have also made an empirical study so that we can find out the reliability of the model proposed by the regression analysis. The proposed hypotheses in our work were tested with some participants. Further nonlinear least squares regression analysis provides considerable support for the proposed set of antecedents and consequences of trust. We are able to verify empirically that the mechanisms of security, utility and SFC indeed influence trust, and we find the effect of these factors are significant. According to many literatures [Marsh, 1994]

[Pavlou and Ramnath, 2002], we can calculate these factors so that we can finally quantify our formula.

Through our formula, our opinion is reflected: trust is subjective, but can be quantifiable. It can compare with risk so that the user can make a decision if the delegation will be executed. Meanwhile, risk can be as the degree to judge trust. Most of our ideas come from the literature having been introduced in above parts in this thesis. They enrich our knowledge about trust and give us a solid foundation.

CHAPTER 5
GENERAL CONCLUSION

The objective of this thesis was to contribute to the advance of the agent trust domain cooperation by the achievement of three fixed objectives. Our prime objective consisted of analyzing and discussing "what is trust?" We have tried to give our hypothesis and formula according to some theories and practical investigation in literatures, which allow measuring the credibility of the trust of a user and his agent acting on his behalf. The second objective consisted in proposing a mutual trust model based on the similarity of opinions of user and his agent. This objective is achieved by re-using the study of comparing viewpoints based on terminological logic initiated by studying Bouzouane's work [Bouzouane, Bouchard and Shen, 2003]. The achievement of this goal made it possible to show that our approach guarantees the semantic equality from these two points of view before measuring theirs. The third and last objective consisted of validating our theoretical approach by a case concert in electronic commerce. This objective was achieved through the development of the prototype of sales and purchases of products through e-commerce. The realization of all of these objectives was carried out in three stages.

Initially, we carried out a review of the literature related to trust through various angles. Such as a computational trust [Marsh, 1994], social trust [Castelfranchi and Falcone, 2001], the relationship between trust and security [Pavlou and Ramnath, 2002], a distributed cognition approach of trust [Chandrasekharan, 2002] and finally the relationship between trust and risk [Griffiths and Luck, 1999] [Gefen *et al.*, 2003]. This literature provides a clear definition of trust and presents some mental ingredients relative with trust. They implied trust is the mental background of delegation, so we can say trust is a

subjective definition. Also these literatures gave us some formulae for trust, which are useful in clarifying discussion of concept.

In fact, due to our focus on the trust between the user and his agent, BDI model [Rao and Georgeff, 1992] supplies a basis for our research. That justifies why we introduced it in the chapter 3. It is an approach to the study of the rational agency, which has received a great deal of attention, being the so-called Belief, Desire, Intention (BDI) model. However, current research on BDI model looks at cognitive issues and, for multi-agent systems, is mostly concerned with the notions of shared beliefs, goals and commitments, and how these are built and communicated [Kinny *et al.*, 1994]. It does not directly address any of the mentioned computational problems of trust.

At the second stage, we carried out a research detailed on the formula of mutual trust. We compare a concept of the viewpoint with all concepts of another viewpoint, by using the subsumption operators offered by the engine of inference of PowerLoom [PowerLoom]. Once the semantic equality between two concepts is guaranteed by a relation of subsumption, we use the formula that we defined in order to measure the mutual trust in an objective way. This study was made while passing the access in review a theoretical research, as above, and a practical investigation. In the investigation, we asked some participants to answer our questions to help us to find if those supposed factors are significant for trust.

In the last stage, our work consists of validating the proposal model to realize an application. We gave an example to illustrate how to work using our formula in order to estimate the mutual trust. It is helpful to understand well our formula and our definition about mutual trust.

It should be noted that we do not consider the proposed method for computing the mutual trust as a definite answer to the various problems related to «human in the loop» [Dautenhanhahn, 2001] but as a step in the direction of the generic evaluation method of mutual trust between the user and his agent acting on his behalf. In this research, we do not cover the integration of two subsumed viewpoints. We aim to work out principles and rules based on necessary formula and terminological constructors to define a new common point of view between the user and his agent, in the perspective of integrating the user and the agent in the same loop of common task realization. Also this thesis does not constitute a firm and final response to the problems about trust stated in our Introduction. However the proposed approach must be regarded as a step ahead contributing to the evolution of the field of the trust between the user and his agent.

From the practical point of view, our theoretical approach could be re-used in the other classes of problems, such as the automated production and automatic piloting the accidents, which is the result of a divergence of point of view between the pilot and the autopilot [Kitano, 1996].

Let us mention that an article, presented in the Seventh international conference on International Association of Science and Technology for Development (IASTED'03) [Bouzouane, Bouchard and Shen, 2003].

REFERENCE

[Abdul-Rahman *et al.*, 2000] Abdul-Rahman A. and Hailes S., Supporting Trust in Virtual Communities. In p*roceedings of the 33rd Hawai International Conference on System Sciences, Maui, Hawaii*, 2000.

[Atreya *et al.*, 2002] Atreya M., Hammoun B., Paine S., Starrett P. and Wu S., Digital Signatures, TechOnLine Article, 2002.

[Baader *et al.*, 2003] Baader F., Calvanese D., McGuiness D., Nardi D. and Patel-Schneider P., The Description Logic Handbook: Theorie, Implementation, and applications, Cambridge University Press, United Kingdom, 2003.

[Ball *et al.*, 1997] Ball G., Ling D., Kurlander D., Miller J., Pugh D., kelly T., Stankosky A., Thiel D., Van Dantzich M. and Wax T., Lifelike Computer Characters: The Personal Project at Microsoft, In *Bradshaw, J.M. ed., software Agents*, AAI Press / MIT Press, 1997, pages 191-222.

[Bouchard, 2003] Bouchard, B.: La mesure de la similarité entre les points de vue de l'usager et de son agent artificiel à l'aide de la logique terminologique, mémoire de maîtrise, l'université du Québec à Chicoutimi, 2003.

[Bouron, 1992] Bouron T., Structures de communication et d'organisation pour la cooperation dans un univers multi-agents, thesis of Paris University, 1992.

[Bouzouane, 2002] Bouzouane, A.: Co-habitation between human and intelligent agent. *In: Proc. of Fifth International Conference on E-Commerce Research (ICECR),* ACM/SIGAS-IFIP, Montreal, Canada, Oct. 24-27, 2002, pages 1-5.

[Bouzouane, Bouchard and Shen, 2003] Bouzouane, A., Bouchard, B., Shen F.: Mutual trust model for human in the loop, In: *Proc. of 7th IASTED International Conference ARTIFICIAL INTELLIGENCE AND SOFT COMPUTING, Banff Canada,* July 2003, pages 133-137.

[Bratman, 1987] Bratman M.E., Intentions, Plans and Practical Reason. In *Harvard University Press, Cambridge, MA*, 1987.

[Brenner *et al.*, 1998] Brenner W., Zarnekow R. and Wittig H., Intelligent Software Agents: foundations and applications, Springer, Germany, 1998.

[Brooks, 1990] Brooks R.A., Elephants don't Play Chess, In *Maes. P. editors, Designing Autonomous Agents, MIT Press / Elsevier*, 1990, pages 3-16.

[Busetta and Ramamohanarao, 1998] Busetta P. and Ramamohanarao K., An Architecture for Mobile BDI Agents, the 1998 ACM Symposium on Applied Computing, 1998.

[Castelfranchi, 2000] Castelfranchi C., Conflicts ontology, Computational conflict, in *Computational conflicts – Conflicts Modeling for Distributed Intelligent Systems*, Müller, H.J. and Dieng R. editors, 2000, pages 21-40.

[Castelfranchi and Falcone, 1997] Castelfranchi C. and Falcone R., Delegation Conflicts, in *Lecture Notes in Artificial Intelligence*, Springer-Verlag, 1997, pages 234-254.

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

[Castelfranchi and Falcone, 1998b] Castelfranchi C. and Falcone R., Toward a Theory of Delegation for Agent-Based Systems, in *Robotics and Autonomous Systems (Multi-Agent Rationality)*, 24(3), 1998, pages 141-157.

[Castelfranchi and Falcone, 2000] Castelfranchi C. and Falcone R., Trust and Control: A Dialectic Link, in *Applied Artificial Intelligence Journal*, 14(8), 2000, pages 799-823.

[Castelfranchi and Falcone, 2001] Castelfranchi C. and Falcone R., Social Trust: A Cognitive Approach. *Trust and Deception in Virtual Societies, Kluwer Academic Publishers, Printed in the Netherlands*, 2001, pages 55-90.

[Chaib-Draa *et al.*, 1992] Chaib-Draa B., Moulin B., Mandiau R. and Millot P., Trends in *distributed artificial intelligence, in Artificial Intelligence Review,* 6(1), 1992, pages 35-66.

[Chandrasekharan, 2002] Chandrasekharan S., Trust: A Distributed Cognition Approach, Carleton University Cognitive Science Technical Report 2002-12, 2002.

[Chaudron *et al.*, 2000] Chaudron L., Fiorino H., Maille N. and Tessier C., Difference : a key to enrich knowledge. Concepts and models, Müller H.J. and Dieng R. editor, in *Computational conflicts-Conflict Modeling for distributed Intelligent Systems*, 2000, pages 82-102.

[Ciancarini *et al.*, 1999] Ciancarini P., Omicini A. and Zambonelli F., Coordination Models for Multi-Agent Systems, in *Agentlink News Journal*, 3, 1999, pages 3-6.

[Dautenhahn, 2001] Dautenhahn K., Socially Intelligent agents: The Human in the loop, in *IEEE Trans on systems*, Vol. 31(5), 2001, pages 345-348.

[Davis and Smith, 1987] Davis R., and Smith R.J., Negotiation as a Metaphor for Distributed Problem Solving, in *ALAN H. BOND and LES Gasser editor*, Distributed Artificial Intelligence, 1987, pages 333-356.

[Deitel and Deitel , 2000] Deitel H.M. and Deitel P.J., Comment Programmer en Java, editions Reynald Goulet inc., 2001, pages 1184-1289.

[Demazeau and Müller, 1991] Demazeau Y. and Müller J.P, Decentralized Artificial Intelligence 2, Elsevier North-Holland, 1991.

[Deutsch, 1958] Deutsch M., Trust and Suspicion. In *Journal of Conflict Resolution 2*, 1958, pages 265-279.

[Deutsch, 1962] Deutsch M., Cooperation and trust: Some theoretical notes. In *M.R. Jones, editor, Nebraska Symposium on Motivation. Nebraska University Press*, 1962, pages 275-320.

[Diday *et al.*, 2000] Diday E., Kodratoff Y., Brito P. and Moulet M., Induction symbolique numérique à partir de donnmées, editions Cépaduès, Toulouse, France, 2000, pages 168-196.

[Diffie and Hellman, 1976] Diffie, W. and Hellman, M., New Directions in Cryptography, in *IEEE Transactions on Information Theory*, IT-22(6), 1976, pages 644-654.

[Doran *et al.*, 1997] Doran T.E., Franklin S., Jennings N.R. and Norman T.J., On cooperation in multi-agent systems, in *The Knowledge Engineering Review*, 12(3), 1997, pages 309-314.

[Durfee *et al.*, 1989] Doran T.E., Lesser V.R. and Corkill D.D., Cooperative Distributed Problem Solving, in *The Handbook of Artificial Intelligence*, vol. 4, Barr A., Cohen P.R. and Feigenbaum E.A. editors, 1989, pages 83-148.

[Durfee and Lesser, 1991] Durfee E. and Lesser V.R., Partial Global Planning: A Coordination Framework for Distributed Hypothesis Formation, in *IEEE Transactions on Systems, Man, and Cybernetics*, 21(5), 1991, pages 1167-1183.

[Elfoson, 1998] Elfoson G., Developing trust with intelligent agents: an exploratory study, in: *Proc. of the 1st Intl. Workshops on Trust*, 1998.

[Ferber, 1995] Ferber, J., Les systèmes multi-agents : vers une intelligence collective, éditions InterEditions, Paris, 1995, pages 6-95.

[Finin *et al.*, 1994] Finin T., Fritzson R., McKay and McEntire R., KQML as an Agent Communication Language, in *Proc. 3rd International Conference on Information and Knowledge Mangement (CIKM'94)*, 1994.

[Galliers, 1991] Galliers J. R., Modeling Autonomous Belief Revision in Dialogue, in Decentralized Artificial intelligence 2: Proc. Of the Second European Workshop on Autonomous Agents in a Multi-Agents World (MAAMAW'90), Demazeau Y. and Müller J.P. editors, Elsevier North-Holland, 1991.

[Gambetta, 2000] Gambetta D., Can we trust trust? In D. Gambetta, editor, Trust, Making and Breaking Cooperative Relations, Basil Blackwell, Oxford. 1988, pages 213–237.

[Gasser *et al.*, 1987] Gasser L., Braganze C. and Herman N., Implementing Distributed Artificial Intelligence Systems using MACE, in *IEEE Conference on Artificial Intelligence Applications*, 1987, pages 315-320.

[Gefen *et al.*, 2003] Gefen D., Rao V.S. and Tractinsky N., The Conceptualization of Trust, Risk and Their Relationship in Electronic Commerce: The Need for Clarifications, in *Proc. of the 36th Hawaii International Conference on System Sciences (HICSS'03),* January 06 - 09, 2003.

[Grandison and Sloman, 2000] Grandison T. and Sloman M., A survey of trust in Internet application, *IEEE Communications Surveys, Fourth Quarter*, 2000.

[Griffiths and Luck, 1999] Griffiths N. and Luck M., *Cooperative plan selection through trust.* In *F. J. Garijo and M. Boman, editors, Multi-Agent System Engineering: Proceedings of the Ninth European Workshop on Modelling Autonomous Agents in a Multi-Agent World (MAAMAW'99).* Springer-Verlag, 1999.

[Grislin-Le, 1998] Grislin-Le S., Interaction Homme-SMA : réflexions et problématique de conception. In *Proc. of JFIADSMA '98, ed.. Hermes*, 1998, pages 133-146.

[Grosz and Kraus, 1998] Grosz B.J. and Kraus S., The evolution of SharedPlans, Wooldridge M. and Rao A. editors, in *Foundations and Theories about Rational Agency*, 1998, pages 1-31.

[Haddadi, 1996] Haddadi A., Communication and Cooperation in Agent Systems, Springer-Verlag, 1996.

[Halpern and Moses, 1985] Halpern J.Y. and Moses Y.O., A guide to the modal logics of knowledge and belief. In *Proceedings of the Ninth International Joint Conference on Articial Intelligence (IJCAI-85)*, Los Angeles, CA, 1985, pages 480-490.

[He *et al.*, 1998] He Q., Sycara K.P. and Finin T.W., Personal Security Agent: KQML-Based PKI, in *Proceedings of the 2nd International Conference on Autonomous Agents (Agents'98)*, 1998, pages 377-384.

[Hertzberg, 1988] Hertzberg L., On the Attitude of Trust, Inquiry, 31(3), 1988, pages 307–322.

[Hutchins, 1995] Hutchins E., How a cockpit remembers its speeds, Cognitive Science, vol. 19, 1995, page 265-288.

[Ingersoll, 2000] Ingersoll, Jonathan E., Jr., Digital Contracts: Simple Tools for Pricing Complex Derivatives, *Journal of business*, University of Chicago Press, 2000, pages 67-88.

[Jarvenpaa *et al.*, 1999] Jarvenpaa, S.L. and Tractinsky, Consumer trust in an Internet store: A Cross-cultural validation, Journal of Computer Mediated Communication, Vol. 5, 1999.

[Jarvenpaa *et al.*, 2000] Jarvenpaa, S.L., Tractinsky, N. and Vitale M., Consumer Trust in an Internet Store, Information Technology and Management, Vol. 1, Issue 12, pages 45-71.

[Jennings, 1992] Jennings, N.R., Joint Intentions as a Model of Multi-Agent Cooperation in Complex Dynamic Environment, doctor thesis, London University, 1992, pages 1-179.

[Jennings, 1993] Jennings, N.R., Commitment and conventions: the foundation of coordination in multi-agent system, in *The Knowledge engineering Review*, 8(3), 1993, pages 223-250.

[Jonker and Treur, 1999] Jonker, C. M. and Treur, J., *Formal analysis of models for the dynamics of trust based on experiences.* In *Garijo, F. J. and Boman, M., editors, Proceedings of the 9th*

European Workshop on Modeling Autonomous Agents in a Multi-Agent World: Multi-Agent System Engineering (MAAMAW-99), volume 1647, Berlin, 1999, pages 221- 231.

[Kerstin, 2000] Kerstin D., Socially Intelligent Agents - The Human in the Loop, *Special Issue of IEEE Transactions on Systems, Man, and Cybernetics,Part A: Systems and Humans,* University of Hertfordshire, UK, 2000.

[Khare and Rifkin, 1997] Khare, R. and Rifkin A. Weaving a Web of Trust, World Wide Web, 2(3), 1997, pages 77-112.

[Kinny *et al.*, 1994] Kinny, D., Ljungberg M., Rao A., Sonenberg E., G. Tidhar and Werner E. Planned team activity. In C. Castelfranchi and E. Werner, editors, Artificial Systems, volume 830 of Lecture Notes in Computer Science, Springer Verlag, 1994, pages 227-256.

[Kitano, 1996] Kitano, H.: New Directions, Nausicaä and the Sirens: A Tale of Two Intelligent Autonomous Agents. In *IEEE Computer Society*, Vol. 11(6), 1996.

[Kleinbaum *et al.*, 1998] Kleinbaum D.G., Kupper L.L., Muller K.E. and Nizam A.D., Duxbury, Applied Regression Analysis and Multivariable Methods, Press An Imprint of Brooks/Cole Publishing Company, 1998.

[Lagenspetz, 1992] Lagenspetz O., Legitimacy and Trust, Philosophical Investigations, 15(1), January 1992, pages 1-21.

[Lamsal, 2001] Lamsal P., Understanding Trust and Security, in *IEEE Internet Computing*, 2001.

[Larousse, 1980] Larousse, Dictionnaire encyclopédique pour tous : le petit Larousse illustré, librairie Larousse, Paris, 1980, page 948.

[Lenstra and Verheul, 1999] Lenstra A.K. and Verheul E.R., Selecting Cryptographic Key Sizes, in *Journal of Cryptology: the journal of the International Association for Cryptologic Research*, 1999.

[Luhmann, 1979] Luhmann, N., Trust and Power, Cichester: John Wiley, New York, 1979.

[Luhmann, 1990] Luhmann, N., Familiarity, Confidence, Trust: Problems and Alternatives, In *D. Gambetta (ed.) Trust: Making and breaking cooperative Relations, Basil Blackwell, N.Y.*, 1990.

[Maes, 1997] Maes P., Agents that Reduce Work and Information Overload, In *Bredshaw, J.M. ed., software Agents. AAAI Press / M IT Press*, 1997, pages 146 -164.

[Malone *et al.*, 1997] Malone T.W., Lai K.Y. and Grant K.R., Agents for Information Sharing and Coordination: a History and Societal Reflections, In *Bradshaw. J.M. ed., Software Agents, AAAI Press I MIT Press*, 1997, pages 109-143.

[Marsh, 1992] Marsh S., Trust and Reliance in Multi-Agent Systems: a Preliminary Report, in *Proceedings of the 4 European Workshop on Modeling Autonomous Agents in a Multi-Agent World (MAAMAW'92),* Rome, 1992.

[Marsh, 1994] Marsh S., Formalising Trust as a Computational Concept, PhD thesis, Department of Computing Science and Mathematics, University of Stirling, April 1994.

[Mayer *et al.*, 1995] Mayer, R.C., Davis, J.H. and Schoorman, F.D., An Integrative Model of Organizational Trust, Academy of Management Review, Vol. 20, No. 3, 1995, pages 709-734.

[McKnight *et al.*, 1998] McKnight D.H., Cummings L.L. and Chervany N.L., Initial Trust Formation in New Organization Relationships, Academy of Management Review Vol. 23, No. 3, 1998, pages 473-490.

[Miceli *et al.*, 1995] Miceli M., Cesta A. and Rizzo P., Distributed Artificial Intelligence from a Socio-Cognitive Standpoint: Looking at Reasons for Interaction. In *Artificial Intelligence and Society*, 1995, pages 287-320.

[Minsky, 1985] Minsky M., The Society of Mind, Touchtone Book. Simon and Schuster, 1985.

[Mitchell *et al.*, 1998] Mitchell J. C., Shmatikov V. and Stern U., Finite-State Analysis of SSL 3.0, In *Seventh USENIX Security Symposium*, 1998, pages 201-216.

[Nebel, 1995] Nebel B., Reasoning and Revision in Hybrid Representation Systems, in *Lecture Notes in Artificial Intelligence*, editions Springer-Verlag, vol. 422, 1995.

[Negroponte, 1997] Negroponte N., Agents: From Direct Manipulation to Delegation, In *Bradshaw. J.M. ed., Software Agents. AAAI Press I MIT Press*, Menlo Park, Calif., 1997, pages 57-66.

[Nelson and Cooprider, 1996] Nelson K.M. and Cooprider J.G., The Contribution of Shared Knowledge to IS Group Performance. MIS Quarterly, 20(4), 1996, pages 409-434.

[Pavlou and Ramnath, 2002] Pavlou P.A. and Ramnath K., Perceived information security, financial liability and consumer trust in electronic commerce transactions. *In journal of Logistics Information Management, Special Issue on Information Security, 15, 5/6*, 2002, pages 358-368.

[PowerLoom] PowerLoom. Knowledge Representation System, ISI, University of Southern California, http://www.isi.edu/isd/LOOM/PowerLoom.

[Rao, 1994] Rao A.S., Means-end plan recognition: Towards a theory of reactive recognition. In *Proceedings of the Fourth International Conference on Principles of Knowledge Representation and Reasoning (KR-94),* Bonn, Germany, 1994.

[Rao and Georgeff, 1991] Rao A.S. and Georgeff M.P., Modeling rational agents within a BDI-architecture. In *R. Fikes and E. Sandewall, editors, Proceedings of the Second International Conference on Principles of Knowledge Representation and Reasoning (KR&R-91)*, Morgan Kaufmann Publishers: San Mateo, CA, April 1991, pages 473–484.

[Rao and Georgeff, 1992] Rao A.S. and Georgeff M.P., An abstract architecture for rational agents. In *W. Swartout C. Rich and B. Nebel, editors, Proceedings of the Third International Conference on Principles of Knowledge Representation and Reasoning, San Mateo*, CA, 1992, pages 439-449.

[Rao and Georgeff, 1995] Rao A.S. and Georgeff, M.P., BDI Agents: From Theory to Practice, In *Proc. of the 1st International Conference on Multi-Agent Systems (ICMAS-95)*, San Francisco, USA, 1995, pages 312-319.

[Rosenschein, 1985] Rosenschein. Rational Interaction: Cooperation among agent. PhD thesis, Stanford University, 1985.

[Schillo *et al.*, 2000] Schillo M., Funk P. and Rovatsos M., Using trust for detecting deceitful agents artificial societies, in *Applied Artificial Intelligence Journal, Special Issue edited by C. Castelfranci, Y.Tan, R. Fancone, and B. Firozabadi, on Deception, Fraud and Trust in Agent Societies,* 2000, pages 825-848.

[Shen *et al.*, 2001] Shen W., Douglas H.N. and Jean-Paul A.B., Multi-agent systems for concurrent intelligent design and manufacturing, published by Taylor and Francis Inc, London, 2001.

[Solomon, 1960] Solomon, L., The Influence of Some Types of Power Relationships and Game Strategies Upon the Development of Interpersonal Trust. In *journal of Abnormal and Social Psychology,* Vol. 61, No. 2, 1960, pages 223-230.

[Spaan, 2002] Spaan M.T.J., Team play among soccer robots, master thesis, Amsterdam University, 2002.

[Stoneburner, 2001] Stoneburner, G., Underlying Technical Models for Information Technology Security, Recommendations of the National Institute of Standards and Technology, NIST Special Publication 800-33, December 2001.

[Tambe, 1997] Tambe M., Towards flexible teamworks, in *Journal of Artificial Intelligence Research (JAIR),* vol. 7, 1997, pages 83-124.

[Tambe *et al.*, 1997] Tambe M., Adibi J., Al-Onaizan Y., Kaminka A., Marsella S., Muslea I. and Tallis M., ISIS: Using an Explicit Model of Teamwork in Robocup'97, in *Proc. of the Robocup'97simulation league tournament,* Information Sciences Institute, University of Southern California, 1997, pages 1-7.

[Tambe *et al.*, 2002] Tambe, M. Scerri P., Pynadath, D.V., Adujstable Autonomy: From Theory to implementation, AAAI'02 Workshop, July 28-August 1, Alberta, Canada, AAAI Press, 2002.

[Tessier *et al.*, 2001] Tessier, C., Müller H. J., Fiorino, H., Chaudron, L. : Agents' conflicts : new issues. In *Tessier. C. Müller, H-J. (eds.): Conflicting agents. Kluwer Academic Pub, Boston, Dordreche. London*, 2001, pages 1-28.

[Tuomas, 1997] Tuomas A., On the Structure of Delegation Networks, in *Proc. Of The 11th Computer Security Foundations Workshop*, 1997, pages 1-13.

[WebSphere] WebSphere. IBM WebSphere Technology for E-Commerce. http://www.ibm.com/us/

[Werner, 1990] Werner E., Cooperating ahents: a unified theory of communication and social structure, in *Distributed Artificial Intelligence: II,* Cohen P., Morgan J. and Pollack M. editors, 1990.

[Wooldridge and Jennings, 1995] Wooldridge, M. and Jennings, N., Intelligent Agents: Theory and Practice, in *The Knowledge Engineering Review* 10 (2), 1995, pages 115-152.

[Wooldridge, 1996] Wooldridge, M., A Logic of BDI Agents With Procedural Knowledge, In J. L. Fiadeiro and P.Y. Schobbens, editors, *Proceedings of the Second Workshop of the MODELAGE Project. Sesimbra, Portugal, January 15-17th,* 1996.

[Wooldridge, 2000] Wooldridge M., Reasoning about rational agents, *MIT Press*, Cambridge, Massachusetts, London, England, 2000.

APPENDIX 1

THE RESULT OF INVESTIGATION

```
   1: Title "example forTrust ";
   2:
   3: Variable Security;
   4: Variable Utility;
   5: Variable SFC;
   6: Variable  Trust;
   7:
   8: Parameter B0;
   9: Parameter B1;
  10: Parameter B2;
  11: Parameter B3;
  12:
  13: Function Trust = B0 + B1* Security + B2* Utility + B3* SFC ;
  14: Data;

Beginning computation...
Stopped due to: Both parameter and relative function convergence.


   ----  Final Results  ----

NLREG version 6.1
Copyright (c) 1992-2004 Phillip H. Sherrod.

example forTrust
Number of observations = 30
Maximum allowed number of iterations = 500
Convergence tolerance factor = 1.000000E-010
Stopped due to: Both parameter and relative function convergence.
Number of iterations performed = 3
Final sum of squared deviations = 2.1789566E-002
Final sum of deviations = 5.5511151E-017
Standard error of estimate = 0.0289493
Average deviation = 0.0226907
Maximum deviation for any observation = 0.0633329
Proportion of variance explained (R^2) = 0.9745  (97.45%)
Adjusted coefficient of multiple determination (Ra^2) = 0.9716  (97.16%)
Durbin-Watson test for autocorrelation = 1.476
Analysis completed 14-Jan-2004 17:18.  Runtime = 0.05 seconds.
```

---- Descriptive Statistics for Variables ----

Variable	Minimum value	Maximum value	Mean value	Standard dev.
Security	0	1	0.5233333	0.3757139
Utility	0	1	0.5566667	0.3578512
SFC	0	1	0.4533333	0.3766763
Trust	0	1	0.5103333	0.1716549

---- Calculated Parameter Values ----

Parameter	Initial guess	Final estimate	Standard error	t	Prob(t)
B0	1	0.00739905428	0.01701085	0.43	0.66718
B1	1	0.360754244	0.01491643	24.19	0.00001
B2	1	0.294660925	0.01557039	18.92	0.00001
B3	1	0.331128624	0.01505602	21.99	0.00001

---- Analysis of Variance ----

Source	DF	Sum of Squares	Mean Square	F value	Prob(F)
Regression	3	0.8327071	0.277569	331.20	0.00001
Error	26	0.02178957	0.0008380602		
Total	29	0.8544967			

Figure 1 : The result of investigation