MONOGRAFIES DE L'INSTITUT D´INVESTIGACIÓ
EN INTEL·LIGÈNCIA ARTIFICIAL

Milking the Reputation Cow: Argumentation, Reasoning and Cognitive Agents

Isaac Pinyol Catadau

Consell Superior d´Investigacions Científiques

MONOGRAFIES DE L'INSTITUT D'INVESTIGACIÓ
EN INTEL·LIGÈNCIA ARTIFICIAL
Number 44

IIIA
Institut d'Investigació en
Intel·ligència Artificial

GOBIERNO
DE ESPAÑA

MINISTERIO
DE CIENCIA
E INNOVACIÓN

CSIC

Milking the Reputation Cow: Argumentation, Reasoning and Cognitive Agents

Isaac Pinyol Catadau

Foreword by Jordi Sabater Mir

2011 Consell Superior d'Investigacions Científiques
Institut d'Investigació en Intel·ligència Artificial
Bellaterra, Catalonia, Spain.

Series Editor
Institut d'Investigació en Intel·ligència Artificial
Consell Superior d'Investigacions Científiques

Foreword by
Jordi Sabater Mir
Institut d'Investigació en Intel·ligència Artificial
Consell Superior d'Investigacions Científiques

Volume Author
Isaac Pinyol Catadau
Institut d'Investigació en Intel·ligència Artificial
Consell Superior d'Investigacions Científiques

ISBN: 978-84-00-09352-5
ISBN (online.): 978-84-00-09353-2
NIPO: 472-11-150-3
NIPO (online.): 472-11-149-0
D.L.: B.33336-2011

Ordering Information: Text orders should be addressed to the Library of the IIIA, Institut d'Investigació en Intel·ligència Artificial, Campus de la Universitat Autònoma de Barcelona, 08193 Bellaterra, Barcelona, Spain.

Foreword

If you look at the literature on computational reputation systems you will notice that, in spite of being a human inspired social control mechanism, very little attention has been given to the work already existent coming from social sciences. It seems like if researchers from computer science were ignoring the amount of work that for many years social scientists have been accumulating regarding this topic. Fortunately, the last few years this has been changing, albeit slowly. The work you have in your hands is one of the best exponents of this new trend in the computational trust and reputation community.

In the following chapters, and taking as foundation a solid cognitive theory of reputation, you will go from the logic formalization of this theory and its integration into a cognitive agent architecture to its use in the context of argumentation dialogs.

To the contrary of what happens in other similar attempts where the original theory is diluted an almost unrecognizable in the implemented system, here it maintains its essence and identity from the beginning to the end. Dr. Isaac Pinyol shows how, in an area that traditionally has been dominated by game theory, the use of a cognitive approach opens a world of new possibilities.

I'm sure that the work you are about to read will become one of the references in the area in a near future.

Bellaterra, June 2011

Jordi Sabater Mir
IIIA - CSIC

Als meus pares, al meu germà Jordi,
a la meva dona Mao-Mei, al nostre fill Noah
i... als seus germans/es?

Acknowledgments

Aquest treball ha estat possible gràcies a la paciència i col·laboració de moltes persones i institucions que amb dedicació i sense remugar massa han posat el seu gra de sorra tan en els moments àlgids d'eufòria, com en els més esquerps de superar. Sense cap mena de dubte, el suport de la meva família durant tota la meva trajectòria personal, acadèmica i professional ha estat un dels pilars més important que m'han fet arribar fins aquí. Sempre els hi estaré agraït. Tampoc hagués pogut finalitzar aquesta monografia sense la paciència i suport de la meva estimada muller Mao-Mei, qui sempre ha estat al meu costat i amb qui sempre he confiat. No puc deixar de donar les gràcies tampoc el nostre fill Noah, nascut el 10 de març d'aquest mateix any 2010. El seu somriure, la seva energia i el seu entusiasme per tot el que es fa i es desfà han estat un valuós combustible en l'última etapa.

Estaré eternament agraït al doctor Jordi Sabater i Mir per haver-me donat la oportunitat de començar aquesta aventura i comprometre's a supervisar la meva recerca. Han estat moltes reunions i molta feina, però també moltes cerveses i molts bons records que mai podré oblidar. L'esforç i dedicació per revisar tot el meu treball així com els seus consells han estat la guia sense la qual no hagués pogut arribar a bon port.

M'és imprescindible agrair a totes les persones de l'Institut d'Investigació en Intel·ligència Artificial, qui des del començant em van acollir, i amb qui tantes bones estones he passat. Sense ells, sense aquest centre on es prima l'excel·lència investigadora i la noblesa personal, tot el procés hagués estat molt més arduós. En particular, no em puc oblidar dels doctorand i amics amb qui he compartit la major part del temps al IIIA i que de ben segur es convertiran en grans científics. Sr. Villatoro, Sra. Vinyals, Sra. Fàbregues, Sra. Delgado i Sr.(Dr.) Nin us tindré sempre present. De la mateixa manera, no puc oblidar els començaments a la UPC on l'Íñigo, l'Atif i en Manel es van convertir en el suport que necessitava.

I am also grateful for the European project eRep (CIT5-028575) and all the partners of the consortium: University of Bayreuth (UBT), University of Groningen (RuG) and the Institute of Cognitive Science and Technology (CNR-ISCT): great work, great people and great meals. In particular, I would like to thank all the people in the Laboratory for Agent-based Social Simulation (LABSS) of the CNR-ISCT for their warm hospitality during my three-month stay in Rome: Mario Paolucci for giving me the opportunity to participate in the LABSS activities, and Rosaria Conte for letting me share her office. I am very thankful to both for introducing me to the world of agent-based social simulations and for their contribution to the preliminary experiments described in chapter 2.

M'agradaria agrair també a la Pilar Dellunde (UAB) la paciència, dedicació i bons consells en el desenvolupament del capítol 4, i a Pablo Noriega (IIIA-CSIC) per a les converses de cafè que em van permetre desencallar el desenvolupament inicial del capítol 6.

No quiero dejar de agradecer a Ramon Hermoso y a Roberto Centeno de la Universidad Rey Juan Carlos (URJC), y a Viviane Torres da Silva de la Universidade Federal Fluminente de Brasil (UFF) por sus comentarios en el desarrollo del modelo BDI+Repage+Norm descrito en la sección 5.4 del capítulo 5.

Finalment vull agrair a totes les persones que de ben segur haurien d'estar mencionades però que per espai i descuit personal no he pogut incloure.

A tots, moltíssimes gràcies.

Isaac Pinyol,
Vilafranca del Penedès
28 d'Agost de 2010

Abstract

Computational trust and reputation models have been recognized as one of the key technologies required to design and implement agent systems. These models manage and aggregate the information needed by agents to efficiently perform partner selection in uncertain situations. For simple applications, a game theoretical approach similar to that used in most models can suffice. However, if we want to undertake problems found in socially complex virtual societies, we need more sophisticated trust and reputation systems, that not only focus on the construction and inference of social evaluations (epistemic decisions), but on their role in the practical reasoning performed by the agents (pragmatic-strategic decisions) and on communications and dialectical processes (memetic decisions). Most of the current state-of-the-art models struggle with epistemic decisions, on how agents evaluate other agents according to certain criteria. Curiously, pragmatic-strategic and memetic decisions are traditionally left apart, either because they are implicit in the model or because it is too dependent on the domain. This work explores this gap, arguing that in complex scenarios where more cognitive approaches are needed, both pragmatic-strategic and memetic decisions are as important as epistemic ones.

Firstly, we construct an ontology of reputation and a reputation language that captures the information that most of the current state-of-the-art computational trust and reputation models manage. This starting point serves, by the one hand, to define a *belief-desire-intention* (BDI) agent architecture that integrates reputation information. Then, desires and intentions interact with beliefs that contain information coming from reputation models. The architecture is flexible enough to model a wide number of agents' families and precisely determines the practical reasoning process that leads to the best *reasonable* action. On the other hand, we exploit the reputation language and use it to define an argumentation-based protocol that allows two parties to engage in dialog processes and exchange reputation-related information. The protocol permits agents justify their social evaluations, endowing them with the capability to intendedly decide whether communicated social evaluations are reliable according to their own knowledge.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

1.1 Motivation

The importance of reputation and trust is out of question in both human and virtual societies. The sociologist Luhmann wrote [Luhmann, 1979]: *"Trust and trustworthiness are necessary in our everyday life. It is part of the glue that holds our society together"*. Luhmann's observation was also contrasted in virtual societies. The proliferation of electronic commerce sites started the need for mechanisms that ensure and enforce *normative* behaviors and at the same time, increase electronic transactions by promoting potential users' *trust* towards the system and the business agencies (agents) that operate in the site.

Along with it, reputation arises as a key component of trust, becoming an implicit social control artifact [Conte and Paolucci, 2002]. Humans rely on reputation information to *choose* partners to cooperate with, to trade, to form coalitions etc. and it has been studied from different perspectives, such as psychology (Bromley [Bromley, 1993], Karlins *et al.* [Karlins and Abelson, 1970]), sociology (Buskens [Buskens, 1998]), philosophy (Plato [Plato, 1955], Hume [Hume, 1975]) and economy (Marimon *et al.* [Marimon et al., 2000], Celentani *et al.* [Celentani et al., 1966]). Every society has its own rules and norms that members should follow to achieve a *well-fare* society. The social control that reputation generates emerges implicitly in the society, since non-normative behaviors will tend to generate bad reputation that agents will take into account when selecting their partners, and therefore it can cause exclusion due to social rejection.

One of fields that most is using these concepts is the field of multi-agent systems (MAS). These systems are traditionally composed of discrete unites called agents that are autonomous and that need to interact to each other to achieve their goals. The parallelism with human societies is obvious, and also the problems, specially when we are talking about *open* MAS. The main feature that characterizes open multi-agent systems is that the intentions of the agents are unknown. Hence, due to the uncertainty of their potential behavior we need mechanisms to control the interactions among the agents, and protect *good* agents from fraudulent entities. Traditionally, three approaches have been followed to solve such problems:

- **Security Approach:** At this level, basic structural properties are guaranteed, like authenticity and integrity of messages, privacy, agents' identities, etc. They can be secured by means of cryptography, digital signatures, electronic certificates etc. However, this approach does not tell anything about the quality of the information, although the established control is more than valuable.

- **Institutional Approach:** This approach assumes a central authority that observes, controls or enforces agents' actions, and might punish them in case of *non-desirable* behaviors. It is indisputable that this approach ensures a high control in the interactions, but it requires a centralized hub. Moreover, the control is bounded to structural aspects of the interactions: allowed, forbidden, obliged actions can be checked and controlled. However, the quality of the interactions is left apart, in part, because a *good* or *bad* interaction has a subjective connotation that can depend on the current goals of each individual agent.

- **Social Approach:** Reputation and trust mechanisms are placed at this level. In this approach agents themselves are capable of punishing non-desirable behaviors, y for instance, not selecting certain partners. To achieve such distributed control agents must model other agents' behaviors, and following the similitude with human societies, trust and reputation mechanism arise as a good solution. This requires however the development of computational models of trust and reputation, which must cover not only the generation of social evaluations in all the dimensions, but on dealing with how agents use reputation information to select partners, how agents communicate and spread reputation, and how agents handle communicated reputation information, etc. It is important to remark that these approaches are complementary and that each one covers a different typology of problems, all related to the control of interactions in open MAS.

This work is framed in the field of computational reputation and trust models for open MAS. In the recent years, the scientific research in this field has considerably increased, and in fact, reputation and trust mechanisms have been already considered a key elements in the design of MAS [Luck et al., 2005]. Nowadays, most of the computational models use game theoretical approaches that suffice for simple environments. However, if we want to undertake problems found in socially complex virtual societies, more sophisticated trust and reputation systems based on solid cognitive theories are needed.

Taking the cognitive theory of reputation developed by Conte and Paolucci [Conte and Paolucci, 2002] as a base, we deal with problems that traditionally have been left apart when facing such complex systems. On the one hand, we deal with pragmatic-strategic decisions by defining an agent architecture capable of integrating reputation information into its deliberative process. On the other hand, we face memetic decisions by specifying a family of argumentation-based dialog protocols that allows the agents to analyze the internal elements used to infer reputation-related concepts, and exchange them with other agents. In the next section we detail the scientific contributions.

1.2 Main Contributions

This work contributes to the field of computational trust and reputation for multiagent systems in three lines:

First - An Ontology of Reputation and the L_{rep} Language

We present an ontology of reputation and the language L_{rep} to capture the reputation information that computational trust and reputation models manage in terms of social evaluations (evaluations about the social performance of an entity in a specific context). It serves to precisely determine the elements that compose a social evaluation and at the same time, provides a clear conceptualization of the involving terms. The main features are:

- The ontology considers computational aspects, such like the representation type used to evaluate other agents performances. For instance, some models use a set of linguistics labels like *Very Bad, Bad, Neutral, Good, Very Good*, while others use probabilistic distributions. We propose four types of representations that capture most of the representations used in the current state-of-the-art models, and define transformation functions to move from one type to another.

- The ontology introduces a taxonomy of social evaluations extracted mainly from the cognitive theory of reputation by Conte and Paolucci [Conte and Paolucci, 2002]. Even when the specific terms may not have a direct connection with the terminology used by other reputation models, the information that most of the current models manage fits into the terms of the ontology.

- The ontology serves as a base to define L_{rep}, a many-sorted first-order language that we use to characterize the reputation information that agents hold. We assume that agents use L_{rep} to write and reason about reputation concepts and associate an inference relation $\vdash_i$ that represents a particular reputation model. With it, we can formalize the fact that even when agents use the same language to express reputation concepts, agents can infer them in multiple and different ways.

Second - The BDI+Repage Model

We introduce the BDI+Repage [Pinyol and Sabater-Mir, 2009a] agent architecture, a *belief-desire-intention* (BDI) architecture that integrates the information that the computational reputation system Repage [Sabater-Mir et al., 2006] provides into the practical reasoning process of the agent. Differently from most of the current state-of-the-art systems that focus on epistemic aspects (how evaluations are calculated), our model deals mainly with the pragmatic aspects of reputation information. The main characteristics of the system are:

- It is modular. The model is defined as a multi-context system (MCS) [Giunchiglia and Serafini, 1994], a framework that allows several distinct theoretical components to be specified together, with a mechanism to relate these components. From a software engineering perspective, MCS supports modular architectures and encapsulation. From a logical modeling perspective, it allows the construction of agents with different and well-defined logics, keeping all formulas of the same logic in their corresponding context. This increases considerably the representation power of logical agents, and at the same time, simplifies their conceptualization. In our model, each main attitude (Belief, Desire and Intention) is specified as an independent context. Also, the Repage system is introduced as a context. Our model specifies then how such contexts are related to each other, defining the practical reasoning path of the agent. This modular architecture permits easy integrations of possible modules that could extend the functionalities of the original one.

- It is based on solid logical frameworks. We use an existing complete logic of preferences based on Lukasiewicz [Casali, 2008] to model desires and intentions, and we introduce a new logic to deal with the beliefs of the agent. The belief logic is a classical first-oder many-sorted logic, deals with probabilities and is capable of representing and combine the information that the reputation model Repage computes. Differently from other probabilistic logics, it handles multiple probability distributions under some restrictive settings, and because it is specified as a first-order logic, it permits a smooth implementation.

- It handles *image* and *reputation*. The Repage model is based on a cognitive theory of reputation that states a main difference between image and reputation. While both objects are social evaluations, image refers to a simple evaluative belief that tells how agents are in a certain context, and reputation is a metabelief, telling that a given social evaluation circulates in the society. The belief logic that we develop captures both concepts and combine them, defining a family of agents depending on how such combination is performed.

- It can be seen as an instantiation of a cognitive trust model. Some cognitive theories of trust suggest that trust is a mental state composed of a set of beliefs and goals that describe the *decision* to rely on someone, so, it is the result of a practical reasoning process. Our model fits into this description and becomes, as far as we know, the only cognitive trust model that describes each step of the reasoning process.

- It is generic. The model is not attached to any specific domain ontology nor network typology, and inherits the properties and characteristics of the underling reputation model. We use Repage as a paradigmatic example, but any model whose information can be captured by the reputation language L_{rep} could be placed into the system.

Third - An Argumentation-Based Protocol for Reputation Exchange

We develop an argumentation-based dialog protocol for the exchange of reputation-related information. Due to the subjectivity of reputation information, a social evaluation totally reliable by an agent A may not be reliable for B, because the bases under which A has inferred the social evaluation cannot be accepted by B. This can happen because agents have different inference rules, have had different experiences, have different goals, etc. When such information is communicated this can become very problematic, specially if the reputation model assigns a reliability measure to the communicated information, because of the reasons above.

The argumentation-based protocol we develop offers a possible solution for this, and can complement already existing methods. We suggest that, in communicated social evaluations, the reliability measure cannot be dependent on the source agent, but must be fully evaluated by the recipient agent accordingly to its own knowledge. Then, taking advantage of the internal structure of reputation-related information, rather than allow only single communications, we allow agents to *justify* their communications following the guidelines of the argumentation-based protocol. Then, the agent can incrementally construct a tree of arguments with their attack relations that can be used to decide on the reliability (and thus acceptance) of a communicated social evaluation. The main characteristics of the system are:

- Only the recipient agent decides about the reliability of a communicated evaluation. This differs from other approaches in which the source agent attaches a reliability measure to the communicated social evaluation. This makes more difficult for dishonest agents to intentionally send fraudulent information, because they must be aware of the knowledge of the recipient and justify the *lie* accordingly.

- It uses argumentation frameworks to give semantics to the dialog. We exploit the L_{rep} language to completely define how arguments are constructed and how arguments influence one another. We instantiate a weighted abstract argument framework to define the acceptability semantics of a communicated social evaluation.

- It handles quantitative and qualitative graded information. One of the main characteristics of reputation information is that it is graded. Nowadays it is strange to find a model that provides crisp evaluations of the agents. For instance, an agent A may be *bad*, *very bad* or *very good* etc. as a car driver, and this has to be taken into account when arguing about evaluations.

- It permits dialogs between parties that use different reputation models. Even when we assume that agents use the same language to talk and reason about reputation information (L_{rep} language), we suppose that they can use different inference rules (different reputation models) without having to exchange the exact rules that each agent uses for the inferences.

Next section provides a detailed explanation of the structure of the book.

1.3 Overview and Structure of the Work

The book is structured in seven chapters and two appendixes:

Chapter 2: We present the theoretical bases of our work and a survey of the most relevant computational trust and reputation models that currently exists in literature. On the one hand, in the first part of the chapter, we introduce the cognitive theory of reputation presented by Conte and Paolucci [Conte and Paolucci, 2002], relating their definition of *image* and *reputation* with other definitions and with the notion of *cognitive trust* pointed out by some authors. Furthermore we explain Repage [Sabater-Mir et al., 2006], a computational reputation model based on [Conte and Paolucci, 2002] and explore some of its advantages by detailing empirical results that we obtained through simulations. On the other hand, the second part of the chapter is devoted to a survey of the current state-of-the-art reputation and trust models. We describe three other surveys and examine the different dimensions of analysis that each one of them proposes. At the end of the chapter, we also propose a complementary classification.

Chapter 3: The objective of this chapter is to establish a taxonomy of reputation-related concepts. First, we define an ontology of reputation to explicitly state the elements that according to us, are important in the field. Our ontology has a clear computational perspective and serves as a taxonomy of the concepts that our work uses. Second, we introduce the L_{rep} language, a first-order language to express reputation-related concepts described in the ontology, and that agents use to write statement and reason about reputation information. We provide examples to show how the language captures a wide range of state-of-the-art models, specially the Repage model, which currently is one of the most expressive models.

Chapter 4: We introduce the BC-logic, a belief logic capable of integrating reputation information coming from reputation models like Repage, with the normal beliefs that the agent holds about the domain. It is a sorted first-order logic that manages probability predicates and that subsumes all possible inconsistencies in terms of probabilities. We prof that the proposed theory used by the agents to reason is consistent and decidable, since it can be seen as a set of universal horn clauses.

Chapter 5: The chapter proposes the BDI+Repage architecture. We specify the architecture using multicontext systems [Giunchiglia and Serafini, 1994] and use the logic defined in chapter 4 to manage the belief base of the agent. We specify one context for each main attitude of the agent (Belief, Desire Intention) and design the links (bridge rules) among those contexts, designing a practical reasoning process. Even when we use the Repage reputation model in the integration, it should be taken as a paradigmatic example, since the only requirement is to manage reputation models whose information can be captured by the language L_{rep}.

Chapter 6: While the previous chapter deals with pragmatic-strategic decisions, on how agents use reputation information to decide what to do, this chapter struggles with memetic decisions. We face a particular problem attached to the fact that reputation information is subjective. We define a protocol specifically designed for the exchange of reputation-related information between two-parties that uses argumentation techniques. We exploit the L_{rep} language and use it to build an argumentation system capable of providing a semantics to decide whether a communicated social evaluation can be considered reliable for the agents.

Chapter 7: We conclude our analysis and provide some future research lines.

Appendix A: In this appendix we introduce the concept of conversion uncertainty (CU), a measure of information loss produced when transforming from one representation type to another. We define it as a conditional entropy. We provide the detailed CU calculations for all possible transformations.

Appendix B: We present some implementation details of the BDI+Repage model using the Jason platform [Bordini et al., 2007], and we present preliminary results that empirically proof the differences in behavior of some of the families of agents that we presented in chapter 5.

1.4 Related Publications

The following publications are a direct consequence of the development of this work.

- I. Pinyol, J. Sabater-Mir, P. Dellunde and M. Paolucci. Reputation-Based Decisions for Logic-Based Cognitive Agents. In Journal of Autonomous Agents and Multi-Agent Systems (JAAMAS). In Press.
- I. Pinyol and J. Sabater-Mir. An Argumentation-Based Dialog for Social Evaluations Exchange. In proceedings of the 19th European Conference on Artificial Intelligence (ECAI'10), Lisbon, Portugal. In Press.
- I. Pinyol, R. Centeno, R. Hermoso, V. Torres da Silva and J. Sabater-Mir. Norms Evaluation through Reputation Mechanisms for BDI Agents. In proceedings of the 13th International Congress of the Catalan Association for Artificial Intelligence (CCIA'10). In Press.
- I. Pinyol and J. Sabater-Mir. Metareasoning and Social Evaluations in Cognitive Agents. In Autonomic Computing and Communication Systems,volume 23 of LNISCT, pages 220-235. Springer Berlin Heidelberg, 2010.
- I. Pinyol and J. Sabater-Mir. Pragmatic-strategic reputation-based decisions in bdi agents. In proceedings of the 8th International Conference on Autonomous Agents and Multiagents Systems (AAMAS'09), Budapest, Hungary,pages 1001-1008, 2009.
- I. Brito, I. Pinyol, D. Villatoro, and J. Sabater-Mir. HIHEREI: human interaction within hybrid environments regulated through electronic institutions. In proceedings of the 8th International Conference on Autonomous Agents and Multiagents Systems (AAMAS'09), pages 1417-1418, 2009 (best student demo award),
- I. Pinyol and J. Sabater-Mir. Towards the definition of an argumentation framework using reputation information. In Proceedings of the workshop Trust in Agent Societies (TRUST@AAMAS09), Budapest, Hungary., pages 92-103, 2009.
- S. Konig, I. Pinyol, D. Villatoro, J. Sabater-Mir, and T. Eymann. An architecture for simulating Internet-of-services economies. In Proceedings of 7th German conference on Multiagent System Technologies (MATES'09), Hamburg, Germany. Volume 5774 of LNCS Science. Springer, 2009.
- I. Pinyol, J. Sabater-Mir, and P. Dellunde. Probabilistic dynamic belief logic for image and reputation. Frontiers in Artificial Intelligence and Applications, IOS Press, vol 184 p197-205, Spain, 2008.
- I. Pinyol and J. Sabater-Mir. Cognitive social evaluations for multi-context bdi agents. In 9th Annual International Workshop Engineering Societies in the Agents World (ESAW'08), 2008.
- I. Pinyol and J. Sabater-Mir. Arguing about reputation. the lrep language. In 8th Annual International Workshop Engineering Societies in the Agents World, volume 4995 of LNCS, pages 284-299. Springer, 2007.

- I. Pinyol, M. Paolucci, J. Sabater-Mir, and R. Conte. Beyond accuracy.Reputation for partner selection with lies and retaliation. Multiagent-based Simulation (MABS'07). Volume 5003 of LNCS, pages 128-140.Springer, 2007.

- I. Pinyol, J. Sabater-Mir, and G. Cuni. How to talk about reputation using a common ontology: From definition to implementation. In Proceedings of the Ninth Workshop on Trust in Agent Societies (TRUST@AAMAS'07). Hawaii, USA., pages 90-01, 2007.

- A. di Salvatore, I. Pinyol, M. Paolucci, and J. Sabater. Grounding reputation experiments. a replication of a simple market with image exchange. In Proceedings of the Model to Model Workshop (M2M07), Marseille, France, pages 32-45, 2007.

- J. Sabater-Mir, I. Pinyol, D. Villatoro and G. Cun. Towards Hybrid Experiments on Reputation Mechanisms: BDI Agents and Humans in Electronic Institutions. Proceedings of the 12th Conference of the Spanish Association for Artificial Intelligence (CAEPIA'07),Salamanca, SPAIN, 2007.

Chapter 2

Reputation In Multiagent Systems

2.1 Introduction

In this chapter we describe the theoretical framework under which all our work holds: the cognitive theory of reputation developed by Conte & Paolucci [Conte and Paolucci, 2002]. We put this theory in contrast with cognitive visions of trust, and show the Repage system, a computational model of reputation based on this theory. In the last part of this chapter we provide a survey of the most representative computational trust and reputation models. We first analyze other reviews showing the different characteristics or dimensions that such models can be classified on, and then we provide our own classification to part of our contribution.

2.2 A Cognitive Theory of Reputation

The vision we have about reputation is set in the cognitive theory developed in [Conte and Paolucci, 2002]. From this theory, reputation cannot be seen as a single-dimensional concept, but need to be understood as a multi-faced artifact that not only focuses on the evaluative dimension but on the process and the effect of transmission. For this, it is crucial the distinction between image and reputation. While image is an evaluative belief that tells how a given target agent behaves according to certain criteria, reputation is a belief on the circulation of social evaluations in the society.

According to [Conte and Paolucci, 2002], reputation proceeds from the individual level to the propagation at a social level, and from the social level, it comes back to the individual. This dynamism makes reputation susceptible to changes, and thus, the accuracy of reputation information is more than questionable. However, as described in some simulation experiments

[Pinyol et al., 2007a, W. Quattrociocchi, 2008b] even with a high rate of *cheating* when transmitting reputation (around 60%) the society seems to perform better when the distinction between image and reputation is taken into account rather than when only image is taken into account. This is also analyzed at the end of this chapter.

As we mentioned in the first chapter, we focus our efforts on the individual and interaction level. Both levels are explicitly taken into account by the cognitive theory when it describes the three levels of mental decisions that agents can perform regarding social evaluations:

- **Epistemic decisions** cover the dynamics of beliefs regarding image and reputation, or in other words, decisions about updating and generating social evaluations, for instance, accepting the beliefs that form images or acknowledging certain reputation. From our point of view, this level fits within a *theoretical reasoning*, which embraces reasoning about *what* to believe from the individual perspective[1].

- **Pragmatic-strategic decisions** are decisions that use image to decide how to behave/interact with potential partners. From our perspective this kind of decisions are done through a process of *practical reasoning*, a kind of reasoning to decide how to act, and very used in logical approaches. Also, we recall here that when deciding to whom to interact with, agents are deciding in fact to whom to rely on to achieve their goals. This brings us to the notion of social trust [Castelfranchi and Falcone, 1998b]. Later in this chapter we analyze the relation between image, reputation and trust from a cognitive perspective.

- **Memetic decisions** refer to the decisions of how and when to spread social evaluations. These decisions are produced also by a process of practical reasoning, but focusing on communication actions.

The cognitive theory holds on two important concepts: image and reputation. Although both are social evaluations, they are distinct objects and involve different groups of agents. In the next subsections we summarize the main ideas behind these two constructs extracted from [Conte and Paolucci, 2002].

2.2.1 Image

Image is an evaluative belief, a belief that describes an evaluation of a target, that can be a single agent or supra-agent (like groups or institutions), towards a specific context. In fact, from both [Conte and Paolucci, 2002] and [Miceli and Castelfranchi, 2000] an image requires the context to be a goal that the agent wants to achieve. Hence, an agent A evaluates another agent B when A

[1] In philosophy, practical reasoning is the capacity to deliver (or reason) about how to act. In contrast, theoretical reasoning is the capacity to reason about what to believe. Therefore, epistemic decisions can be considered theoretical reasoning, while pragmatic-strategic decisions, practical reasoning

thinks that B is *good* or *bad* for achieving the goal. We will see how this constrain is relaxed when considering Repage [Sabater-Mir et al., 2006], the computation model inspired in this cognitive theory and the base for our work. For instance, in the most simplified scenario, an agent can hold a *very good* image of *John* in the context of *obtaining 2 boxes of high-quality wine*.

The theory describes three sets of agents that participate in a given social evaluation as image:

- **Evaluators**: A nonempty set of agents that share the evaluation. Hence, they must share the same goal.
- **Targets**: A nonempty set of agents or supra-agents that are evaluated by the set of evaluators.
- **Beneficiaries**: A nonempty set of agents that use the evaluations, and thus, share the same goal.

It is important to notice that the sets of evaluators and beneficiaries do not necessary are the same. This is very clear in online reputation mechanism, like eBay [eBay, 2002], where buyers evaluate sellers and these evaluations are used by other buyers.

The cognitive theory [Conte and Paolucci, 2002] makes some predictions on the quality of the evaluations when assuming overlapping of roles. For instance, they argue that when there is a low overlapping between the targets and the beneficiaries, an overestimation is produced in the evaluations. Instead, when there is a high overlapping, evaluations are accurate. Moreover, when beneficiaries and evaluators are lowly overlapped evaluations are mostly underrated, while when they are highly overlapped, accuracy is the norm in the evaluations. In [eRep, 2007] the authors show an empirical validation of such predictions, checking different online user-oriented reputation systems found in electronic auctions, recommender systems, discussion forums and social networks where the overlapping of roles is known.

From the individual perspective, agents can be partially aware of such sets and can in fact act in all the roles. We assume that our cognitive agent i is endowed with goals and beliefs and therefore is able to generate evaluations about other agents. Then, i acts as evaluator when performing epistemic decisions. As well, i can act as beneficiary when receives evaluations from a set of agents S. In this case, i knows that the agents in S act as evaluators. Also, when i decides to send its own evaluations to a set of agents D, i is aware that agents in D may act as beneficiaries, and that they will know that i is an evaluator. Curiously, i may not be aware that she is actually being targeted by others, but must be aware of such possibility and the consequences of achieving bad evaluations. Because of that, cognitive agents have the *motivation* to act accordingly to well-established social behaviors.

2.2.2 Reputation

The theory considers reputation as a belief about others's evaluations. From a broad sense, it can be considered a meta-belief, although when focusing on the individual level this is not necessary true. The theory analyses the roles of agents participating in a given social evaluation as reputation:

- **Evaluators**: A nonempty set of agents that share the evaluation. Hence, they must share the same goal.
- **Targets**: A nonempty set of agents or supra-agents that are evaluated by the set of evaluators.
- **Beneficiaries**: A nonempty set of agents that use the evaluations, and thus, also share the same goal.
- **Third Parties**: A nonempty set of agents that acknowledge that some evaluators share the evaluation.

The first three roles are the same as for image. Here though, the theory introduces a third party agents group. This group shares the belief that a group of evaluators is endowed with the social evaluation. Third parties are the holders of the reputation, and often they completely include the set of evaluators. Third parties are those aware of the effects of reputation transmission and the ones that transmit reputation (so called *gossip*).

At the individual level, when a third party transmit reputation information about a given target to a set of agents, does not necessary believe the corresponding image of the target. This is because reputation moves to a level above of image, the belief about the circulation of an evaluation. Thus, when an individual agent accepts or acknowledges a given reputation it indicates that the agent assumes that the nested evaluation circulates in the society, that most of the members of the society, if asked, would acknowledge the existence of a circulating voice about the target.

The cognitive theory suggests that reputation information circulates more in the society than image information. The authors argue that the transmission of image is in fact the transmission of a set of evaluative beliefs owned by the source of the communication. Then, when i communicates her image, she is communicating her beliefs, implicitly committing to the truth of the evaluation. Instead, when transmitting reputation there is no such commitment. From this, it can be deduced that agents will transmit image only when they are very *secure* about the truth of the evaluation. Reporting on reputation implies a minor responsibility. However, the counter-effects are obvious: falsifying reputation information is not costly either.

2.2.3 Why Is Reputation a Meta-belief?

The introduction of the third parties set in the previous definition of reputation must be reinterpreted when talking about the mental state of an individual agent that accepts a given reputation. In this sense it is important to remark that reputation is the belief about the existence of a communicated evaluation. Since the notion of evaluation implies a sort of beliefs, reputation is considered a meta-belief. It should not be confused with the idea of shared image, which is also a meta-belief but a particular case of image.

Cognitive agents can have in their minds beliefs about others' evaluations, so, beliefs about others' images. When these images correspond to the same target in the same context we say that the agent is aware of a shared image, that is also a meta-belief. Probably this information is part of the own image of the agent. The difference with reputation is twofold:

1. The agent can totally identify the owner of each image, while in reputation this is lost in the generalization.

2. There is no reference to the transmission. As said before, reputation is a belief about the existence of a voice about the target, that most of the people in the society communicates such evaluation, but where the evaluation does not have any concrete referent. From this point, it can be deduced that *without explicit communication, no reputation is possible.*

In fact, the epistemic decision towards the acceptance of a reputation must go through a generalization process. First, an agent can be aware that a determined set of agents report the existence of a voice about a target, so, the agent is aware of a shared voice. When the agent generalizes the shared voice and assumes that instead of a concrete set of agents most agents would report the existence of the voice, the agent is, in that moment, acknowledging a reputation.

2.3 Reputation and Social Trust

Reputation and social trust have a close relationship. In fact, we can argue that social trust is built on the bases of reputation and image, and at the same time, reputation is constructed on the social trust. But, what is trust[2]?

As almost always happens with concepts that have a strong common sense component, there is not a global accepted definition of trust. May be the most classical one was given by Gambetta [Gambetta, 1990]: *"Trust is a subjective probability by which an individual A expects, that another individual B performs a given action on which its welfare depends"*. However, to exemplify better the relationship between reputation and trust, we need to rely on the internal components of trust. For this, we base our argument on the definition of trust provided by Castelfranchi and Falcone [Castelfranchi and Falcone, 1998b, Castelfranchi and Falcone, 1998a], where trust is analyzed in terms of the cognitive components that it is based on. According to the authors, trust is a mental

[2] We refer to social trust as the trust towards other entities, groups of entities or social organizations, not about the trust on objects. From now on, we will use trust in the former sense.

state, a complex attitude composed of beliefs and goals that determine the expectations towards certain behaviors of the trustee agents. Furthermore, they defend that trust is scalable, and that the *degree* of trust is based on the subjective certainty of the composing beliefs and the utility (or importance) of the goals. From this idea we can deduce that trust is in fact a practical reasoning process.

In a more formal way, let i, j be two agents, the cognitive components that make i trusts j regarding the goal g are the following [Castelfranchi and Falcone, 1998b][3]:

- **Goal Seeking**: i has the goal g.
- **Competence Belief**: i believes that j is capable of obtaining g from a set of actions (summarized in the action α)
- **Disposition Belief**: i believes that j will actually perform α to obtain g. This belief makes agents predictable.
- **Dependence Belief**: i believes that she needs/depends on j to perform the task.

Competence and disposition beliefs, together with the goal are the *core trust*. They model the ability and willingness of the agent j to achieve g. They are evaluative beliefs and are constituents of the image and reputation of j in the sense described in subsections 2.2.1 and 2.2.2. This property was already mention in [Miceli and Castelfranchi, 2000], where the authors exemplify which kinds of beliefs compose evaluations, and the capabilities that cognitive agents must achieve in order to be evaluators.

We can conclude then that image and reputation are necessary conditions, although not sufficient, to achieve trust on j. For this it is also necessary the *decision* to rely on j, to not search for any other alternative resource to achieve g, so, to achieve g through j. This is summarized in the dependence belief, and it is the reason that justifies the consideration of trust as a practical reasoning, and very related to pragmatic-strategic decisions.

2.3.1 Occurrent vs. Dispositional Trust

Other contributions [Herzig et al., 2008] refine this notion of social trust by differentiating *occurrent* from *dispositional* trust. The former is understood as the trust on other agents to act here and now, and coincide with the core trust definition given by Castelfranchi and Falcone. In contrast, dispositional trust denotes the disposition of the trustee to perform an action in order to obtain a potential goal when some conditions hold [Herzig et al., 2008].

[3] The authors describe other beliefs and goals that are part of the trust mental state, like fulfillment belief or wishes. However, for the sake of clarity we obviate them because they are a direct cause of the beliefs shown in the list.

From a more technical perspective the authors define occurrent trust with the predicate $OccTrust(i, j, \alpha, \varphi)$, indicating that i trusts j here and now to perform action α to obtain goal φ. As in the definition of core trust from Castelfranchi and Falcone [Castelfranchi and Falcone, 1998b], the components embrace an occurrent goal, an occurrent capability belief, an occurrent power belief and an occurrent intention belief. More formally:

$$\begin{aligned} OccTrust(i, j, \alpha, \varphi) \quad =_{def} \quad & OccGoal_i(\varphi)\wedge \\ & Belief_i(OccCap(j, \alpha))\wedge \\ & Belief_i(OccPower(j, \alpha, \varphi))\wedge \\ & Belief_i(OccIntends(j, \alpha)) \end{aligned}$$

The beliefs on the occurrent capability and occurrent power correspond to the competence beliefs, while occurrent intention to the disposition belief. Regarding dispositional trust, the background components are the same but we move from *occurrent* goals to *potential* goals, and from *occurrent* beliefs to *potential* beliefs. Following [Herzig et al., 2008], dispositional trust is defined as follows:

$$\begin{aligned} DispTrust(i, j, \alpha, \varphi) \quad =_{def} \quad & PotGoal_i(\varphi)\wedge \\ & Belief_i(CondCap(j, \alpha))\wedge \\ & Belief_i(CondPower(j, \alpha, \varphi))\wedge \\ & Belief_i(CondIntends(j, \alpha)) \end{aligned}$$

On the one hand, the potential goal refers to a goal that *currently* is not the case, but that at some point it may be occurrent. Hence, from a logical sense, occurrent trust implies dispositional trust but not the opposite. On the other hand, the notions of conditional capability, power and intention are the conditioned versions of the occurrent trust components. They describe under which conditions agent i believes that j has the capability to execute α, the power to achieve φ from α, and the intention to perform α. The work [Herzig et al., 2008] also presents a formalization of these predicates using some of the existing logics for autonomous agents and multiagent systems. We get into more technical details in chapter 4.

Dispositional trust is important for our work due to the relation with reputation. The authors argue that the notion of reputation described by Conte and Paolucci in [Conte and Paolucci, 2002] is the equivalent dispositional trust but at a group level. Their definition of reputation involves also four parameters, $Rep(G, j, \alpha, \varphi)$, where G is a group of agents, and its components are:

$$\begin{aligned} Rep(G, j, \alpha, \varphi) \quad =_{def} \quad & PotGoal_G(\varphi)\wedge \\ & GroupBel_G(CondCap(j, \alpha))\wedge \\ & GroupBel_G(CondPower(j, \alpha, \varphi))\wedge \\ & GroupBel_G(CondIntends(j, \alpha)) \end{aligned}$$

Group belief should not be confused with the standard notion of common belief. In general, we say that when a group G has a common belief that φ, each member of the group believes φ and is aware that the other members also believe φ [4]. Instead, when the group G has a group belief that φ, it only implies that each agent of the group believes that G has a group belief that φ, and there is a mutual belief of this fact among the members of the group:

$$GroupBel_G\varphi$$

implies that for each $i, j \in G$,

$$Bel_i(GroupBel_G\varphi) \wedge Bel_i(Bel_j(GroupBel_G\varphi))$$

A logical account for such operator can be found in [Gaudou et al., 2006] (operator G) and as the same authors mention, it is closely related to the group belief described in [Tuomela, 1992]. From this definition it is important to notice that $GroupBel_G\varphi$ does not imply $Bel_i\varphi$ even when $i \in G$. The authors argue that this notion of reputation coincides with the reputation concept from Conte and Paolucci [Conte and Paolucci, 2002], since accepting a group belief does not mean to accept the individual belief.

From our point of view, the definition of reputation that we have given in this work, based on [Conte and Paolucci, 2002], goes beyond group beliefs, since one dimension of reputation is the transmission effects, and the acknowledge of a voice that *circulates* in the society. In this sense, it is more closely related to the notion of grounded information established by a group of agents [Gaudou et al., 2006]. In this work, $\mathcal{G}_G\varphi$ stands for *it is publicly grounded for the group G that φ holds*. The interesting element here is the expression *publicly grounded*, which implies some kind of transmission through the group. It refers to an objective notion, to what it can be observed by the agents in terms of social commitments [Walton and Krabbe, 1995] that are public and observed by the members of the group. Nevertheless our notion of reputation requires a generalization in terms of identifying the individual members of the group.

Some issues remain still unclear in this approach, being the bootstrapping the most relevant. Although Conte and Paolucci's theories have at some extend the same problem, the definition of the computational model Repage helps in this enterprise.

2.4 Towards the Reputing Agent: The Repage System

The Repage system [Sabater-Mir et al., 2006] is a computational reputation model that tackles epistemic decisions and that supports the computation of image and reputation information.

It is important to remark that Repage does not determine the actual image and reputation in terms of beliefs as described in the theory. It gives components that, from the actual cognitive theory, should *support* or *feed* the creation of

[4] Formally, $CommonBel_{\{i,j\}}\varphi =_{def} Bel_i\varphi \wedge Bel_j\varphi \wedge Bel_iBelj\varphi \wedge Bel_jBel_i\varphi \wedge \ldots$

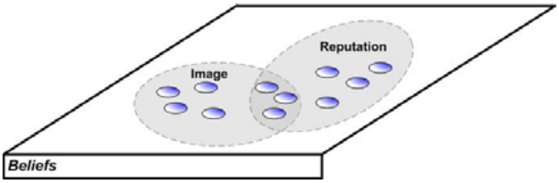

Figure 2.1: Image and reputation as a mental state of a cognitive agent. Parts of the elements that form the social evaluations believed by the agent (image) may come from reputation.

social evaluations as image and reputation. From the cognitive theory it can be deduced that a cognitive agent is endowed with beliefs that describe the mental state in which image and reputation are created. It is possible that part of the beliefs that represent reputation in the mind of the agent are used to create the image of the agent (what she thinks), or even the opposite (see figure 2.1). For instance, *John* may belief that *Ann* is a good social worker because she acknowledges a reputation of *John* as such. The cognitive theory does not make any statement about the interplay between image and reputation, but certainty, it is crucial in the design of cognitive agents. It may require more pragmatic approaches that rely on concrete architectures.

Repage [Sabater-Mir et al., 2006] is a computational model that gives support to cognitive agent architectures that want to distinguish between image and reputation. Figure 2.2 shows how Repage gives support to the beliefs that in the mind of the agent are image and reputation. Notice that inside Repage, supporting information for image never collides with supporting information for reputation. Repage implements a way to compute social evaluations, not how such information interferes. This is done at the belief level, because other information may pay a crucial role (see figure 2.2). In the following subsections, we get in touch with the internal elements of Repage and show some simulation results that illustrate the importance to keep the difference between image and reputation.

2.4.1 The Repage Architecture

In the Repage architecture we find three main elements, a memory, a set of *detectors* and the *analyzer* (see figure 2.3). The memory is composed of a set of references to the predicates hold in the main memory of the agent. Only those predicates that are relevant for the calculus of reputation and image are considered.

In the memory, predicates are conceptually organized in different levels of abstraction and inter-connected. Each predicate that belongs to one of the main types (image, reputation, shared voice, shared evaluation[5], valued communica-

[5] Repage's authors use the term *shared evaluation* instead of *shared image*, even though they refer to the same conceptualization

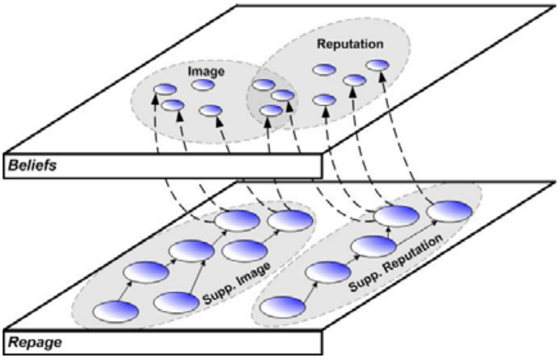

Figure 2.2: The Repage Integrated in the Beliefs of the Agents.

tion and outcome) contains an evaluation that refers to a certain agent in a specific role. For instance, an agent may have an image of agent A (target) as a seller (role), and an image of the same agent A as informant.

The value (evaluation) associated to a predicate is a tuple of five numbers summing to one, plus a strength value. Each number has an associated label in the rating scale: very bad (VB), bad (B), neutral (N), good (G) and very good (VG). We call this representation a *weighted labeled tuple* and it represents a probability distribution. In the new Repage implementation, this is generalized to t partitions and it is associated to the role being evaluated. For instance, the role *car sellers* can have a binary evaluation (*bad* and *good*), and the role *fruit seller* can have four (*very bad*, *regular*, *good*, *very good*).

The network of dependencies specifies which predicates contribute to the values of others. Each predicate (except those at the bottom level) has a set of antecedents and at the same time contributes to the calculation of other predicates. The *detectors*, inference units specialized in each particular kind of predicate, receive notifications from predicates that have changed or that appear in the system, like new communications or new fulfillments, and use the dependences to recalculate the new values and to populate the memory with new predicates. Moreover, each predicate has associated a strength that is a function of its antecedents and of the intrinsic properties of each kind of predicate. As a general rule, predicates that resume or aggregate a bigger number of predicates will hold a higher strength. In the next subsection we formally define how Repage aggregates strengths and probability distributions.

At the first level of the Repage memory we find a set of predicates not evaluated yet by the system.

- Contract: agreements of a future interaction between two agents. For instance, in an e-Commerce environment, an agent expects to obtain a certain quality of a product after paying for it a determined price.

- Fulfillment: the result of the interaction. In the same e-Commerce example, the fulfillment would be the real quality of the product the agent

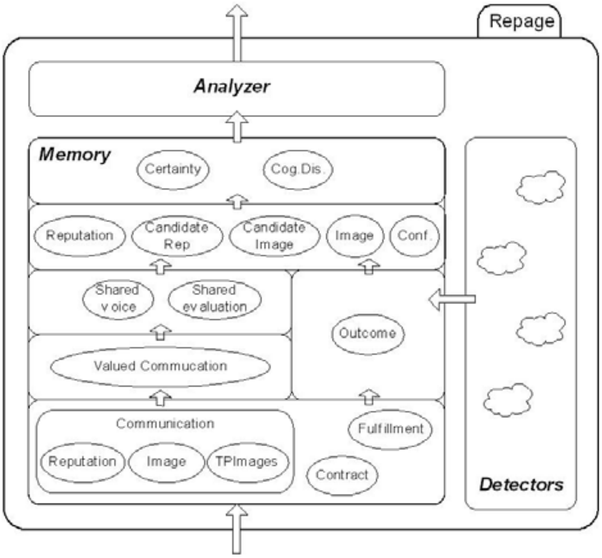

Figure 2.3: The Repage Architecture (Source: [Sabater-Mir et al., 2006]).

got.

- Communications: Information that other agents may communicate about others evaluations. These communications may be related to three different aspects: the image that the informer has about a target, the image that according to the informer a third party agent has, and the reputation that the informer assigns to the target.

Contracts and fulfillments implement direct experiences. The contract represents the *agreement* or *expectations* that the initiator of the interactions expects to obtain after the interactions. The fulfillment contains what it has been obtained after the interactions.

In level two we have two kind of predicates:

- Valued communication: The subjective evaluation of the communication received that takes into account, for instance the image the agent may have of the informer as informant. Communications from agents whose credibility in terms of image or may be reputation are low, will not be considered as strongly as the ones coming from well reputed informers.

- Outcome: The agent's subjective evaluation of the direct interaction. From a fulfillment and a contract a *detector* builds up an outcome predicate that evaluates the particular transaction.

In the third level we find two predicates that are only fed by valued communications. On the one hand, a shared voice will hold the information received about the same target and same role coming from communicated reputations. On the other hand, shared evaluation is the equivalent for communicated images and third party images.

Shared voice predicates generate candidate reputations, and shared evaluations together with outcomes, candidate images. In this fourth level, candidate reputation and candidate images are not strong enough to become a full reputation and image respectively. New communications and new direct interactions will contribute at this level to enrich these predicates and therefore "jump" to images and reputations.

The last level implements cognitive dissonances and certainties. From the point of view of the agent, different pieces of relevant information may conclude in contradictory information (cognitive dissonance) or the opposite, certain information. In the case of dissonance, the *analyzer* will propose actions to the agent in order to solve the contradiction. We refer to ([Paolucci et al., 2005]) for a more detailed explanation about how the *analyzer* works. Nevertheless, the work proposed here suggests that such capabilities take place outside Repage, at the belief level of the agent.

Aforesaid, the computation of evaluations and strength for each predicate is done through aggregation functions. The network of dependencies determines which evaluations and strength must be aggregated. In the next subsection we detail such aggregation functions.

2.4.2 Aggregation Functions for Repage

We have two elements to aggregate: probability distributions and strengths. Lets consider n evaluations to aggregate with their respective strength:

$$v_1, s_1, \ldots, v_n, s_n$$

And let v_{ij} be the weight j of the evaluation i.

Aggregation of Evaluations

The current implementation of Repage system aggregates the previous evaluations in a new evaluation r as

$$\forall k : 1 \leq k \leq t : r_k = \frac{\prod_{i=1}^{n}(s_i v_{ik} + \frac{1-s_i}{t})}{\sum_{j=1}^{t}\prod_{i=1}^{n}(s_i v_{ij} + \frac{1-s_i}{t})} \tag{2.1}$$

As the authors claim in [Sabater-Mir et al., 2006] the previous aggregation function is associative and distributive. Indeed, the evaluation e_I (which all the weights are $1/t$, the maximum randomness) acts as the identity element of the function. Thus, $f(v, e_I) = v$ for any arbitrary evaluation v. However, the previous evaluation has some problems if weights or strengths are 0. In this case a rectification function is applied to both factors (see [Sabater-Mir et al., 2006] for the details). Also, in [Sabater-Mir and Paolucci, 2007] another aggregation function is proposed, solving the problems of weights and strength of value 0.

Aggregation of Strengths

In Repage, the strength models uncertainty in the evaluations, the reliability. The current implementation of Repage considers that uncertainty reduces as the number of evaluations being aggregated increases. Then, let s_r be the resulting strength, the aggregation function is quite simple and defined as

$$s_r = 2\arctan(\sum_{i=1}^{n} s_i)/\pi \tag{2.2}$$

Notice that

$$\lim_{x\to\infty} 2\arctan(x)/\pi = 1 \tag{2.3}$$

The function "arctan" is used to normalize the value. Of course, this function can be manipulated by adding an exponent to it. Then, with higher exponent, the new strength value would need more aggregations to get closer to 1, which should be considered the maximum certainty.

The previous calculation of strengths does not consider at all the weights. In [Sabater-Mir and Paolucci, 2007] an alternative version of such aggregation is given, by considering that aggregations of evaluations representing very different values should in fact decrease the strength of the resulting aggregation, and the other way around. Therefore the new proposed aggregation function takes into account the *differences* among evaluations.

2.4.3 Experimental Results

Several simulations have been performed to show the importance of keeping the distinction between image and reputation [Pinyol et al., 2007a, W. Quattrociocchi, 2008b, W. Quattrociocchi, 2008a, di Salvatore et al., 2007] using the Repage system. In this subsection we detail the exploratory work done in [Pinyol et al., 2007a], and summarize the results found in [W. Quattrociocchi, 2008b], since the latter are based on the underlying simulation model presented in the former.

To remark the difference between the effects of reputation and image, in [Pinyol et al., 2007a] we explored through agent-based simulations the use of both kind of social evaluations in two experimental conditions:

- **L1**, where there is only exchange of image between agents
- **L2**, where agents can exchange both image and reputation.

While L1 is comparable with a large body of similar literature (ex. [Sen and Sajja, 2002a]), the introduction of reputation (L2) as a separate object in a simulative experiment was presented in [Pinyol et al., 2007a] for the first time.

Description of the Simulation

The simulation experiment was designed as the simplest possible setting where accurate information is a *commodity*, meaning that information is both valuable and scarce. This simplified approach is largely used in the field [Sen and Sajja, 2002a], both on the side of the market and agent design.

The experiment includes only two kind of agents, the buyers and the sellers. All agents perform actions in discrete time units (turns from now on). In a turn, a buyer performs one communication request and one purchase operation. In addition, the buyer answers all the information requests that it receives.

Goods are characterized by an utility factor that we interpret as quality (but, given the level of abstraction used, could as well represent other utility factors as quantity, discount, timeliness) with values between 1 and 100.

Sellers are characterized by a constant quality and a fixed stock, that is decreased at every purchase. They are essentially reactive, their functional role in the simulation being limited to providing an abstract good of variable quality to the buyers. Sellers leave the simulation when their stock is exhausted or when for certain number of turns they do not sell anything, and are substituted by a new seller with similar characteristics.

The disappearance of sellers makes information necessary. Reliable communication allows for faster discovering of the better sellers. This motivates the agents to participate in the information exchange. In a setting with permanent sellers (infinite stock), once all buyers have found a good seller, there is no reason to change and the experiment freezes. With finite stock, even after having found a good seller, buyers should be prepared to start a new search when the good seller's stock ends.

At the same time, limited stock makes good sellers a scarce resource, and this constitutes a motivation for the agents not to distribute information. One of the interests of the model is in the balance between these two factors. There are four parameters that describe an experiment: the number of buyers NB, the number of sellers NS, the stock for each seller S, and the distribution of quality among sellers.

While sellers are totally reactive and sell products on demand, each buyer is endowed with the Repage system to help them figure out image and reputation predicates, and use them to decide (1) to whom to ask, (2) to whom to buy, (3) how to answer inquires from other buyers. All three decisions determine the performance of the whole system, which is calculated through the overall quality obtained by the buyers at each turn. The definition of the decision making procedure is key for the obtained results. Thus, we detail it in the next section. Notice though that it has been designed ad-hoc, following the direction extracted from the cognitive theory of reputation by Conte and Paolucci [Conte and Paolucci, 2002].

```
1. Candidate_Seller := Select randomly one image's seller

2. If Candidate_Seller is empty or decided to risk then Candidate_Seller := select randomly
   one seller without image

3. Buy from Candidate_Seller
```

Figure 2.4: Buying action: Decision procedure for L1

```
1. Candidate_Seller := Select randomly one good enough seller image.

2. If Candidate_Seller is empty then Candidate_Seller := select randomly one good enough
   seller reputation

3. if Candidate_Seller is empty or decided to risk then Candidate_Seller := select randomly
   one seller without image

4. Buy from Candidate_Seller
```

Figure 2.5: Buying action - Decision procedure for L2

Buying action

For this action the question is: which seller should I choose? The Repage system provides information about image and reputation of each one of the sellers. The easiest option would be to pick the seller with *better* image, or (in L2) better reputation if image is not available. We set a threshold for an evaluation to be considered *good enough* to be used to make a choice. In addition, we keep a limited chance to explore other sellers, controlled by the system parameter *risk*[6]. Figures 2.4 and 2.5 describe the reasoning procedure that agents use to pick the seller in the situations L1 and L2 respectively. Notice that image has always priority over reputation, since image implies an acknowledge of the evaluation itself while reputation only an acknowledge of what is said. We can find however scenarios where this preference of image over reputation can be questioned.

Asking action

As in the previous action, the first decision is the choice of the agent to be queried, and the decision making procedure is exactly the same as for choosing a seller, but dealing with images and reputation of the agents as informers (*informer image*) instead of as sellers.

Once decided who to ask, the kind of question must be chosen. We consider only two possible queries: Q1 - Ask information about a buyer as informer (basically, how honest is buyer X as informer?), and Q2 - Ask for some good or bad seller (for instance, who is a good seller, or who is a bad seller?). Notice that this second possible question does not refer to one specific individual, but to the whole body of information that the queried agent may have. This is in order to allow for managing large numbers of seller, when the probability to choose a

[6] *Risk* is implemented as a probability (typically between 5% and 15%) for the buyer to try out unknown sellers

target seller that the queried agent have some information about would be very low. The agent will ask one of these two questions with a probability of 50%. If Q1 is chosen, buyer X as informer would be the less known one, that is, the one with less information to build up an image or reputation of it.

Answering action

Let agent S be the agent asking the question and R the agent being queried. R can lie, either because she is a cheater or because she is retaliating. When R is a cheater whatever information being answered is changed to its opposite value. Instead, retaliation is accomplished by sending inaccurate information, for instance, by sending "Idontknow" when really R has information, or simply giving the opposite value, like in the cheating case. In both cases, retaliation is done when R has a bad image as informer of S. In this case, in L1 condition R sends an "Idontknow" message even when she has information. Instead, in L2 condition R sends inaccurate reputation information. R converts the possible image to send to reputation, putting the opposite value. In this way, R avoids possible retaliation from S.

Because of the fear of retaliation, sending an image takes place only when an agent is very secure of the evaluation (reflected in the Repage *strength* parameter included in every evaluation). We include then the *thStrength* parameter, a threshold that allows to implement *fear of retaliation* in the agents. When *thStrength* is zero, there is no fear since whatever image formed will be a candidate to be sent, no matter its strength. As we increase *thStrength*, agents will become more conservative, less image and more reputation will circulate in the system.

Research Questions and Results

We have two experimental conditions, with image only (L1) and with both Image and Reputation (L2). We will explore several values of the parameters in order to show how and where there is an advantage in using reputation. The hypotheses are:

H1 Initial advantage: L2 shows an initial advantage over L1, that is, L2 grows faster. Intuitively, this would indicate that when reputation is present, agents are able to discover faster *good* sellers. This is related to the bootstrapping problem. Initially, agents have a maximum uncertainty, and need to interact or exchange information to model the behavior of sellers. The accomplish of this hypotheses indicates that when entering in a society with unknown partners, the exchange of reputation information helps agents reduce uncertainty in a faster way than without it.

H2 Performance: L2 performs better as a whole, that is, the average quality at regime is higher than L1. Note that to obtain this result we are hardwiring a limitation in image communication, based on the theory that foresees large amounts of retaliation against mistaken image communications but not on the reputation side. Intuitively, the hypothesis suggests that with reputation, agents adapt better to the unpredictable scenarios. We recall here that sellers disappear when they are out of stock.

We run simulations to examine the relationship between L1 and L2 with different levels in some parameters. The stock is fixed at 50, the number of buyers at 25, and the number of sellers at 100. We included cheaters as well with percentages of 0%, 25% and 50%.

We run the simulations for 100 steps, and explore the variation of good and bad sellers, from the extreme case of 1% of good sellers and 99% of bad sellers(A1), going trough 5% good sellers and 95% bad sellers(A2),and 10% good sellers and 90% bad sellers(A3), and finally, to another extreme where we have 50% of good sellers and 50% of bad sellers(A4). For each one of these situations and for every experimental condition (L1 and L2) we run 10 simulations. Figures show the accumulated average per turn of a concrete settings in both L1 and L2 experimental conditions.

Experiments without cheaters In figure 2.6 we show results for the four situations without cheaters. Both hypotheses H1 and H2 are verified. With the increase of good sellers the difference between L1 and L2 gets smaller to the point it disappears in situation A4. Because of the good sellers increase, they can be reached by random search and the necessity of communicating social evaluations decreases. In the extreme situation A4, statistically every buyer would find a good seller at the second turn (there is a probability of 50% to get one in one turn). In A3 the probability to reach one good seller per turn is 0.1, then, in 10 turns approximately every one would reach a good one. In L1 the amount of useful communications (different from *"Idontknow"*) is much lower than in L2, due to the fear of retaliation that governs this situation. In conditions where communication is not important (A4), the difference between the levels disappears.

Experiments with cheaters Figure 2.7 shows results for situations A1, A2, A3 and A4 with 50% of cheaters. The increased amount of false information produces a bigger impact in situations and conditions where communication is more important. Quality reached in L1 shows almost no decrease with respect to the experiment without cheaters, while L2 quality tends to drop to L1 levels. This shows how the better performance of L2 over L1 is due to the larger amount of information that circulates in L2.

Final Remarks

The results show that using reputation and image instead of only image improves the average quality reached in the whole system. These results should be considered as a proof of concept about the usefulness of the reputation model Repage[Conte and Paolucci, 2002], under a set of assumptions that we discuss with a perspective on future works:

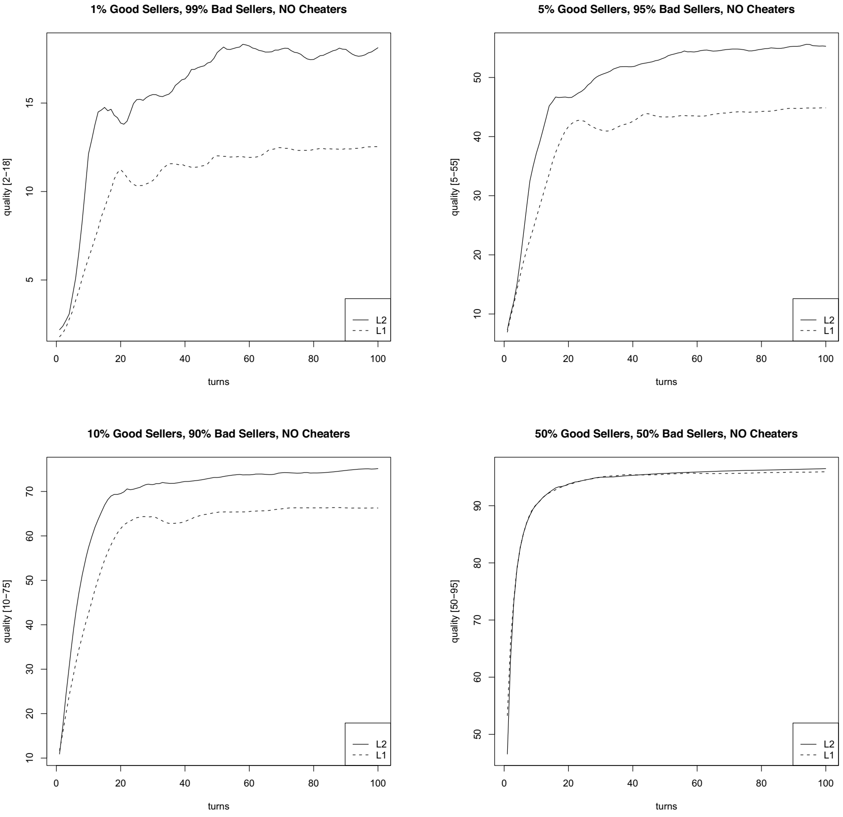

Figure 2.6: Accumulated average quality per turn without cheaters for situation A1, A2, A3 and A4 respectively

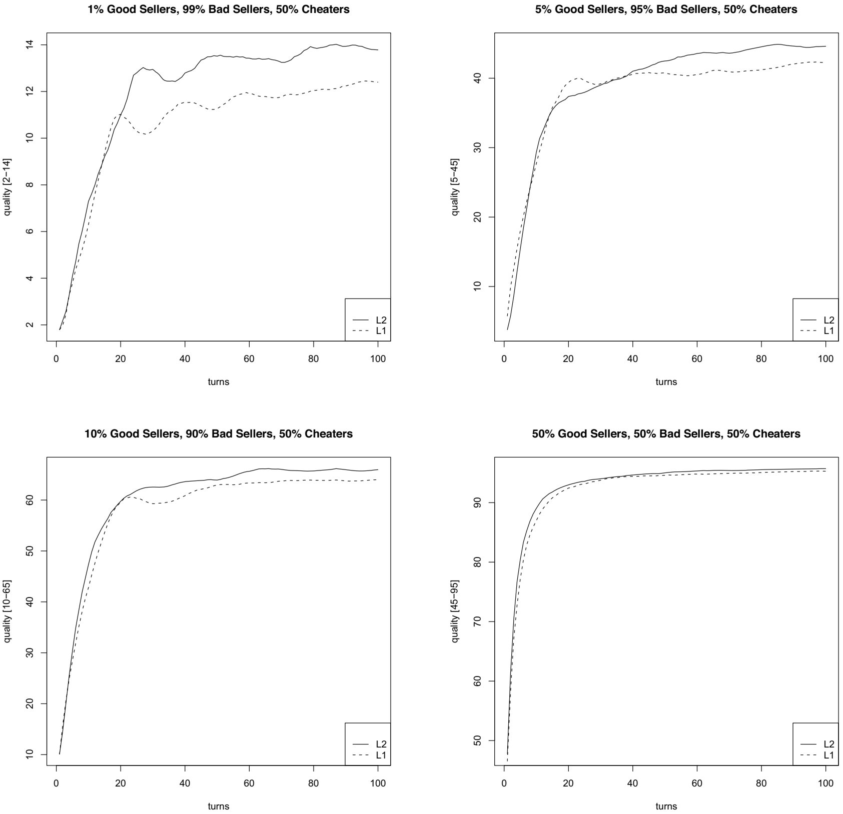

Figure 2.7: Accumulated average quality per turn with cheaters for situations A1, A2, A3 and A4 respectively

Retaliation: The presence of retaliation is crucial for the present results. We claim that the fact of communicating a social evaluation that is an image implies a commitment from the source agent. From the theory, image is a believed evaluation and sharing it implies that the source agent is informing of what he/she *accepts* as true. This differs from reputation, since accepting a reputation do not imply to accept the nested belief. Because of that, sharing what an agent acknowledges as a reputation does not imply a personal commitment. Here we assume that the personal commitment associated to image transmission exposes the agent to a possible retaliation if inaccurate information was sent.

Communication and reputation: There is no reputation without communication. Therefore, scenarios with lack of communication or few exchange of information cannot use reputation. However, in virtual societies with autonomous agents that have the freedom to communicate, that need to cooperate and have the right to choose partners, the separation between image and reputation considerably increases the circulation of information and improves the performance of their activities. In these experiments, even when there is no punishment for direct interactions and considering at each turn only one possible question, the introduction of this difference already improves the average quality per turn. In scenarios where *quality* is scarce and agents are completely autonomous is where this social control mechanism makes the difference.

Decision making procedure: The decision making schema determines the performance of the system. In fact, this is where the agent is taking advantage of the distinction between image and reputation. However, notice that the use of image and reputation from Repage has been determined ad-hoc, without any well-defined formalism. In our work we focus in this gap and propose (1) a BDI agent architecture fed by Repage to reason using social evaluations (dealing with pragmatic-strategic decisions) and (2) an argumentation framework to argue about social evaluations (facing memetic decisions).

As mentioned earlier, some other research has been conducted using the same simulation environment. In the work described in [W. Quattrociocchi, 2008b, W. Quattrociocchi, 2008a] the authors rely on studying the effect of cheating in the previously mentioned environment. The results are summarized in figure 2.8. It can be observed that the performance of the society in the L2 situation is better than in L1, even with high percentage of cheaters (around 60%). This result is very interesting because it shows that the massive circulation of inaccurate information[7] still produces benefits in the society even with a high rate of cheaters.

Notice that we are not talking about how from this elements agents are able to actually *trust* somebody, so, to reason using such information (pragmatic-strategic decisions). As mentioned in the first chapter, this is one of the contributions of this book. Also, the model does not say anything about the memetic decisions. The other contribution of this work covers this aspect.

[7] Notice that the agents send reputation when the reliability (or strength) of image predicates is not higher than certain threshold.

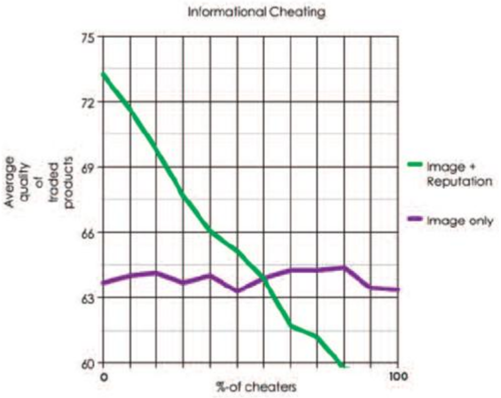

Figure 2.8: Average quality obtained by the buyers against the percentage of cheaters. It can be observed that when such percentage is lower that 60% the use of image and reputation achieves better accuracies than only using image. Source:[Paolucci et al., 2009].

2.5 Review on Computational Trust and Reputation Models

We have exposed so far the bases of our work: (1) The cognitive theory of reputation proposed by Conte and Paolucci [Conte and Paolucci, 2002], and (2) the computational reputation model Repage [Sabater-Mir et al., 2006]. Furthermore we have explored the bases of social trust in relation with reputation and image. In this section we provide a non-exhaustive but representative vision of the computational trust and reputation systems that according to us are relevant in the current state-of-the-art. Instead of providing detailed descriptions for each model, we summarize their main characteristics and background ideas, showing different dimensions of analysis that current surveys on computational reputation models have provided. Moreover, we provide four new classification dimensions that have not been directly treated so far and that enhance part of the contributions of this work.

Many surveys exist in literature and along with them, different dimensions to classify and characterize reputation models. Some of them are based on online trust and reputation related systems [Jsang et al., 2007a, Grandison and Sloman, 2000, Artz and Gil, 2007, Grabner-Kruter and Kaluscha, 2003], others on trust and reputation in peer-to-peer systems [Koutrouli and Tsalgatidou, 2006, Suryanarayana and Taylo, 2004]. Some reviews focus on concrete aspects or functionalities like attack and defense techniques [Hoffman et al., 2007] or reputation management [Ruohomaa et al., 2007]. Others are more general [Sabater and Sierra, 2005, Herzig et al., 2008, eRep, 2007, Lu et al., 2007, Mui et al., 2002a].

2.5.1 Dimensions of Analysis: Current View

We examine the most representative classification dimensions proposed in literature that, from our perspective, clarify the general view of reputation and trust models present in literature. All the models that appear in the following classifications are briefly described in sections 2.5.3 and 2.5.4.

Sabater *et al.*'s Classification Dimensions

The first classification dimensions we introduce are those defined by Sabater *et al.* [Sabater-Mir, 2003, Sabater and Sierra, 2005]. This is one of the most used classification in the current literature, and has been used as a base in other reviews. The proposed dimensions give a rather general view of the characteristics that reputation and trust models achieve. Even though nowadays the classification could be extended and refined, it is still a good starting point. These dimensions are:

Paradigm Type: Following [Sabater and Sierra, 2001], models are classified as *cognitive* and *numerical*. The former refers to models in which the notion of trust or reputation is built on beliefs and their degrees. Also, in these models how trust and reputation values are calculated can be as relevant as the final value. The model of social trust presented by Castelfranchi and Falcone [Castelfranchi and Falcone, 1998b] and the Repage model [Sabater-Mir et al., 2006] are good examples. On the opposite, the numerical paradigm includes models that use game theoretical approaches.

Information Sources: Models can be also classified depending on the information sources they use to infer trust or reputation:

- Direct experiences are one of the most valuable sources of information for the agents. The author differentiates between direct interactions (DI) and direct observations (DO).

- Witness information (WI) is information gathered from other agents. Even though this particular item could be extended with a full typology of witness information, like third-party observations, third-party interactions, reputation communication etc., the work remains at this level.

- Sociological information (SI) is based on the analysis of social relations among the agents, and can be computed through social network analysis if relational data is available [Sabater and Sierra, 2002].

- Prejudice (P) is an information source that allows bootstrapping of trust and reputation when no other information is available, and that coincide with the human notion without account for the negative connotation. Stereotyping is a very related notion that has been also used in this sense.

Visibility: From this dimension, the trust/reputation information of an agent can be considered a global property that all other agents can observe (G), or can be considered a private and subjective property that each agent builds (S). This dimension is one of the most popular and has been used also in multiple surveys, in particular, in [eRep, 2007], which refine this dimension. We explain it in detail in this section.

Online reputation models fit perfectly in the global category, while reputation and trust models that are part of individual agent's architectures are considered subjective. Often this dimension is categorized as centralized and distributed models.

Granularity: It refers to the context-dependence of trust/reputation information. Some models consider that trust is associated to a concrete context. For instance, it is not the same to trust somebody to drive a car than to play a good soccer game. In general, single-context models can be considered as a particular case of multi-context ones, because the context is implicit in the environment. Online reputation models are a good example of single-context models.

Cheating Behavior: This dimension explores the models' assumptions regarding the behaviour of communicating agents. Three levels are defined:

- Level 0: Cheaters are not considered. Hence, third-party information comes from honest agents that, in the case they send false information is because they are also mistaken.

- Level 1: Agents can hide or bias communicated information, but never lie.

- Level 2: Cheating is considered.

Type of Exchanged Information: Sabater in [Sabater and Sierra, 2005] considers that models that assume exchange of information can be separated in two groups. Those that send boolean information and those that use continuous measures. Nowadays this dimension can be also analyzed in much more detail.

Table 2.2 shows the summary of the models reviewed in [Sabater and Sierra, 2005] classified in the dimensions explained above. The table also includes a column indicating whether the model uses reliability measures for trust and reputation values, and the type of model that the authors of the models claim to be (trust or reputation). Furthermore we include the Repage model and the BDI+Repage model, which originally were not present. We classify Repage as a numerical and cognitive model, since it uses elements of both approaches. It uses both witness information and direct interaction. It is a subjective contextual model that tolerates cheating. The exchanged information can be communicated images and reputations. It offers reliability measures and it is considered a reputation model. Instead, BDI+Repage model, according to Sabater's definition, should be considered a trust and reputation model.

This classification establishes the bases for some of the most recent reviews, and in particular, for the next one we present in this chapter.

Par - Paradigm	**N** - Numerical **C** - Cognitive
InS - Information Sources	**DI** - Direct Interaction **DO** - Direct Observation **WI** - Witness Information **SI** - Sociological Information **P** - Prejudice
Vis - Visibility	**S** - Subjective **G** - Global
Gra - Granularity	**CD** - Context Dependent **NCD** - Non Context Dependent
Che - Cheating Assumptions	**L0** - No cheating **L1** - Bias information **L2** - Cheating
Rel - Reliability Measure	✓, –
Type - Model Type	**T** - Trust **R** - Reputation

Table 2.1: Legend for the table 2.2. Source: [Sabater and Sierra, 2005]

Model	**Par**	**InS**	**Vis**	**Gra**	**Che**	**BoE**	**Rel**	**Type**
Marsh	N	DI	S	CD	–	–	–	T
eBay	N	WI	G	NCD	0	–	–	R
Sporas	N	WI	G	NCD	0	–	✓	R
Histos	N	DI+WI	S	NCD	0	–	–	R
Schillo *et al.*	N	DI, DO WI	S	NCD	1	✓	–	T
Rahman & Hailes	N	DI, WI	S	CD	2	4 val	–	T,R
Esfandiary *et al.*	N	DI, DO, WI, P	S	CD	0	–	–	T
Yu & Singh	N	DI, WI	S	NCD	0	–	–	T,R
Sen & Sajja	N	DI, DO, WI	S	NCD	2	✓	–	R
AFRAS	N	DI+WI	S	NCD	2	–	✓	R
Carter *et al.*	N	WI	G	NCD	0	–	–	R
Castelfranchi *et al.*	C	-	S	CD	–	–	–	T
Regret	N	DI+WI+SI+P	S	CD	2	–	✓	T,R
Repage	*C/N*	*DI+WI*	*S*	*CD*	2	–	✓	*R*
BDI+Repage	*C/N*	*DI+WI*	*S*	*CD*	2	–	–	*T, R*
ForTrust	*C*	–	*S*	*CD*	–	–	–	*T, R*
Rasmusson& Jason	*N*	*WI*	*G*	*NCD*	2	–	–	*R*
Regan & Cohen	*N*	*DI + WI*	*S*	*NCD*	2	–	–	*T*
Padovan et al.	*N*	*DI + WI*	*S*	*NCD*	0	–	–	*R*
Ripperger	*N*	–	*S*	*NCD*	–	–	–	*T*
LIAR	*N*	*DI + DO*	*S*	*NCD*	2	–	✓	*T, R*
FIRE	*N*	*DI + WI*	*S*	*CD*	0	–	✓	*T, R*
Mui et al.	*N*	*DI, WI*	*S*	*CD*	0	–	–	*R*
Dirichlet	*N*	*WI*	*G*	*NCD*	0	–	–	*R*
Sierra & Debenham	*N*	*DI + WI + SI*	*S*	*NCD*	0	–	–	*T*

Table 2.2: Summary of models' characteristics defined by Sabater-Mir & Sierra (2005)'s dimension. The models in *italic* were not present in the original work. Legend on table 2.1

Balke et al.'s Classification

The classification presented by Balke *et al.* in [Balke et al., 2009] focuses on the five stages process that, from the authors' perspective, exists in reputation and trust models between transactions. According to them, when the transaction i is produced, there is first a recording of cooperative behavior, followed by a ranking and storage stage. The fourth stage refers to the recall for cooperative behavior, concluding with an adaptation or learning of the strategy as the fifth stage. After these five stages, transaction $i + 1$ can proceed.

The authors define several possible behaviors for each stage, generating then a taxonomy of models. In the following lines we briefly detail each one of the stages.

Recording of Cooperative behavior: The first stage deals with the recording of the transaction, and for this, models must be aware of the contextuality of the evaluations. Hence, the authors argue that in this stage models can be considered multi-context or single-context, coinciding with the granularity dimension defined by Sabater (see previous section).

Rating of Cooperative Behavior: After the transaction is recorded, it must be rated and incorporated into the system. In this stage the authors differentiate between pure game theoretical approaches where trust/reputation is considered as a subjective probability (in the sense defined by Gambetta [Gambetta, 1990]), and more cognitive approaches where a new rate affect the mental state of the agent.

The former approach is usually based on aggregation functions that summarize final values of trust/reputation, being them the important element. Instead, even when cognitive approaches may use as well aggregation functions, the information is processed in intermediate steps that can be as important as the final values and that affect a whole mental state of the agent. This classification is related to the paradigm dimension defined by Sabater (see previous sections). At the end of this chapter we get into more details about these differences.

Storage of cooperative behavior: The third stage refers to the storage of the rated information. According to the authors, the information can be stored by the same agent (distributed) of by a global third-party (centralized). This is related to the visibility dimension defined by Sabater, but from our point of view the present one indicates a more representative and descriptive dimension. In this classification it is more clear that when agents store their own rated information, trust/reputation values are considered subjective, while in the centralized approach it must be a global property, because they are publicly seen by all the members of the society.

Recall of Cooperative Behavior: This stage focuses on the information used by the models to calculate/infer trust and reputation. In this stage, the authors intend to define a layered categorization, where the first dimension is whether the model considers somehow witness information or not. According to the authors, the latter are usually considered trust models, while the others, since they require the exchange of information, reputation models. While this differentiation is very questionable and somehow contrary to our vision of trust

Recording	**SC** - Single-context model **MC** - Multi-context model
Rating	**C** - Cognitive base **MA** - Mathematical
Storage	**CS** - Centralized **DS** - Distributed
Recall	**T** - Trust Model **RE** - Reputation Model
Cheating	**L0** - No cheating **L2** - Cheating

Table 2.3: Legend for classification of table 2.4

and reputation, it is interesting to show the different conceptions existing in the literature.

The authors also examine the kind of exchanged information, and the assumptions regarding the behavior of source agents, as in Sabater's assumptions of level 0, 1 and 2.

Learning/Adaptation Strategy: The last stage relies on the final decision, on how to use all the previous information to actually adapt the agent's behavior for future interactions. From our view, this refer to the pragmatic-strategic decisions stated by Conte and Paolucci in [Conte and Paolucci, 2002]. The authors argue that in fact, this stage cannot be classified because the surveyed models do not offer clear strategies due to their high context-dependency. We do not completely disagree with this idea. Some models offer evaluative calculus, degrees of trust or reputation that do not necessary indicate how to use them. Some models though have it implicit, while others, rely on the decision making of the agents. We analyze this aspects in our classification described at the end of this chapter.

Table 2.4 shows the summary of the surveyed models classified according to these dimensions.

eRep Project's Classification

The European project *eRep: Social Knowledge for e-Governance* [eRep, 2006b] aimed at providing both theoretical and empirical guidelines for the design and use of reputation technology. Their first deliverable [eRep, 2007] describes an interesting survey of computational reputation and trust models, and classify them in four different well-defined categories:

1. Agent-Oriented *Solitary* Approaches (AO Sol): In this approach, the agent itself calculates the evaluations regarding other agents taking only into account its own previous experiences. There is no exchange of information. This category corresponds to (1) the definition of trust model of the fourth stage of the Balke's classification (see previous subsection), and (2) the combination of the information source and visibility dimensions of

Model	**Reco**	**Rat**	**Sto**	**Reca**	**Cheat**
Marsh	MC	MA	DS	T	-
Schillo *et al.*	SC	MA	CS	RE	L2
Rasmusson & Jason	SC	MA	CS	RE	L2
Abdul-Rahman & Hailes	MC	MA	DS	RE	L0
Regan & Cohen	MC	MA	CS/DS	RE	L2
Sporas	SC	MA	CS	RE	L0
Histos	SC	MA	DS	RE	L2
Yu & Singh	SC	MA	DS	RE	L2
Padovan *et al.*	SC	MA	CS/DS	RE	L0
Foner	SC	MA	CS	RE	L0
Regret	MC	MA	DS	RE	L2
Repage	SC	CO	DS	RE	L2
eBay	*SC*	*MA*	*CS*	*RE*	*L0*
BDI+Repage	*MC*	*CO*	*DS*	*RE*	*L2*
Sen & Sajja	*MC*	*MA*	*DS*	*RE*	*L2*
Esfandiary et al	*SC*	*MA*	*DS*	*RE*	*L0*
AFRAS	*MC*	*MA*	*DS*	*RE*	*L2*
Carter et al	*MC*	*MA*	*CS*	*RE*	*L0*
Ripperger	*MC*	*MA*	*DS*	*T*	–
ForTrust	*MC*	*CO*	*DS*	*T*	–
FIRE	*MC*	*MA*	*DS*	*RE*	*L0*
LIAR	*SC*	*MA*	*DS*	*RE*	*L2*
Castelfranchi & Falcone	*MC*	*CO*	*DS*	*T*	–
Mui et al.	*MC*	*MA*	*DS*	*RE*	*L0*
Dirichlet	*SC*	*MA*	*CS*	*RE*	*L0*
Sierra & Debenham	*MC*	*MA*	*DS*	*T*	*L0*

Table 2.4: Classification of the models developed by Balke *et al.* (2009). The models in *italic* were not present in the original work. Legend on table 2.3.

Sabater's classification, taking DI (direct interaction) and S (subjective) as values respectively.

2. Agent-Oriented *Social* Approaches (AO Soc): In this category, agents themselves also calculate the evaluations but they may also rely on third-party information. Hence, there must be exchange of information. Regarding Sabater's classification, this corresponds as well to the combination of the information source and visibility dimension, where the latter is set to S (subjective) and the former to WI (witness information) plus (may be) other sources. Regarding Balke's classification, the equivalence embraces again the fourth stage with their definition of *reputation* model.

3. *Objective* External Evaluation Agencies (Obj EEA): In contrast to agent-oriented approaches where agents recollect the information and evaluate themselves other agents, external agencies can compute such evaluations according to certain criteria. This category covers models that compute the evaluations through objective facts, like well-defined quality standards for instance. Not many models fit into this category.

4. *Subjective* External Evaluation Agencies (Sub EEA): In this case, overall evaluations are performed as well in an external agency, but in this case the result is the aggregation of the subjective agents evaluations collected by the system. Online reputation systems perfectly fit in this category. For instance, in eBay users rate their individual experiences with the sellers by giving a positive, negative or neutral point. Then, the eBay system collects the rates and issues an overall punctuation for each seller, in this case by simply summing all the rating scores and presenting it with a system of colored stars.

This classification mainly focuses on two big dimensions, agent-oriented and external evaluation agencies. This division corresponds to the visibility dimension of Sabater, and the storage stage in Balke *et al.*'s classification. The work done fits into the agent-oriented social approach category.

Summarizing, table 2.5 shows the classification of the models reviewed in [eRep, 2006b]. Models that fit into two or more categories indicate that the description of the models does not quite determine the approach, and that in principle, they could be considered in all the marked categories.

2.5.2 Yet another Classification

We could not finish the chapter without including our own classification dimensions that as far as we know, have not been considered yet in detail in the current state-of-the-art surveys, and that frame very well this work. For the nature of the work, this classification only faces distributed models, or in terms of the survey in [eRep, 2007], agent-oriented approaches. The dimensions we define here are the following:

Model	**AO Soc**	**AO Sol**	**Obj EEA**	**Sub EEA**
Abdul-Rahman & Hailes	✓	–	–	–
Kuhlen	–	–	✓	–
Marsh	–	✓	–	–
Padovan *et al.*	✓	✓	✓	✓
Rasmusson & Jason	✓	✓	✓	✓
Rasmusson's Reviewer Ag.	–	–	✓	–
Regan & Cohen	✓	–	–	–
Regret	✓	–	–	–
Repage	✓	–	–	–
Ripperger	–	✓	–	–
Schillo *et al.*	✓	–	–	–
Zacharia *et al.*(SPORAS, HISTOS)	✓	–	–	✓
eBay	–	–	–	✓
LIAR	✓	–	–	–
FIRE	✓	–	–	–
Mui et al.	✓	–	–	–
Yu & Singh	✓	–	–	–
BDI+Repage	✓	–	–	–
ForTrust	–	✓	–	–
Castelfranchi & Falcone	–	✓	–	–
Dirichlet	–	–	–	✓
Carter et al.	–	–	✓	✓
AFRAS	✓	–	–	–
Sen & Sajja	✓	–	–	–
Sierra & Debenham	✓	–	–	–

Table 2.5: Classification of the models according to the eRep project. The models in *italic* were not present in the original work.

Trust Dimension

We do not want to differentiate between models classified as *trust* and others as *reputation*. We strongly believe that the distinction between both *kinds* of models does not rely on a clear consensus in the community. For instance, the *type* dimension that Sabater provides in his classification is not based on any objective fact, but on what the authors of the models claim [Sabater-Mir, 2003].

On the contrary, when facing these concepts from a more cognitive perspective, the distinction becomes clearer, at least in some aspects. From the concept of social trust [Castelfranchi and Falcone, 1998b], occurrent and dispositional trust [Herzig et al., 2008, ForTrust, 2009] and pragmatic-strategic decisions pointed out by Conte and Paolucci in [Conte and Paolucci, 2002], we can deduce that *trust* implies a decision. As mentioned before, trust can be seen as a process of practical reasoning that leads to the decision to interact with somebody. Regarding this aspect, some models provide evaluations, rates, scores etc. for each agent to help the decision maker with a final decision. Instead, others specify how the actual decision should be made. From our point of view, only the latter cases can be considered trust models. We recall here that in this case, the decisions are also pragmatic-strategic, in the sense described in [Conte and Paolucci, 2002].

Table 2.6 summarizes the models that from our definition should be considered trust models. We mark them with '✓'. For instance, the model defined by Marsh [Marsh, 1994] is a trust model because it indicates exactly to whom to interact with. The final decision is made through a well-defined threshold. Another example is the model defined by Sen & Sajja [Sen and Sajja, 2002b]. Even when this model is usually considered a reputation model, the fact is that it defines a decision making process that identifies to whom to interact with, and then, fits in our definition of trust.

Models marked with '$-$' are those that we do not consider trust models. They calculate measures or evaluations to help a decision making process. For instance, the AFRAS model [Carbo et al., 2002b, Carbo et al., 2002a] gives evaluations in terms of fuzzy sets, and the shape of these fuzzy numbers also determines a reliability measure. However, there is no mechanism that tells the agent how to use such evaluations. This situation is similar as in the Repage model. As explained before the model only gives support to the creation of image or reputation predicates, social evaluations.

Finally, we use the mark $\sim$ to indicate that the model does not give an explicit decision mechanism, but that it is rather dependent on the current desires of the agent. For instance, the Regret model [Sabater and Sierra, 2001] provides for each agent and context a *trust* value, together with a reliability measure. The trust value is calculated through aggregation of the information from several sources. One of the sources is defined by an ontology, which already determines which information is considered more important[8]. Hence, the goals of the agent are somehow codified in this ontology, and the final trust value obtained is

[8] For instance, to calculate the trust of agents as sellers, the ontology can define that this is evaluated through the price in an 80% and through the delivery time in a 20%

an indicator of which possible target agent matches better with the desires of the agent. However, since it offers a reliability measure the decision is not yet possible. For instance, lets assume that agent a has a trust value of 0.6 with a reliability of 1. On the other hand, another agent b has a trust value of 0.8 with a reliability of 0.4. Which is the best option? It still requires a decision making process. However, it is clear that with similar reliability measures, the agent with highest trust value is the chosen one. FIRE model [Huynh et al., 2006a] shows a similar situation.

Cognitive Dimension

Although this dimension has already appeared in other surveys, the provided definitions are quite vague. In this dimension we differentiate models that have clear representations of trust, reputation, image etc. in terms of cognitive elements such as beliefs, goals, desires, intentions, etc. From our perspective, models that achieve such representation explicitly describe the epistemic and motivational attitudes that are necessary for the agents to have *trust* or to hold social evaluations. From a human point of view, this allows for a better understanding of the internal components of trust and reputation, and for a clear implication to possible final decisions. From a software agents perspective, this endows the agents with a clear capacity to *explain* their decisions and to reason about the trust structure itself, making a metareasoning [Castelfranchi and Paglieri, 2007] possible. In this sense, in models that achieve a cognitive representation, final values of trust and reputation are as important as the structure that supports them. These models are usually very clear at the conceptual level, but lack in computational aspects.

Often, models that are not endowed with this property consider the model as a black box that receives inputs and issues trust and reputation *values*. Because of that, the internal calculation process cannot be considered by the agent, only the final values. Moreover, the integration with the other elements of the agent remains unclear because motivational attitudes are assumed or mix with the calculus. However, their computational aspects are usually quite well defined.

In table 2.6 we show the summary of the reviewed models against this dimension. We marked with '✓' the ones with such property, and '$-$' the lack of it. We mark the Repage model with '$\sim$' because the internal structure is based on predicates that have associated cognitive notions, but it does not have an explicit representation of them. In fact, Repage integrates into first-order like predicates, mixing also epistemic and motivational attitudes. The model presented in this book, the BDI+Repage model, makes explicit these missing cognitive components.

Procedural dimension

Often, models offer a nice way to represent and deal with trust and reputation, but there is no explanation on how they bootstrap. This is quite common in cognitive models, which focus on the internal components of trust and reputation, but not how such components are built. Also though, some non-cognitive models do not give explicit details on the calculus of their evaluations. We must recall here that we focus on the epistemic decisions, not on the creation and combination of motivational attitudes (goal-based).

The model introduced by Castelfranchi and Falcone [Castelfranchi and Falcone, 1998b] regarding social trust does not give details on how the beliefs are created. ForTrust model [Herzig et al., 2008, ForTrust, 2009] redefines the notion of social trust and introduces cognitive reputation but still epistemic decisions remain unclear. On the contrary, models like AFRAS [Carbo et al., 2002b, Carbo et al., 2002a] and Regret [Sabater and Sierra, 2002, Sabater-Mir, 2003] describe until the last detail how evaluations are created and how they are aggregated.

We mark Marsh [Marsh, 1994] and Abdul-Rahman *et al.* [Abdul-Rahman and Hailes, 2000] models with '∼' to indicate that in general they provide all the calculations, but do not provide some ground calculations. For instance, in the former, the model does not indicate how direct interactions are evaluated. The author indicates that this is left open and dependent of the context (and we totally agree with it). The same happens with the latter model.

Generality dimension

The last dimension we want to analyze refers to the generality of the model. In this dimension we want to classify the models that have a general purpose '✓' versus the ones that focus on very particular scenarios '−'. For instance, the model by Abdul-Rahman *et al.* [Abdul-Rahman and Hailes, 2000] is a non-general model that focuses on the trust on the information provided by witness agents. The same happens with the model by Yu & Singh [Yu and Singh, 2003], which is designed for agents participating in a very structured peer-to-peer network, where evaluations are only done in terms of quality of services. Obviously, the models that have such specification obtain good results and very acceptable computational complexities.

On the contrary, models built for general purposes can be adapted to multiple scenarios and are perfect then for general agent architectures. Regret [Sabater and Sierra, 2001, Sabater-Mir, 2003] and BDI+Repage model [Pinyol and Sabater-Mir, 2009a] are good examples of such models. Again, table 2.6 summarizes in the last column this property against the surveyed models.

2.5.3 Centralized Approaches

In this section we review the reputation and trust models that we classify as centralized and that appear in the reviews above. We remark that in the description of the models, the terms *trust* and *reputation* correspond to the view that the respective authors provide in their articles. Therefore, they may not coincide with the notions we describe in our work.

Model	Trust	Cognitive	Procedural	Generality
Abdul-Rahman *et al.*	–	–	~	–
AFRAS	–	–	✓	✓
Castelfranchi *et al.*	✓	✓	–	✓
Esfandiari *et al.*	–	–	✓	✓
FIRE	~	–	✓	✓
ForTrust	✓	✓	–	✓
Marsh	✓	–	~	✓
Mui *et al.*	✓	–	~	✓
LIAR	✓	–	✓	–
Regret	~	–	✓	✓
Regan & Cohen	✓	–	✓	–
Repage	–	~	✓	✓
Ripperger	✓	–	✓	–
Schillo *et al.*	–	–	✓	✓
Sen & Sajja	✓	–	✓	–
Yu & Singh	✓	–	✓	–
Sierra & Debenham	✓	–	✓	✓
BDI+Repage	✓	✓	✓	✓

Table 2.6: Computational Models against our classification dimensions.

Online reputation models

These models are used in e-commerce sites such as eBay [eBay, 2002], Amazon [Amazon, 2002] and OnSale [OnSale, 2002] among others. These sites work as market places where buyer users buy products from seller users. After a transaction is done, the buyer has the possibility to *rate* the seller, so, to give its opinion about it. In eBay for instance, users have three possibilities, *positive*(1), *neutral*(0) or *negative*(-1). The value of the sum of all the rates is the reputation value, that is public to everybody. eBay presents these results with a system of colored stars.

These systems have a very simple implementation and offer very intuitive representations, making them ideal for human-based applications. However, they lack in robustness; no reliability measures, no consideration of false information or cheating, no temporal issues[9], and in general, lack of the main characteristics that make special each one of the following models.

Sporas and Histos

Sporas and Histos [Zacharia, 1999] are a natural evolution of the online models. The idea is very similar, but in this case, only the most recent feedbacks are considered. Moreover, the aggregation function is not just the sum. It has

[9] For instance, feedbacks remain countable for ever. So, a seller with very high reputation value could start acting as a bad seller without having an immediate effect on its reputation value.

been designed to produce small rating changes for users with very high reputation, and bigger rating changes for users with lower reputation. They also incorporate a measure of the reliability of the users' reputation. Histos incorporates a data structured similar to the trust net used in the Schillo *et al.*'s model [Schillo et al., 1999, Schillo et al., 2000]

Carter *et al.*

The underlying idea of the model introduced by Carter *et al.* [Carter et al., 2002] states that the reputation of an agent is the degree of fulfillments of roles ascribed to it by the society [Sabater and Sierra, 2005]. They state that each society defines the set of roles that a participant can play, and that the reputation of each participant is the result of a weighted aggregation of the fulfillments achieved by the agent on each role. Because of that, for them it is not possible to find a universal way to calculate reputation, since it needs to be in a context of such a society. The value is calculated by a central authority who controls all the transactions.

Kuhlen

The model presented by Kuhlen [Kuhlen, 1999][10] does not come from the area of multiagent systems, but from economics, facing trust management issues to make electronic commerce more reliable. The author's idea considers a trusted third-party agency that objectively evaluates certain quality standards that e-Commerce sites should be endowed with, issuing a certified seal that could be posted in the e-Commerce web place. The important point here is that there exist an implemented version for issuing such certificate based on objective quality measures. It has not been applied to multi-agent systems yet, but the idea should work as well.

Padovan *et al.*

The model introduced by Padovan *et al.* [Padovan et al., 2002] uses a combination of agent oriented approaches and external approaches. The model is designed over the platform Avalanche [Eymann, 2000], an agent-based coordination mechanism for electronic marketplaces. Again, the focus is the e-Commerce. The approach suggests the use of domain-specific rating agents capable to provide reputation information to the buyers. These agents act as external agencies that are able to evaluate transactions in an objective way. Single agents are endowed with certain goals that when they match with the specific reputation information, can be used in their strategies to select partners.

We include the model in this category because in fact, individual agents do not compute reputation, but they query external special entities. Then reputation is seen by the agents as a centralized property. They build their trust

[10] We extracted the explanation of this model from [eRep, 2007] and [trustProject, 2000] because the original article is in German

based on such information. In this sense then, the model could be considered also agent-oriented.

Dirichlet Reputation Systems

This family of reputation systems [Jsang et al., 2007b] works very well in centralized environments where users' ratings are based on a discrete and finite sorted set, for instance, {*very bad, bad, neutral, good, very good*}. These models are capable of giving a probability distribution on this sorted set, representing the probability that the agent has to act as stated in each one of the categories. For example, a seller that just starts selling has a reputation value totally unknown. So, the probability distribution over the sorted set will be $(.2, .2, .2, .2, .2)$. If she is a good seller, users will rate her with good punctuation. So, after a while her reputation value could be $(0, 0, .1, .3, .6)$.

To do so, these models use Dirichlet probability distribution, a multinomial Bayesian distribution. The idea is to approximate the set of evidences (users' rates) to the appropriate Dirichlet distribution and then, extrapolate the value of each category. If we had 2-valued evidences (for instance, {*bad,good*}) and considering evidences as Bernoulli experiments, we could approximate the situation to a binomial distribution. If evidences are multi-valued, we need a multinomial distribution, and Dirichlet distributions seem a good option.

2.5.4 Agent-Oriented Approaches

In this section we show a set of models that share the characteristic of considering reputation or trust as subjective properties.

A-Rahman and Hailes

This model [Abdul-Rahman and Hailes, 2000] uses the term *trust*, and its main characteristic relays on that evaluations are represented with a discrete set of four elements. The model is fed by two sources: direct experiences and third party communications of direct experiences. The representation of the evaluations is done in terms of the discrete set {*vt (very trustworthy), t (trustworthy), u (untrustworthy), vu (very untrustworthy)*}. Then, for each agent and context the system keeps a tuple with the number of past own experiences or communicated experiences in each category. For instance, agent A may have a tuple of agent B as a seller like $(0, 0, 2, 3)$, meaning that agent A has received or experienced 2 results as untrustworthiness and 3 as very untrustworthiness. Finally the *trust* value is computed taking the maximum of the tuple values. In our example for agent A, agent B as a seller would be very untrustworthy. In case of tie between *vt* and *t* and between *u* and *vu* the system gives the values U^+ (mostly trustworthy) and U^- (mostly untrustworthy) respectively. In any other tie case the system returns U^0 (neutral).

AFRAS

The model presented by Carbo *et al.* [Carbo et al., 2002b] uses fuzzy sets to represent reputation values. The idea is that the latest interaction that an agent has with a partner, that is also valued as a fuzzy set, updates the old fuzzy set reputation value through a weighted aggregation. To calculate the weights, they introduce the *remembrance for memory*, a factor that allows the agent to give more weight to the latest interaction or to the old reputation value. The novelty of this approach relies on the reliability of the reputation value, since it is intrinsically represented in the fuzzy set. So, a wide fuzzy set for a reputation value indicates a high level of uncertainty, meanwhile narrow ones, implies more reliability.

The model also deals with the recommendations sent by other members of the society. The recommendations are aggregated together with the direct interactions. The level of reliability of this witness information will depend on the good or bad reputation of the senders. In this case then, recommendations from a very well reputed sender could have the same weight as direct interactions.

Castelfranchi & Falcone and ForTrust

Both models are explained in detail in section 2.3.

ReGreT

The ReGreT system presented by Sabater [Sabater and Sierra, 2001] is maybe one of the most complete reputation and trust models, since it takes into account several advantages of all the models presented so far.

ReGreT uses direct experiences, third party information and social structures to calculate trust, reputation and levels of credibility. In this model, trust is a function of direct trust, only calculated through direct experiences, and reputation. The incorporated reputation model uses transmitted information, social networks analysis, system reputation and prejudices (to infer reputation values of unknown agents from their belonging group). It also incorporates a credibility module to evaluate the truthfulness of witness information, that of course, takes into account the reputation and trust of the information provider. It provides reliability measures for trust, reputation and credibility values.

Finally, an important aspect of this model is the consideration for an ontological dimension. They defined the trust of agent a on b towards certain context φ as $T_{a \to b}\varphi$. The situation φ is totally contextualized, and may depend on other elements. To describe the relationships of contextualized environment, it is assumed an ontology that describes this knowledge, that could be seen as the current *preferred desires* or *goals* of the agent.

Esfandiari *et al.*

Esfandiari & Chandrasekharan defined a model [Esfandiari and Chandrasekharan, 2001] where trust evaluation considers different sources, although no information is provided on how to combine them for a final choice. A *first trust* is based on observations, and it is calculated using Bayesian networks. A *second trust* is based on interactions. For the latter, agents use an exploratory protocol to ask other agents about how to evaluate the degree of trust, and a query protocol to ask for recommendations from trusted agents. The model builds a trust net as a directed graph to deal with the received witness information.

An interesting point of the model relies on the labeling of the edges. Instead of using a single value to determine the trust degree of an agent, the model uses intervals with the minimum an the maximum values received in all paths. By considering colored labels the model can deal with trust on different properties of the agents. Finally, the model also considers institutionalized trust (system reputation in Regret). As mentioned before, no decision making mechanism is specified.

FIRE

The FIRE model introduced by Huynh *et al.* [Huynh et al., 2006a] incorporates similar elements than Regret. It computes as well a *trust* value for each agent and a reliability measure. It uses direct trust computed though direct experiences (extracted from Regret as the same authors claim), witness information (similar to Regret) and certified reputation. The last one is a completely new component. Certified reputations *are ratings presented by the rated agent about itself which have been obtained from its partners in past interactions*[Huynh et al., 2006a]. The authors argue that this could be seen as the recommendation letters or references when applying for a job position.

The model uses role-based trust to determine the elements that contribute to the calculation of trust. This component is similar to the ontology dimension of Regret. Therefore, they can be seen as the desires (or goals) of the agent.

Marsh

This model [Marsh, 1994], one of the first that appeared in the literature, talks explicitly about trust, and only takes into account direct experiences. It defines three kinds of trust.

- **Basic Trust**: T_x^t represents the trust disposition of agent x at time t.
- **General Trust**: $T_x(y)^t$ represents the general trust that agent x has on y at time t without specifying any situation.
- **Situational Trust**: $T_x(y, \alpha)^t$ represents the trust of agent x on the target agent y in the situation α. Marsh defines a basic formula to calculate it:

  $$T_x(y, \alpha)^t = U_x(\alpha)^t \cdot I_x(\alpha)^t \cdot \overline{T_x(y)^t} \tag{2.4}$$

  where $U_x(\alpha)^t$ is the utility that agent x gains from situation α, $I_x(\alpha)^t$ is the importance for agent x in the situation α, and $\overline{T_x(y)^t}$ is the estimation of general trust after taking into account all information related to $T_x(y)^t$. The author proposes three ways to calculate this estimation: the mean, the maximum and the minimum of all past experiences.

Yu and Singh

In this model [Yu and Singh, 2001], the result of direct interactions is stored as what the authors call quality of service (-*QoS*-). Agents only keep the most recent interactions, and each agent defines a threshold for each partner over which she is classified as a trustworthy agent.

Also, the model incorporates for each agent a *TrustNet* structure, in a similar way as Schillo *et al.* [Schillo et al., 2000] and Histos [Zacharia, 1999]. The difference is that agents being queried can refer to other agents. The initial agent will take into account the information only if the refereed agents are not too far in the social tree. The model uses Dempster Shafer evidence theory to aggregate the information from different source agents.

Mui *et al.*

Mui *et al.*'s model [Mui et al., 2001, Mui et al., 2002b] suggests a similar approach as the one proposed by Yu & Singh [Yu and Singh, 2003], where reputation is inferred from propagated ratings through a peer-to-peer network. To combine the information coming from different agents, the model uses Bayesian-like statistics. The model assumes that each interaction is an independent Bernoulli experiment. Then, it defines a random variable as the sum of the Bernoulli distributions whose expectation is exactly the average. Finally, it estimates lower and upper bounds using Chernoff bounds for the probability of success of the next trial. In contrast, Yu & Singh use Dempster Shafer evidence theory for this aggregation. The propagation mechanism for reputation is done in a very similar way than Yu & Singh.

LIAR

The LIAR model presented by Muller & Vercouter [Muller and Vercouter, 2005] focuses on the detection of fraud and reputation management in the communications. The authors use a normative language to formalize prohibited situations in terms of the information sent by the agents and the commitments that they set. Through this, the model defines a procedure capable to detect lies.

The model mainly uses two different *kinds* of reputation: Direct Experience-Based Reputation and Observation-Based Reputation. With this information agents can decide whether to *trust* or *distrust* the information sent by a given source agent. The authors detail the decision making process for the trust decision, and thus, from our perspective, it becomes a trust model. The model is framed in peer-to-peer networks.

Regan & Cohen

Regan & Cohen [Regan and Cohen, 2005] proposed a trust system for online market places where the set of buyers and sellers is well-distinguished. The authors argue that only sellers should be evaluated by the buyers, and not the opposite, because sellers have more control over exchanges and transactions. According to the model, buyers can evaluate sellers by computing what the authors called *direct* reputation (similar to Image) and *indirect* reputation. The former is calculated by dividing the pool of sellers into three groups: Those with good direct reputation (or good image), those with bad image, and those that are unknown. Then, buyers evaluate each transaction with sellers through a satisfaction threshold.

The calculus of indirect reputation is done by the introduction of informer agents (*advisors* in the authors' words). These agents own direct information about the target and can send under request such information to buyers. They use peer-to-peer networks to model the exchange and request of such information.

Sierra & Debenham

Sierra and Debenham [Sierra and Debenham, 2005] presented an information-based trust model for agents involved in negotiation processes. Their main concern is to compute the probability for an agent α to accept a proposition δ from an agent β. For this computation, the model uses three sources of information that are properly weighted:

- The reputation that according to α has β about proposition δ. So, the model accepts witness information.
- The power that β has in the social group. The model incorporates sociological information, similar to the Regret model [Sabater and Sierra, 2002].
- The *trust* that α has on β that δ will be accomplished. The authors calculate such measure only using the history of observations.

The *trust* measure is computed as the conditional entropy (from Shannon's information theory [Shannon, 1948]) of the distribution that tells the probability of β to achieve δ, knowing the previous observations (signed contracts and fulfillments). This measure is somehow related to the *direct trust* in the Regret model.

Schillo *et al.*

The model presented by Schillo *et al.* [Schillo et al., 2000] was designed for societies or environments where the evaluation of interactions between agents has a boolean nature, for instance, *good* or *bad*. For this reason it works perfectly in scenarios like the prisoners dilemma. The idea is that the result of an interaction computes the honesty of the partner by checking what she claimed and what she finally did. Taking into account all the results in the interactions, the model calculates the probability on the honesty in the next interaction, by simply dividing the number of interactions where the agent was honest by the total number of interactions. Then, let A, B be agents, where A has observed B being honest h times on a total of n interactions, the probability for A that B will be honest the next interaction is calculated by $T(A, B) = \frac{h}{n}$.

This naive idea is complemented with a very interesting source of information. They incorporate a social network, a *TrustNet* data structure, for each agent. The idea is that agents can query other agents that have met before. This witness information will be a set of interaction results, not a summary of them, that agents can incorporate to their probability calculus.

Ripperger

In the book by Ripperger [Ripperger, 1998][11] the author describes from a pure economical perspective the *creation* of trust as a mechanism to stabilize uncertain expectations when choosing actions. Thus, the author considers trust as expectations, and use economic theories to calculate it, developing detailed decision making processes for the selection.

Rasmusson & Janson

Rasmusson & Janson [Rasmusson and Janson, 1996] and [Rasmusson and Janson, 1997] propose a mechanism similar to Padovan *et al.* explained above, with the introduction of special agents as trusted third party or reviewer agents. The authors' main focus was as well online marketplaces although their results can be applied to open multi-agent systems. The model considers that agents should use gossiping in order to find out faster their desired information. The interesting part of the model is that to reduce the intentional spreading of false information, agents can pay agents to remember them, not in the case though of asking for information. The idea is to use incentives to ensure that paid agents tell the truth.

Sen and Sajja

In the model presented by Sen and Sajja [Sen and Sajja, 2002b] the authors explicitly talk about reputation. The model considers two kinds of direct experience: direct interaction and direct observation. The idea is that only direct interactions give an exact perception of the performance of the agents. The authors suppose that observations are noisy, and that may differ from reality. Due to this difference, the impact than direct interactions have on the updating rule of reputation values is much higher than direct observations. They represent the reputation values as real numbers in the interval $[0, 1]$ where 0 represents the worst reputation and 1 the best one, following a linear function.

[11] The brief description of this model was extracted from [eRep, 2007], because it only exists a German edition of the book

In addition, in their model agents can query other agents about the performance of other partners, being the answer always a boolean, good or bad. From this witness information, agents calculate the number of positive and negative answers received about the same partner.

Repage and BDI+Repage

We have already explained in detail the Repage model [Sabater-Mir et al., 2006] in this chapter. The BDI+Repage model [Pinyol and Sabater-Mir, 2009a, Pinyol et al., 2008, Pinyol, 2008, Pinyol and Sabater-Mir, 2008] is one of the contributions of this work and thus, it is explained in the following chapters.

2.6 Conclusions

In this chapter we have explored the theoretical bases of our work, the state-of-the-art regarding computational trust and reputation models and the position of our model in the current literature. The classification dimensions that we provide enhance the contribution of the model, and are also summarized in the introductory chapter. We would like to remark though several considerations:

1. Even when we are placing the BDI+Repage model as a trust model, we want to clarify that the architecture is more general and that potentially, it could serve as a global agent architecture. In the model we do not explicitly define trust, but it emerges from a set of beliefs, desires and intentions when a decision is made and such decision involves an action to interact with another agent.

2. Also, when trust emerges from the reasoning, it can be completely defined in terms of a mental state composed of beliefs, desires and intentions. Hence, we classify it as a cognitive model.

3. As far as we know, the BDI+Repage model is the only trust model that has a cognitive representation and at the same time, an analytical formulation to update and calculate the cognitive components of trust.

4. Finally, the model has a general purpose. It is not attached to any underlying network typology nor ontology, and thus, it could and should be adapted to the peculiarities of the environments, although we believe that this knowledge could be codified as beliefs.

The BDI+Repage model somehow proves that *trust* can not be considered apart and independent of the goals of the agent, and that can become an emergent property induced by a mental state of the agent, following Castelfranchi & Falcones' ideas [Castelfranchi and Falcone, 1998b].

We would like to remarkable the proliferation of cognitive models in the last few years. Besides Castelfranchi and Falcone's model, published in 1998, Repage, forTrust and BDI+Repage were published in 2006, 2008 and 2009 respectively. This shows an increasing interest in considering the representation of such complex concepts as mental states.

Chapter 3

An Ontology of Reputation: A Computational Account

3.1 Introduction

In this chapter we present the ontology of reputation mentioned in the introduction, and the language L_{rep}, to capture the reputation information. It serves to precisely determine the elements that compose a social evaluation and at the same time, provides a clear conceptualization of the involving terms. In a more pragmatic perspective, the purpose of this chapter is twofold: (1) to establish the terminology that we use to describe reputation-related concepts, and (2) to formally define a language that captures the information that reputation models compute.

The terminology on reputation concepts is described as a taxonomy, which can be seen as an ontology of reputation. We detail the elements that social evaluations manage according to the current state-of-the-art models and the cognitive theory of reputation explained in the previous chapter. We pay special attention to the representations that computational models use to quantify how *good* or *bad* targets result to be in a concrete context.

Also, we formally specify L_{rep} to capture the elements of the ontology from an individual point of view. It serves both as a communication language and to characterize the reputation information that agents manage.

3.2 The Ontology

In this section we provide a list of concepts and their relationship that play a crucial role in the conceptualization of reputation-related information. The list does not intend to be complete, but as exhaustive as possible. Moreover, the computational description offers a pragmatic approach that we use to formalize in a well-precise terminology the problems we deal with in this work. Figures

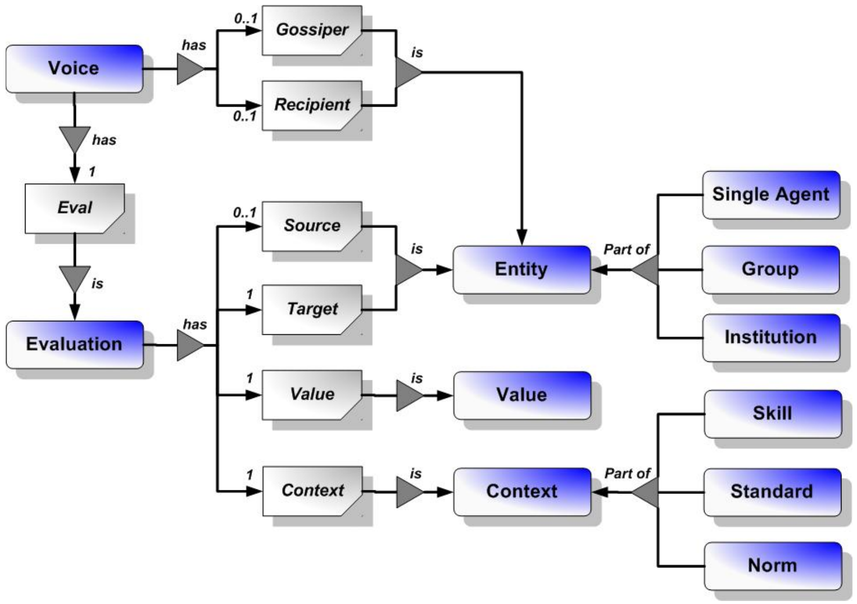

Figure 3.1: The main components of an evaluation and voice. Elements with cardinalities 0..1 indicate that are optional.

3.1, 3.2 and 3.3 show a graphical schema of such elements and relations.

3.2.1 Components of Social Evaluations

In a social evaluation (figure 3.1) we find three compulsory elements: a `target`, a `Context` and a `Value`. Intuitively, the construct describes the common elements that social evaluations include, without indicating the *type* of social evaluation, which we detail later in this section. Also, it is possible to find the `Source` of the evaluation (optional), indicating who is the creator of such evaluation.

Entity

An `Entity` is any element of the society susceptible of either being evaluated or having an active part in the generation and diffusion of social evaluations. From the point of view of the cognitive theory of Conte and Paolucci [Conte and Paolucci, 2002] an `Entity` can participate in the reputation process in four different ways:

1. `Target`: An `Entity` that is being evaluated.

2. `Source`: An `Entity` that generates the evaluation.

3. `Gossiper`: An `Entity` that spreads an `Evaluation`.

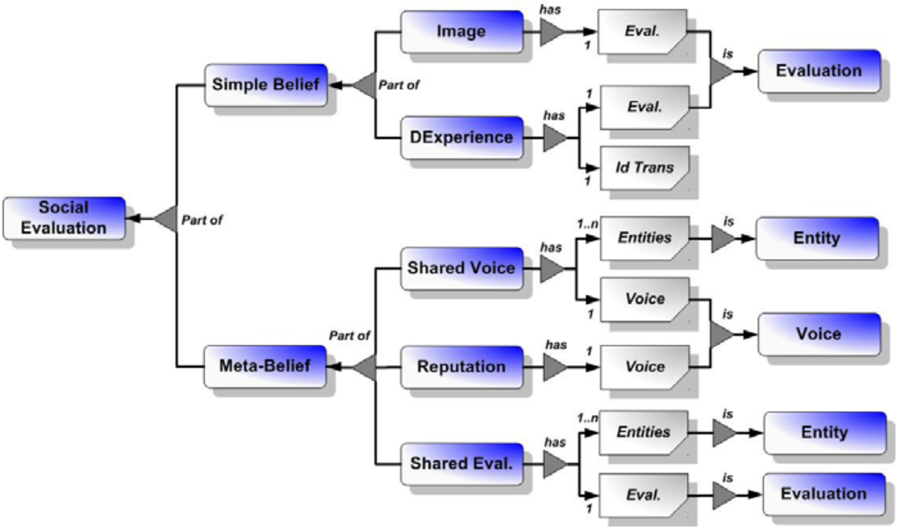

Figure 3.2: The taxonomy of social evaluations

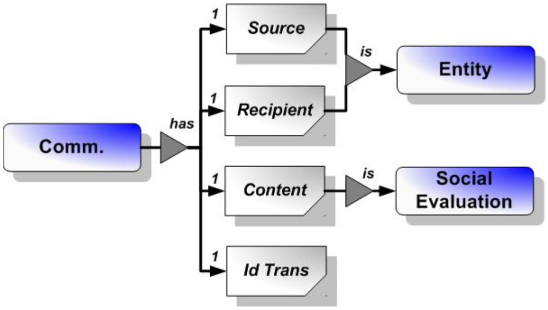

Figure 3.3: The components of a communicated social evaluation

4. Recipient: An Entity that receives an Evaluation

An Entity can be a single agent, a group of agents or an institution.

Context

Agents can evaluate the same Target from different perspectives. For instance, we can have a bad *image* of Agent A as a chess player, but a very good *image* of the same agent as a soccer player. This is what we call the context of the evaluation. In the ontology, the context can describe a Norm, a Standard or a Skill.

Value

The Value describes the quantification of the social evaluation. It describes in a well-defined semantics how *good* or *bad* the target results to be in the specified context. As an example, some models use linguistic labels for this enterprise, *very good*, *good* etc., others, a value between 0 and 1, being 0 the worst evaluation, and 1 the best one. In section 3.2.3 we propose four representation types that are representative according to the current state-of-the-art models. We also define how the conversion between different types can be performed in order to preserve as much as possible the predefined semantics of the original representation.

Evaluation and Voice

Finally, an Evaluation encapsulates all the elements that participate in a social evaluation. It includes two Entities playing the role of Source and Target, the Context and the Value of the evaluation. The source is optional and represents the entity that has generated the evaluation. The target, the context and the value are *sine qua non* elements for the existence of a social evaluation.

Similarly, the notion of Voice, includes the necessary elements to represent the spreading of an Evaluation. A Voice is defined as a “report on reputation”. For instance, “It IS SAID that John is good at playing soccer” is an example of a Voice. Besides the Evaluation itself, it has two Entities that identify the Gossiper and the Recipient of the Voice.

3.2.2 A Taxonomy of Social Evaluations: Beliefs and Meta-beliefs

As we mentioned in the previous chapter, image and reputation are social evaluations. However, some recent work [eRep, 2006b] based on [Conte and Paolucci, 2002], state that concepts like shared voice and share evaluation among others can be also classified in these terms. In this subsection we detail the specific elements that are involved in each *type* of social evaluation, developing a taxonomy.

As shown in figure 3.2, we acknowledge that social evaluations are evaluative beliefs. It means that they have a representation in terms of beliefs, and such representation involves an evaluation. Also, as described in the cognitive theory, `image` is a belief while `reputation` is a meta-belief. In the following lines we briefly define them:

Image

An agent holding it, believes in the `Evaluation` that contains the object. In other words, an `Image` is the believed opinion of the agent about a given `Target` with respect to a given `Context`. The important point here is that the agent believes that the `Evaluation` is *true*.

Reputation

`Reputation` is a generalization and loss of reference of a `Shared Voice`. An agent holding a `Reputation` believes that most of the entities would acknowledge the existence of a `Voice`. It refers to what a target agent "IS SAID to BE" by most of the population or group. From the point of view of the holder agent, it is understood as a belief of others' beliefs in the sense that the holder agent believes that most of the population believe certain evaluation. For instance, taking again our example, to acknowledge that most of the people say that "John is good at playing soccer", can be understood as to believe that most of the people believe that "John is good as a soccer player", but this does not imply to really believe that John is good at it. As explained earlier, we consider `Reputation` as a meta-belief.

Shared Evaluation

In this case, an agent holding a `Shared Evaluation` believes that each member of a perfectly identified set of `Entities` (`Group`) believes the `Evaluation` included in the object. Clearly, this concept is considered also a meta-belief.

Shared Voice

An agent holding a `Shared Voice` has the certainty that a perfectly identified set of `Entities` would acknowledge the existence of the `Voice` if asked. Following the previous example, the fact that agents A, B, C and D inform agent X that "It IS SAID that John is good at playing soccer" is understood as a `Shared Voice`, since agents A, B, C, D share the same `Voice` about *John*. The object only refers to what the agents have reported, it is not a representation of what they believe. We consider that a `Shared Voice` is a meta-belief.

Direct Experience

It is an object that refers to the `Evaluation` that an entity creates from a single interaction or experience with another `Entity`. After an interaction, the generated outcome (the objective result of the transaction) is subjectively evaluated by the agent. As shown in the figure 3.2, a direct experience contains an `Evaluation` and a transaction id `IdTrans`.

Communicated Social Evaluations

Communications are an essential part for the creation of social evaluations. Figure 3.3 shows the components of a communication. We find three main elements: the source and recipient of the communication, which are `Entities`, and the content of the communication, that is a `Social Evaluation`. The source refers to the origin of the communication, and should not be confused with the source of the evaluation, which refers to the creator of a social evaluation. The recipient refers to the entity that receives the communication at a concrete instant of time, and should not be confused with the recipient of a `Voice`. The latter refers to the general conception that a voice or a rumor can be held by an `Entity`, and may not be present (see the cardinality). Instead, the recipient of a communication is always present and refers to the dialectical view of the communication: at instant time t, an `Entity` (recipient) has received a communication from another `Entity` (the source).

Thus, agents are endowed with the capability to communicate images and reputations, but also shared voices, shared evaluations, and even direct experiences. Notice that we do not allow to communicate information about communications. This is a reasonable simplification since no existing reputation models allow for that, although it would be easy to extend the representation to permit it.

We consider that agents are capable to infer images, reputations and the reminding social evaluations from a finite set of direct experiences and communications. The way agents do these inferences depends on the reputation model they are using. For this reason, we define communications and direct experiences as ground elements.

3.2.3 Value Representations and Transformations

The representation of evaluative values is one of the most important element that characterize a reputation model. In the literature we find models that use simple boolean values indicating how *good* or *bad* agents are, others use a numerical ratio, and others, probability distributions. For instance, the eBay model uses a system of colored stars to show the reputation of a seller that could be seen as a simple integer number between 0 and 100.000, meanwhile the Repage model uses a probability distribution over the discrete set Very Bad, Bad, Neutral, Good, Very Good.

In this subsection we give four possible representation types that, although not exhaustive, are representative of the current-state-of-the-art models. Those are *Boolean* (BO), *Rational* (RE), *Discrete Set* (DS) and *Probability Distribution* (PD). We provide a descriptive semantics for each type and conversion functions

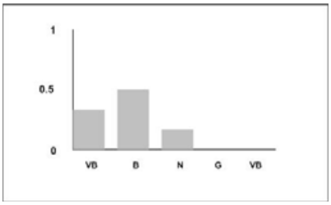

Figure 3.4: Graphical representation of the probability distribution (0.3,0.5,0.2,0,0)

that allows to move from one type to another preserving the semantics *as much as possible*.

Representation Types

- **Boolean Representation** (BO): In this case, evaluations take two possible values, good or bad. We define *true* as Good, and *false* as Bad.
- **Bounded Rational Representation** (RE): Here, the value is a number included in the bounded interval $[0, 1] \cap \mathbb{Q}$ where 0 is the worst evaluation, 1 the best evaluation and 0.5 the absolute neutral evaluation. The curve we have chosen indicating the level of goodness/badness is linear, from 0 to 1. Alternatively, other curves could be defined.
- **Discrete Sets Representation** (DS): In this case, the value belongs to the following sorted discrete set {*Very Bad, Bad, Neutral, Good, Very Good*} ({VB,B,N,G,VG} from now on). Its semantics is intrinsic on the definition of each element of the sorted set. Of course, other linguistic labels could be chosen.
- **Probabilistic Distribution Representation** (PD): Finally, this last representation applies a probability distribution (PD) over the sorted discrete set seen in the DS representation.

  Let L be the vector $[VB, B, N, G, VG]$ where $L_1 = VB, L_2 = B$ and so on. If X is a probability distribution over L then we define X_i as the probability of being evaluated as L_i. Given a probability distribution x we have that $\sum_{i=1..5} x_i = 1$. For instance, we could have the distribution $[0.3, 0.5, 0.2, 0, 0]$ meaning that with probability of 0.3 the target is *Very Bad*, with 0.5 that is *Bad* and with 0.2 that is *Neutral*. Graphically it can be represented as shown in figure 3.4.

We want to remark that the transformations presented here are not unique. They could be presented in multiple ways. We provide though an illustrative set of conversions that try to preserve as much as possible the semantics. Obviously, since some representations are more expressive than others, some transformations imply a loose of information. Appendix A provides a measure to compute such lost, based on the entropy from information theory.

Transformations From PD : This is the most expressive type, being the only one offering probabilities. Because of that, transforming to the other types will imply always some information loss. Here we propose one for each type:

$\rightarrow$ *To Boolean (BO)*

In the BO we only have two values. The idea is that a PD value will converge to *Good* if the probability distribution *tends* to values $\{G, VG\}$, and *Bad* if it *tends* to the values $\{VB, B\}$. In our context, the word *tend* implies that we need an operation capable of transforming a probability distribution element to an unidimensional number, where a threshold can tell us whether to transform the PD value to true or false. This function calculates is the center of mass (CM) of a PD element, $CM : PD \rightarrow [0,1] \cap \mathbb{Q}$. It returns a bounded rational number $\in [0,1] \cap \mathbb{Q}$ indicating in terms of average how *good* (converging to 1) or *bad* (converging to 0) is the evaluation whose value is represented in a probabilistic distribution. Of course 0.5 would be the absolute neutral. Then, it is easy to think that values over 0.5 would indicate mostly good, and below mostly bad. The value 0.5 will be our threshold. Let $x \in PD$, the function CM is defined as follows:

$$CM(x) = \frac{1}{10} \sum_{i=1}^{5} (2 \cdot i - 1) \cdot x_i \tag{3.1}$$

To transform a given PD value x to a boolean it is enough to evaluate the following boolean expression[1]: $CM(x) >= 0.5$

$\rightarrow$ *To Rational (RE)*

Let $x \in PD$ the transformation to a RE is: $CM(X)$

$\rightarrow$ *To Discrete Set (DS)*

Notice that due to the semantics of the bounded rational type (RE), the interval $[0,1]$ could be mapped into the discrete set type (DS) $\{VB, B, N, G, VG\}$ in an easy way, keeping the semantics in the transformation. The function $R : [0,1] \cap \mathbb{Q} \rightarrow \{VB, B, N, G, VB\}$ does this mapping as follows: Let $x \in [0,1] \cap \mathbb{Q}$, then

$$R(x) = \begin{cases} VB & \text{if } 0 \leq x \leq 0.2; \\ B & \text{if } 0.2 < x \leq 0.4 \text{ ;} \\ N & \text{if } 0.4 < x \leq 0.6 \text{ ;} \\ G & \text{if } 0.6 < x \leq 0.8 \text{ ;} \\ VG & \text{if } 0.8 < x \leq 1 \text{ .} \end{cases} \tag{3.2}$$

We have already seen how to transform an element from type PD to type RE. We can apply the R function over the resulting element in type RE, obtaining an element of type DS. Given that, the full transformation of an element $x \in PD$ is calculated by $R(CM(x))$.

[1] The decision of including 0.5 as a good evaluation is totally arbitrary, but consistent in all the transformations. Alternatively, one could suppose a threshold that should be consistent with the reminding transformations.

Transformations From DS :

$\rightarrow$ *To Boolean (BO)*

In this case, the semantics that has $true$ in the boolean representation suggests the mapping to G or VG in a discrete set representation, and the $false$ to VB or B. Following the same decision we made in the previous transformations, the neutral value N should be considered $true$ as well. Formally, we define the function $S : DS \rightarrow [1,5] \cap I\!N$ that returns the index position of a given element in the sorted set $\{VB, B, N, G, VG\}$.

Let $x \in DS$, then

$$S(x) = \begin{cases} 1 & \text{if } x = VB; \\ 2 & \text{if } x = B \ ; \\ 3 & \text{if } x = N \ ; \\ 4 & \text{if } x = G \ ; \\ 5 & \text{if } x = VG \ . \end{cases} \tag{3.3}$$

The transformation to BO is calculated as $S(x) \geq 3$.

$\rightarrow$ *To Rational (RE)*

For this transformation we recall here that function R (equation 3.2) divides the interval $[0,1]$ into five parts, each of them assigned to one of the values of the type DS. For instance, all the values between 0.2 and 0.4 are mapped into the element B of DS. Thus, given an element of type DS, the rational equivalent value should be included in the interval defined in function R. For example, a VB value as rational would be in the interval $(0.2, 0.4]$. Although whatever value in the interval would be fine, we pick the one just in the middle, 0.3, the average between the maximum and the minimum. To formalize the transformation we use a function that gives this central point. $C : [1,5] \cap I\!N \rightarrow \{0.1, 0.3, 0.5, 0.7, 0.9\}$. Let $y \in [1,5]$ the function C is defined as $C(y) = \frac{2 \cdot x - 1}{10}$. We can now describe the transformation: Let $x \in DS$, the transformation to RE is exactly $C(S(x))$.

$\rightarrow$ *To Probabilistic Distribution (PD)*

This case is simple, since a DS can be seen as a particular case of a PD, assigning the probability of 1 to the corresponding element of the set. We define the function $B : [1,5] \cap I\!N \rightarrow PD$, that creates a PD element assigning a probability of 1 to the corresponding element and zero to the rest. Let $x \in [1,5] \cap I\!N$ the function B is defined as:

$$B(x) = \{y \in PD : \forall_{z \neq x} y_z = 0 \wedge y_x = 1\} \tag{3.4}$$

Then, let $x \in DS$, its transformation to PD is calculated with the expression $B(S(x))$.

Transformations From RE :

$\rightarrow$ *To Boolean (BO)*

Let $x \in RE$, the transformation between a RE type to a BO type is calculated evaluating the expression $x \geq 0.5$.

$\rightarrow$ *To Discrete Set (DS)*

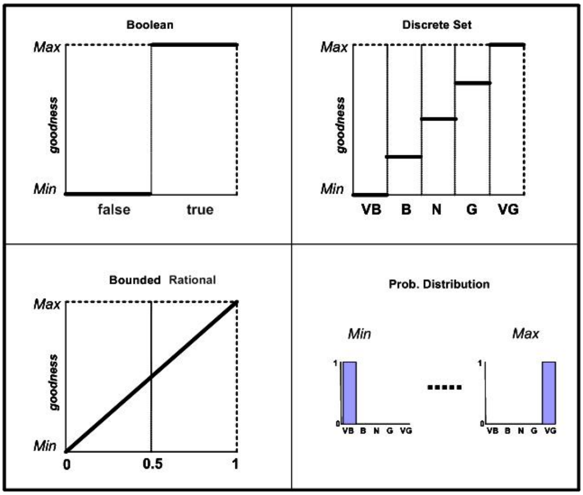

Figure 3.5: Semantic representation of the different types

Let $x \in RE$ and the function R (eq 3.2), the transformation to a DS is calculated using the expression $R(x)$.

→ *To Probabilistic Distribution (PD)*

The idea for converting an element $x \in RE$ to a PD type is to generate a PD element whose center of mass is equal to x. There are though an infinite number of possible combinations. We decide to choose the representation in which two contiguous elements of the PD set have probabilities greater than 0 and assign the corresponding probabilities in order to achieve the desirable center of mass.

Let i_1 and i_2 be the two index positions of the elements of PD that we choose to create the PD value. To calculate them we use the function $R' : [0,1] \cap \mathbb{Q} \rightarrow [1,5] \cap \mathbb{N}$ defined as

$$R'(x) = \begin{cases} 1 & \text{if } 0 \leq x \leq 0.2; \\ 2 & \text{if } 0.2 < x \leq 0.4 \ ; \\ 3 & \text{if } 0.4 < x \leq 0.6 \ ; \\ 4 & \text{if } 0.6 < x \leq 0.8 \ ; \\ 5 & \text{if } 0.8 < x \leq 1 \ . \end{cases} \tag{3.5}$$

This equation indicates to which interval certain RE element belongs to[2]. We calculate i_1 as follows:

$$i_1 = min\{R'(x), R'(max\{x - 0.1, 0\})\} \tag{3.6}$$

The index i_2 is the next following number, taking into account that the maximum allowed number is 5

$$i_2 = min\{5, i_1 + 1\} \tag{3.7}$$

[2] Notice that the equality $R'(x) = S(R(x))$ holds.

Let $x \in [0,1] \cap \mathbb{Q}$, we need to find two probabilities, z and y, such that $z+y=1$ and its center of mass (considering a PD element) is the original x. So, we need to solve the following equation:

$$x = C(i_1) \cdot z + C(i_2) \cdot y \tag{3.8}$$

Solving it, we have that $z = 1-y$ and $y = \frac{x-C(i_1)}{C(i_2)-C(i_1)}$. Now, we are aware of the probabilities that we need to assign and the two index positions of the elements of PD. We only need a function that creates a PD element from an index position and a probability. We use the function $B' : [1,4] \cap \mathbb{N} \times [0,1] \cap \mathbb{R} \rightarrow PD$ defined as

$$B'(i,p) = \{y \in PD : \forall_{r \neq i,i+1} y_r = 0 \wedge y_i = p \wedge y_{i+1} = 1-p\} \tag{3.9}$$

For instance, $B'(3,0.3)$ returns as a PD element [0,0,0.3,0.7,0], $B'(1,0.8)$ returns [0.8,0.2,0,0,0]. Finally, we have all the element to calculate the transformation from a given $x \in RE$ to a PD:

$$B'(i_1, 1 - \frac{x - C(i_1)}{C(i_2) - C(i_1)}) \tag{3.10}$$

Transformations From BO **:**

In this case we start from the most expressive type (PD) .

→ *To Probabilistic Distribution (PD)*

Again, several PD representations could represent a $true$ or $false$ value. For this enterprise we define two constants, $B_F = [2/5, 2/5, 1/5, 0, 0]$ and $B_T = [0, 0, 1/5, 2/5, 2/5]$ belonging to PD. The transformation function from a BO to PD is then quite simple. Let $x \in BO$:

$$\begin{array}{l} B_T \text{ if } x \\ B_F \text{ if } \neg x \end{array} \tag{3.11}$$

We have chosen such constants because they represent a stable agent that behaves mostly bad and mostly good respectively. When other constants are chosen, the other transformations should be adapted accordingly, as we will see in the following lines.

→ *To Discrete Set (DS)*

Here we should decide which values of the considered bad evaluations or good evaluations correspond to the $false$ and $true$ values respectively. However, once fixed the constants B_F and B_T to represent $false$ and $true$ values in probabilistic distribution, we have to take them as a base to decide the transformation. The idea is that if B_F represents a $false$, its center of mass (function CM) has to indicate the position in the interval $[0,1]$ that the $false$ value represents, and having it, the function R would determine which element of the discrete set represents the $false$ value. The same reasoning can be made for the $true$ value. Then, let $x \in BO$, the transformation to DS is:

$$\begin{array}{l} R(CM(B_T)) \text{ if } x \\ R(CM(B_F)) \text{ if } \neg x \end{array} \tag{3.12}$$

	BO	*RE*	DS	PD
$x : BO$	x	$CM(B_T)$ if x $CM(B_F)$ if $\neg x$	$R(CM(B_T))$ if x $R(CM(B_F))$ if $\neg x$	B_T if x B_F if $\neg x$
$x : RE$	$x \geq 0.5$	x	$R(x)$	$i_1 = min\{R'(x)$ $R'(x - 0.1)\}$ $i_2 = min\{5, i_1 + 1\}$ $B'(i_1, 1 - \frac{x - C(i_1)}{C(i_2) - C(i_1)})$
$x : DS$	$S(x) \geq 3$	$C(S(x))$	x	$B(S(x))$
$x : PD$	$CM(x) \geq 0.5$	$CM(x)$	$R(CM(x))$	x

Table 3.1: Conversion table between representation types. Table 3.2 summarizes the defined functions

Actually, the $false$ value goes to B, and $true$ to G.

→ *To Rational (RE)*

If we have used the function CM in the previous transformation, it is clear that to keep consistency, having $x \in BO$ the transformation must be:

$$\begin{array}{l} CM(B_T) \text{ if } x \\ CM(B_F) \text{ if } \neg x \end{array} \tag{3.13}$$

Table 3.1 summarizes the conversions, and table 3.2 the defined functions.

3.3 The L_{rep} Language

The work presented in the previous section serves as a descriptive analysis of the elements involved in social evaluations. However, we want to specify a formal language that captures the reputation-related information that individual agents use to write statements (and reason) about reputation-related information. The language is based on the ontology defined above and we use it to characterize the reputation information that single agents manage. In this sense, we assume from now on that agents talk about social evaluation in terms of the L_{rep} language.

3.3.1 Defining L_{rep}

As shown in the ontology, social evaluations incorporates three main elements: the target, the context, and the value of the evaluation [Pinyol et al., 2007b]. For instance, an evaluation may say that an agent a (target), as a car driver (context) is very good (value). The language we define in this section takes these elements into account. Following [Grant et al., 2000] where languages are built as a hierarchy of first-order languages, we define $L_{context}$, and L_{rep}. Both are classical first-order languages with equality and contain the logical symbols $\wedge, \neg$ and $\rightarrow$[3]. $L_{context}$ is the language that the agents use to describe the context

[3] For the sake of clarity we reduce the first-order languages to facts, conjunctions of facts, and rules

Domain	**Definition**
$R' : [0,1] \cap \mathbb{Q} \to [1,5] \cap \mathbb{N}$	$R'(x) = \begin{cases} 1 & \text{if } 0 \leq x \leq 0.2; \\ 2 & \text{if } 0.2 < x \leq 0.4; \\ 3 & \text{if } 0.4 < x \leq 0.6; \\ 4 & \text{if } 0.6 < x \leq 0.8; \\ 5 & \text{if } 0.8 < x \leq 1. \end{cases}$
$S : \{VB, B, N, G, VG\} \to [1,5] \cap \mathbb{N}$	$S(x) = \begin{cases} 1 & \text{if } x = VB; \\ 2 & \text{if } x = B; \\ 3 & \text{if } x = N; \\ 4 & \text{if } x = G; \\ 5 & \text{if } x = VG. \end{cases}$
$R : [0,1] \cap \mathbb{Q} \to \{VB, B, N, G, VG\}$	$R(x) = \begin{cases} VB & \text{if } 0 \leq x \leq 0.2; \\ B & \text{if } 0.2 < x \leq 0.4; \\ N & \text{if } 0.4 < x \leq 0.6; \\ G & \text{if } 0.6 < x \leq 0.8; \\ VG & \text{if } 0.8 < x \leq 1. \end{cases}$
$CM : PD \to [0,1] \cap \mathbb{Q}$	$CM(x) = \frac{1}{10} \sum_{i=1}^{5} (2i-1)x_i$
$B' : [1,4] \cap \mathbb{N} \times [0,1] \cap \mathbb{R} \to PD$	$B'(i,p) = \{y \in PD : \forall_{r \neq i,i+1} y_r = 0 \wedge$ $y_i = p \wedge y_{i+1} = 1 - p\}$
$C : [1,5] \cap \mathbb{N} \to \{0.1, 0.3, 0.5, 0.7, 0.9\}$	$C(x) = \frac{2x-1}{10}$
- -	$B_F = [2/5, 2/5, 1/5, 0, 0]$ $B_T = [0, 0, 1/5, 2/5, 2/5]$

Table 3.2: Summary of functions for the transformation types

of the evaluations, like norms, or skills, while L_{rep} is used to write statements about social evaluations.

Definition ($L_{context}$ - *Domain Language*) $L_{context}$ is an unsorted first-order language that includes predicates, constants and functions, necessary for writing statements about the domain. Even when we do not provide any specific language for describing the context of the evaluations, we suggest that a first-order language should be enough to express norms, standards or skills.

Definition (L_{rep} - *Reputation Language*) L_{rep} is a sorted first-order language used to reason about social evaluations. It includes $L_{context}$ and special first-order predicates that are identified by their sorts. These special predicates describe the types of social evaluations, (*Image, Reputation, Shared Voice, Shared Evaluation, Direct Experience*) and Communications (Img, Rep, ShV, ShE, DE and Comm from now on). We call direct experiences and communications *ground elements*, and are the basic elements from which social evaluations are inferred.

The sorts that the language uses are the following:

- S_A: It includes a finite set of target identifiers $\{i_1, \ldots, i_n\}$, which embraces single agents, group of agents and institutions. In fact, we assume that each possible group has assigned an identifier.

- S_F: It contains the set of constant formulas representing elements of $L_{context}$ and L_{rep} itself. The idea is that well-formed formulas from $L_{context}$ and L_{rep} are introduced in L_{rep} as constants for the language[4]. In this way, they can be nested in a first-order predicate. Regarding embedded L_{rep} formulas we only allow one nested level. We use it to capture the idea of communicated social evaluations.

- S_V: It represents the values of the evaluation. In the previous sections we have described four representation types that could be mapped in this sort (BO, RE, DS, PD). Later in this section we provide the characteristics that, according to our needs, such representation values should have. We require that the set of possible values is countable, and that a linear order is defined between the values.

- S_T: It incorporates discrete time instants. We use them to express that direct experiences and communications take place in a *discrete* unit of time. In a more pragmatic view, it also serves as a unique identifier for the communication and direct interactions.

We pay special attention to the sort S_V, which represents values of a totally ordered set $M = \langle G, \leq \rangle$. It includes the set of constants C_V containing a label v for each $v \in G$. Examples of M are $\langle [0,1] \cap \mathbb{Q}, \leq \rangle$ (RE), where $\leq$ is the standard pre-order binary function for rational numbers, or $\langle \{VB, B, N, G, VG\}, \leq_s \rangle$ (DS)

[4] It can be built recursively and simultaneously with S_F. We add the constant $\lceil \varphi \rceil$ for each $\varphi \in wff(L_{context})$ and the constant $\lceil \Psi \rceil$ for each formula $\Psi \in wff(L_{rep})$

referring to the linguistic labels *Very Bad, bad, Neutral, Good, Very Good*, and where $VB \leq_s B \leq_s N \leq_s G \leq_s VG$. With the probabilistic distribution representation (PD), a possible pre-order could be defined by considering the center of mass.

The set of well-formed formulas of L_{rep} (wff(L_{rep})) is defined using the standard syntax of classical first-order logic, and it has special first-order predicates. Those are Img, Rep, ShV, ShE, DE, $Comm$. As mentioned before, the last two predicates (DE and $Comm$) are what we call *ground elements*.

- $Img(S_A, S_F, S_V)$: Represents an image predicate, an evaluation that is believed by an agent. For instance,

  $$Img(j, \lceil Provider(service(X)) \rceil, VG)$$

  indicates that the agent holding the predicate has a VG image of agent j as a provider of service X. In terms of the mental state of the agent, it indicates that the holder of the predicate *believes* such evaluation. In this case and in future examples, we take M as $< VB, B, N, G, VG, \leq_s >$ where the elements represent linguistic labels indicating *very bad*, *bad*, *neutral*, *good* and *very good*.

- $Rep(S_A, S_F, S_V)$: Represents a reputation predicate. A reputation refers to an evaluation that is known to circulate in the society. For example

  $$Rep(j, \lceil Provider(service(X)) \rceil, VG)$$

  indicates that "it is said" that agent j is VG as a provider of service X. In this case, the agent holding the predicate believes that the evaluation circulates in society, but this does not imply that the agent believes the evaluation.

- $ShV(S_A, S_F, S_V, S_A)$, $ShI(S_A, S_F, S_V, S_A)$: Represents a shared voice and a shared image respectively. A shared voice is also an evaluation that circulates in society, like reputation. The difference is that in this case, the members of society that *say* it, are identified ($\mathcal{G}$). A shared image is a belief about the beliefs of other agents. It indicates that the holder of the predicate believes that a certain group of identified agents ($\mathcal{G}$) *believe* an evaluation. Both predicates include the group that shares the voice or image respectively.

- $DE(S_A, S_F, S_V, S_T)$: Represents a direct experience. For instance,

  $$DE(j, \lceil Provider(service(X)) \rceil, VG, t_2)$$

  indicates that the agent had a VG direct experience with j as a Provider of service X at the time t_2.

- $Comm(S_A, S_F, S_T)$: Represents a communication. For example,

  $$Comm(j, \lceil Img(j, k, Provider(service(X)), VG) \rceil, t_2)$$

  indicates that the agent received a communication at time t_2 from agent j saying that its image about k as a Provider of service X is VG.

Often we will write a subindex to explicitly state the agent holding the predicate. For instance, $DE_i(j, \ulcorner Provider(service(X)) \urcorner, VG, t_2)$ indicates that agent i has had a direct experience with j as a provider of service X at the time t_2 and it was VG.

3.3.2 Reputation Theories

To characterize all the reputation information that an agent i holds we define the concept of reputation theory. Intuitively, we consider that from a set of direct experiences (DE) and communications (Comm) (what we call ground elements) agents are able to infer the remaining information (image, reputation, shared voice and shared evaluation) through a consequence relation $\vdash_i$, associated with agent i. The consequence relation represents agent i's reputation model. Formally:

Definition (*Reputation Theory*) Let $\Delta \subset wff(L_{rep})$, we say that Δ is a reputation theory when $\forall \alpha \in \Delta$, α is a ground element (a direct experience or communication). Then, letting $d \in wff(L_{rep})$, we write $\Delta \vdash d$ to indicate that from the reputation theory Δ, it can be deduced d via $\vdash$.

The reputation-related information that agent i holds is characterized then by the tuple $\langle \Delta_i, \vdash_i \rangle$, where Δ_i is the set of ground elements gathered by i through interactions and communications (i's reputation theory), and $\vdash_i$ the consequence relation (i's reputation model).

In the next section we show how L_{rep}, together with the representation types we have defined can capture the reputation-related information provided by three well-known reputation models.

3.4 L_{rep} on work: Examples

3.4.1 eBay Reputation Model

eBay site [eBay, 2002] is one of the most concurred (if not the most) online marketplace in the world with more than 50 million registered users. eBay reputation model considers reputation as a public and centralized value in which the context is implicit. In this case, users rate sellers after each transaction, with values of +1, 0 , -1. The reputation value of the sellers then is calculated as the sum of all the ratings over the last six months, and presented to potential buyers with a system of colored stars.

From this definition, we can conclude that in this model, the reputation theory is composed of a set of communicated direct experiences, where the ratings from the buyers are the direct experiences. We can consider that the context is the constant C, and the value representation is the bounded rational type ($[0, 1] \cap \mathbb{Q}$). We can easily normalize the values $-1, 0, 1$ to $0, 0.5, 1$ respectively. As a matter of example, let $b_1, b_2, \ldots$ be users, and $s_1, s_2, \ldots$ sellers, a reputation theory for the eBay system could have the following elements:

$$\begin{array}{l} Comm(b_1, DE(s_1, C, 0, t_1)) \\ Comm(b_2, DE(s_1, C, 0, t_2)) \\ Comm(b_2, DE(s_1, C, 0.5, t_3)) \end{array}$$

$$\begin{array}{l} Comm(b_1, DE(s_2, C, 1, t_4)) \\ Comm(b_4, DE(s_2, C, 1, t_5)) \\ Comm(b_3, DE(s_2, C, 1, t_6)) \end{array}$$

Then, the model is able to compute the general reputation of each one of the sellers. Since eBay punctuation goes from 0 to 100000, a simple normalized transformation to the interval [0,1] seems to be enough. However, notice that the colored stars representation does not follow a linear curve. From a semantic point of view and in our value representation, 0 means very bad reputation, 0.5 neutral reputation, and 1 very good reputation, with a totally linear function. In eBay, having more that 10 point is already considered a good reputation. The next step in the scale is more than 100 points (with a different colored star), and the next is more than 500. In conclusion there is no lineal relation between the punctuation and the semantic representation of the stars. Then, it is necessary a transformation from the ontology representation value to the eBay scale. A possible transformation function is described in the following equation:

$$H : [0, 100000] \rightarrow [0, 1] \tag{3.14}$$

$$H(X) = \begin{cases} 0 & \text{if } X < 10; \\ 1 & \text{if } X > 100000; \\ \frac{log(X)-0.5}{8} + 0.5 & otherwise. \end{cases} \tag{3.15}$$

The idea is that from a set of communicated direct experiences reputation predicates can be inferred. According to the previous reputation theory example, the generated predicates would be

$$\begin{array}{l} Rep(s_1, C, 0) \\ Rep(s_2, C, 0) \end{array}$$

In the example, s_2 gets a punctuation of 0 because its punctuation is still lower than 10.

3.4.2 Abdul-Rahman and Hailes Model

The distributed model presented by Abdul-Rahman and Hailes [Abdul-Rahman and Hailes, 2000] uses the term *trust* even though as shown in the previous chapter, it cannot be considered a trust model. In this case, social evaluations take into account the context. The model is fed by two sources: direct experiences and third-party communications of direct experiences. The representation of the evaluations is done in terms of the discrete set {*vt (very trustworthiness), t (trustworthiness), u (untrustworthiness), vu (very untrustworthiness)*}. Then, for each agent and context the system keeps a tuple with the number of past own experiences or communicated experiences in each category. For instance, agent A may have a tuple of agent B as a seller like $(0, 0, 2, 3)$, meaning that agent A has received or experienced 2 results as untrustworthiness and 3 as very untrustworthiness. Finally the *trust* value is computed taking the maximum of the tuple values. In our example for agent A,

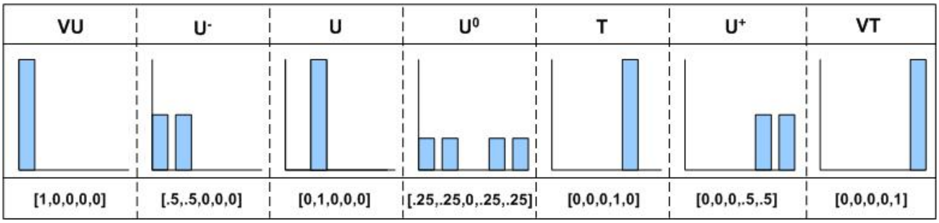

Figure 3.6: Abdul-Rahman and Hailes model values expressed in terms of a probabilistic distribution(PD)

agent B as a seller would be very untrustworthy. In case of tie between *vt* and *t* and between *u* and *vu* the system gives the values U^+ (mostly trustworthy) and U^- (mostly untrustworthy) respectively. In any other tie case the system returns U^0 (neutral).

Each agent holds its own reputation theory. In this case, we could define a specific representation type for the evaluation values of the model. Notice that they are linguistic labels. However, to illustrate the use of the probabilistic distribution type, we chose it and establish the relation shown in figure 3.6. An example of a reputation theory for the agent *i* could be:

$$
\begin{array}{l}
DE_i(b_1, seller, [1,0,0,0,0], t_1) \\
DE_i(b_1, seller, [0,1,0,0,0], t_2) \\
DE_i(b_2, seller, [0,0,0,0,1], t_3) \\
DE_i(b_2, seller, [0,0,0,0,1], t_4) \\
Comm_i(u_1, DE_{u_1}(b_1, seller, [1,0,0,0,0], t_x), t_5) \\
Comm_i(u_2, DE_{u_1}(b_1, seller, [0,1,0,0,0], t_y), t_6)
\end{array}
$$

From this theory, agent *i* is able to infer image predicates. The *trust* measure that the model provides, in terms of the ontology, coincide with the concept of image, because agents accept the measure as *true*. Then, using the transformation shown in figure 3.6, the following image predicates can be inferred:

$$
\begin{array}{l}
Img_i(b_1, seller, [0.5, 0.5, 0, 0, 0]) \\
Img_i(b_2, seller, [0, 0, 0, 0, 1])
\end{array}
$$

3.4.3 The Repage Model

In chapter 2 we have already explained the Repage model in detail. It can be observed that the internal elements coincide quite well with the predicates defined in L_{rep}. First, the ground elements in the model are communicated images, communicated reputations, communicated third-party images, and outcomes (direct experiences in terms of L_{rep}). For instance, a set of communicated images and communicated reputations gathered by agent *i* could be:

$$
\begin{array}{l}
Comm_i(u_1, Img_{u_1}(s_1, seller, [0.2, 0.3, 0.5, 0, 0]), t_1) \\
Comm_i(u_2, Img_{u_2}(s_2, seller, [0, 0, 0, 0.3, 0.7]), t_2) \\
Comm_i(u_1, Rep_{u_1}(s_1, seller, [0.5, 0.3, 0.1, 0.1, 0]), t_3) \\
Comm_i(u_2, Rep_{u_2}(s_2, seller, [0.5, 0, 0, 0, 0.5]), t_4)
\end{array}
$$

Note that in the example, the source of the communication and the source of the communicated predicate is the same agent in all the communications. This is the difference with third-party communicated images. For instance:

$$\begin{array}{l} comm_i(u_1, Img_{u_5}(s_1, seller, [0.2, 0.3, 0.5, 0, 0]), t_1) \\ comm_i(u_2, Img_{u_6}(s_2, seller, [0, 0, 0, 0.3, 0.7]), t_2) \end{array}$$

In the first communication, agent u_1 communicates to agent i the image that u_5 holds about s_1 as a *seller*. The model does not consider third-party communicated reputations even when the language could capture it. Regarding direct experiences, we consider that outcome predicates coincide with the definition of a direct experience that we have given. Recalling that a direct experience is the subjective evaluation of a direct interaction, outcome predicates from Repage represents it by evaluating the difference between a given contract and the fulfillment. In any case, a reputation theory that captures the Repage grounding information is represented by a set of communications as explained above, and direct experience predicates that capture the outcomes that Repage generates.

From such information, Repage is able to infer predicates like shared evaluation, shared voice, image and reputation. Notice that Repage considers other intermediate predicates, like candidate images, candidate reputations etc. If necessary, the language could be easily extended with them. We keep it as it is now, and assume that from a reputation theory, the agent is able to infer, for instance the following predicates:

$$\begin{array}{l} Img_i(s_1, seller, [0.3, 0.3, 0.3, 0.1, 0]) \\ Img_i(s_2, seller, [0, 0, 0.1, 0.2, 0.7]) \\ Rep_i(s_1, informer, [0, 0, 0, 0, 1]) \\ Rep_i(s_2, informer, [0, 0.5, 0.5, 0, 0]) \\ ShE_i(s_1, seller, [0.3, 0.3, 0.3, 0.1, 0], \{u_1, u_2\}) \\ ShV_i(s_1, seller, [0.3, 0.3, 0.3, 0.1, 0], \{u_1, u_2\}) \\ \ldots \end{array}$$

In [Pinyol and Sabater-Mir, 2009b] we redefine Repage in terms of a finite set of deductive rules that implements and characterizes the consequence relation $\vdash_i$, associated to the language L_{rep}.

3.5 Related Work

The ontology presented in this chapter is based on the set of terms about reputation concepts defined in the European project eRep [eRep, 2006b, eRep, 2006a]. The aim of this effort was to define an ontology that all partners participating in the project would use as a consensual starting point. This ontology describes in detail all the elements participating in social evaluations, as well as the processes of transmitting them. We took a subset of these elements and provide a more computational view.

Nevertheless, this is not the only work referring to the definition of a reputation ontology. Casare et al. [Casare and Sichman, 2005] propose a functional ontology whose goal is to put together at a conceptual level all the knowledge about reputation. It is based on the concepts defined in the Functional Ontology of Law [Valente, 1995]. The approach is interesting from a theoretical point of view because it offers a structured definition of reputation and its related concepts, including processes of transmission, but it does not detail the internal elements. The main difference between the presented ontology and the ontologies in eRep [eRep, 2006b] and Casare et al. [Casare and Sichman, 2005] is the computational focus which deals with representation types.

3.6 Conclusions

In this chapter we have described a taxonomy of social evaluations, including the elements that in general reputation models manage. Summarizing, *Image, Reputation, Shared Evaluation, Shared Voice* and *Direct Experiences* are social evaluations. The common elements that social evaluations have are a *target*, a *context* and a *value*. In the ontology, we group these elements in the object `Evaluation`. Intuitively, the context describes the property being evaluated of the target, which can be a single agent, a group of agents or even an institution. The value quantifies the evaluation. Also, we have introduced four representative representation types for evaluation values. In the appendix A we explain a direct application of such ontology and the conversion among representation types that we have presented in the chapter. The application deals with the interoperability of agents using different reputation models.

Also, we introduce direct experiences and communications as ground elements. We suggests that from a set of direct experiences and third-party communications agents infer social evaluations. Hence, agents can communicate images, reputations or even their own direct experiences. As seen in chapter 2, most of the current state-of-the-art reputation models use communication of social evaluations as a source for computing evaluations.

The chapter also serves to characterize, from an individualistic point of view, the reputation information that agents manage. For this, based on the previously defined ontology, we introduce the L_{rep} language, a first-order language that captures the elements of the ontology. We associate to this language a consequence relation $\vdash_i$ for each agent i. We state that agents can use the same language to express social evaluations but use different rules to infer them.

We show how the language can successfully capture the reputation information managed by three current reputation models. In concrete the Repage model [Sabater-Mir et al., 2006], whose internal elements already coincide with the elements in the ontology, the eBay system [eBay, 2002] and the model by Abdul-Rahman and Hailes [Abdul-Rahman and Hailes, 2000].

We provide in the following chapters a BDI agent architecture that integrates image and reputation information. Since the most expressive representation type that we use is the probabilistic distribution used by the Repage model, we take it as the paradigmatic example for the integration, although potentially we could use it with models that use other representations, as shown in this chapter.

Chapter 4

Image, Reputation and Beliefs: A Logical Account

4.1 Introduction

In this chapter we define L_{BC}, the belief language and logic that the BDI+Repage model uses for the grounding of image and reputation information, and for reasoning about acquired knowledge. Obviously, the focus of this work is to represent and reason about the information computed from the Repage system, although the logic is generic enough and allows extensions. The main characteristics of L_{BC} are:

- **Reasoning with probabilities**: Since Repage provides social evaluations in terms of probability distributions, the logic must accept beliefs on probabilities. This is substantially different than probability beliefs. The later refers to beliefs whose evaluations are not crisp (true or false), but rely on probability measures (see [Casali et al., 2004, Casali et al., 2008] for instance). We refer to crisp beliefs that contain probability information.

- **Different representation for Image and Reputation**: Image and reputation are distinct objects, and thus their representation in terms of beliefs can differ. In the language, information from Image is represented with the special predicate E, while reputation with the predicate S. Then, the axiomatic we present relate both predicates to the belief predicate B which determines what the agent believes in a given instant of time.

- **Completeness**: The completeness of the logic was a must for us. For this, we do not move away from first-order logic and define L_{BC} as a distinguishable subset of many-sorted first-order logic. We show then the existence of consistent theories. This approach has advantages and limitations that we comment in the chapter.

4.2 Defining the Belief Logic

In this section we describe the language L_{BC} to express agents' beliefs and to reason about them. The language must be able to capture the semantics that Repage predicates bring over formulas. Since a social evaluation in Repage describes a behavior of a target agent in a role as a probability distribution, L_{BC} must capture probabilities over some underlying language of the agents' ontology.

Agents also need to perform basic epistemic inferences. In general, agents observe and interact with the environment, incorporating knowledge to their respective bases. Obviously, we focus on the knowledge that comes from Repage system, which provides evaluations in terms of probabilities and that can be combined with other knowledge of the agent through logical inferences. This allows the agents to combine such knowledge with their desires to finally generate intentions and act in consequence to fulfill them. The idea is that L_{BC} must capture all the knowledge that agents believe at a given instant of time.

To define L_{BC} we use the approach described in [Grant et al., 2000] where languages are structured as a hierarchy. A different approach that also uses hierarchies of languages is the one taken by [Giunchiglia and Serafini, 1994], that could be alternatively used for our purposes. Both works suggest that first-order logic is enough to define consistent theories of propositional attitudes for rational agents. In these papers, formulas φ from a certain propositional language A can be *embedded* into another language B as constants for the language, usually written as $\lceil\varphi\rceil$. For instance, we can have a language that describes possible weather events in cities: $Rain(Barcelona)$, $Sunny(Rome) \wedge Sunny(Berlin)$, and another language can talk about these events in terms of date/time: $Forecast(10/11/2010, \lceil Rain(Barcelona)\rceil)$.

4.2.1 Preliminaries: An Intuitive Idea

We want to illustrate with an example the kind of reasoning we are expecting from the logic of belief. First, we recall that the Repage system provides probability distributions over the different roles that an agent plays. For example, in a scenario with buyers and sellers, a buyer can decide to evaluate sellers in two *roles*: the quality of the products they sell and the delivery time of the products.

Role	**Possible Outcomes**				
Seller(Q)	$VeryGood\_Q$	$Good\_Q$	$Neutral\_Q$	Bad_Q	$VeryBad\_Q$
Seller(dTime)	$dTime \leq 5$	$5 < dTime \leq 10$	$10 < dTime$		

Note that the possible outcomes for each role cover all the possibilities. How such information is finally codified as beliefs is one of the contributions of this work and it is explained in detail later. For the example, it is enough to realize that part of the information that the agent manages comes from an evaluation process that the Repage provides, while other comes from the general knowledge of the agent. This is what justifies such integration.

In our model, the desires of our agent i lead the practical reasoning process. The main idea is that for each desire, the belief logic should determine which *actions* allow the agent to achieve the desire and with which probability. For instance, agent i can desire the following with a strength of 0.9

$$(D^+(VeryGood\_Q \vee Good\_Q) \wedge dTime \leq 5 \wedge payLess(500), 0.9)$$

indicating that i desires to obtain a very good or good quality product delivered in less that 5 days and paying less than 500. Then, the logic of beliefs should provide which actions are capable of producing it. In concrete we would like the system to provide beliefs like

$$B(buy(Bob), (VeryGood\_Q \vee Good\_Q) \wedge dTime \leq 5 \wedge paidLess(500), 0.45) \quad (1)$$
$$B(buy(Alice), (VeryGood\_Q \vee Good\_Q) \wedge dTime \leq 5 \wedge paidLess(500), 0.8) \quad (2)$$
$$B(buy(Charlie), (VeryGood\_Q \vee Good\_Q) \wedge dTime \leq 5 \wedge paidLess(500), 0.4) \quad (3)$$

For instance, (1) indicates that after executing the action $buy(Bob)$, agent i will obtain

$$(VeryGood\_Q \vee Good\_Q) \wedge dTime \leq 5 \wedge paidLess(500)$$

with a probability of 0.45. The belief logic should deduce such information from more simple beliefs. For example, to deduce (1) agent i can hold the following predicates:

$$B(buy(Bob), (VeryGood\_Q \vee Good\_Q), 0.9) \quad (4)$$
$$B(buy(Bob), dTime \leq 5, 0.5) \quad (5)$$
$$B(buy(Bob), paidLess(500), 1) \quad (6)$$

Our approach suggests that formulas like (4) and (5) are *generated* from Repage. Note that (4) comes from the evaluation that agent i has about *Bob* in the role *Seller(Q)*, while (5) from the evaluation of the same agent *Bob* in the role *Seller(dTime)*. The key idea is that Repage gives a probability distribution for each agent and role, and such probabilities can be combined under the assumption that distributions are stochastically independent. Then, the system should be able to infer from (4) and (5) the following:

$$B(buy(Bob), (VeryGood\_Q \vee Good\_Q) \wedge dTime \leq 5, 0.45) \quad (7)$$

where the probability of $0.45 = 0.9 \cdot 0.5$ is calculated following the standard probability computation for independent events. Also, the system should know that if a formula is always true (probability 1), like the case of (6) and it does not belong to any particular distribution, it can be combined using conjunction, to finally generate (1). Also, the formula (6) should be calculated from the knowledge that i has about how much it cost to buy at *Bob*. In this sense, it is feasible and reasonable to assume that i should deduce (6) from:

$$
\begin{array}{ll}
B(\iota, buy(Bob), paid(350), 1) & (8)\\
B(\iota, paid(350) \rightarrow paidLess(500), 1) & (9)
\end{array}
$$

where ι stands for an empty action. The beliefs are able then to codify knowledge that always holds after an action is executed (like formula (8)) and knowledge that always holds independently from the action (like formula (9)). If we want to keep an uniform notation, both kinds of formulas can be codified with probability 1.

The previous example illustrates the kind of reasoning we are looking for and the properties of the belief logic, which we enumerate in the following lines:

(i) Evaluations from Repage codify the knowledge about the probabilities. This includes not only the assignment of probabilities for each agent and role, but the correct construction of the probability spaces. For instance, regarding Bob and role $Seller(dTime)$ the following beliefs could be generated:

$$
\begin{array}{l}
B(buy(Bob), dTime \leq 5, 0.5)\\
B(buy(Bob), 5 < dTime \leq 10, 0.3)\\
B(buy(Bob), 10 < dTime, 0.2)\\
\hline
B(buy(Bob), dTime \leq 10, 0.8)\\
B(buy(Bob), dTime \leq 5 \vee 10 < dTime, 0.7)\\
B(buy(Bob), 5 < dTime, 0.5)\\
\hline
B(buy(Bob), dTime \leq 10 \vee 10 < dTime, 1)
\end{array}
$$

(ii) Repage provides evaluations for each agent and role in terms of image and reputation, which define two probabilistic distributions over the same agent and role that must be combined to finally generate beliefs. To avoid inconsistencies, we introduce besides the belief predicate B two more predicates, E (image) and S (reputation). Through the appropriate axioms we combine them to finally generate beliefs that do not fall into inconsistencies.

(iii) When combining two formulas, in order to preserve a correct semantics and accuracy of the probabilities, we only can ensure that the resulting probability is correct when such formulas refer to the same action (so, the same agent) and talk about different roles, which we assume are stochastically independent. For this we need to codify into the belief predicates also the roles that are involved in the formula, and permit the combination of beliefs only when the intersection of such set of roles is empty. For instance, following the above example, the beliefs should be codified in the following way:

$$\begin{array}{l}
B(buy(Bob), dTime \leq 5, 0.5, \{Seller(dTime)\}) \\
B(buy(Bob), 5 < dTime \leq 10, 0.3, \{Seller(dTime)\}) \\
B(buy(Bob), 10 < dTime, 0.2, \{Seller(dTime)\}) \\
\hline
B(buy(Bob), dTime \leq 10, 0.8, \{Seller(dTime)\}) \\
B(buy(Bob), dTime \leq 5 \vee 10 < dTime, 0.7, \{Seller(dTime)\}) \\
B(buy(Bob), 5 < dTime, 0.5, \{Seller(dTime)\}) \\
\hline
B(buy(Bob), dTime \leq 10 \vee 10 < dTime, 1, \{Seller(dTime)\})
\end{array}$$

Then the belief

$$B(buy(Bob), dTime \leq 5, 0.5, \{Seller(dTime)\})$$

can be combined with

$$B(buy(Bob), (VeryGood\_Q \vee Good\_Q), 0.9, \{Seller(Q)\})$$

because the intersection of the respective set of roles is empty and the action is the same. The resulting conjunction could be:

$$B(buy(Bob), (VeryGood\_Q \vee Good\_Q) \wedge dTime \leq 5, 0.45, \{Seller(Q), Seller(dTime)\})$$

Note that we could also combine them with a disjunction:

$$B(buy(Bob), (VeryGood\_Q \vee Good\_Q) \vee dTime \leq 5, 0.95, \{Seller(Q), Seller(dTime)\})$$

where $0.95 = 0.9 + 0.5 - 0.45$ is calculated following standard probabilistic computations. In both cases the set of roles is the same, since it is an indication of the roles that effect the formula. This mechanism prevents the logic to combine formulas which are not independent, so that the intersection of their respective set of roles is not empty.

The following subsection formalizes the syntax and semantics of the belief logic L_{BC}.

4.2.2 L_{BC} Syntax and Semantics

Following [Grant et al., 2000] we define two languages. The first one, denoted by L_{basic}, is the object language. L_{basic} is a classical propositional language that contains the symbols needed by the agents for writing statements about the application domain. The second language, denoted by L_{BC}, is the language the agents use to reason about beliefs, image and reputation. L_{BC} is a first-order many-sorted language that contains constant symbols for the formulas of the language L_{basic}.

For instance, in the example stated above, L_{basic} could be composed of the set of elementary propositions that we use to describe the possible outcomes of each role: $VeryGood\_Q$, $Good\_Q$, …, $dTime \leq 5$, $5 < dTime \leq 10$, … and the propositions $Paid(X)$ and $PaidLess(X)$ for each rational number X. Then, the language is constructed with the standard syntax of propositional logic that includes the symbols $\neg$, $\wedge$, $\vee$ and $\rightarrow$ necessary to express the base domain, as shown in the example.

L_{BC} is a first-order many-sorted language and contains four sorts:

- S_A: the sort representing actions.
- S_F: the sort representing formulas of the language L_{basic}.
- S_R: the sort representing the power set of roles.
- S_P: the sort representing probability values.

We use different letters for variables of different sorts of L_{BC}:

- $a, a_1, a_2, \dots$ for variables of sort S_A
- $x, x_1, x_2, \dots$ for variables of sort S_F
- $r, r_1, r_2, \dots$ for variables of sort S_R
- $p, p_1, p_2, \dots$ for variables of sort S_P

Constants and predicate symbols of L_{BC} are identified by their sorts. The sort S_A includes a finite set of constant symbols C_A to denote actions. It also contains the constant ι to denote the special empty action. The sort S_F includes a set of constant symbols C_F to denote all formulas of the language L_{basic}. The set C_F contains constants of the form $\lceil\sigma\rceil$, where σ is a formula of L_{basic}. The sort S_P includes a set of constant symbols C_P to denote rational numbers in the unit interval $[0,1] \cap \mathbb{Q}$. For each $p \in [0,1] \cap \mathbb{Q}$ we introduce the constant $\overline{p}$ in the sort. However, in general, for the sake of clarity, we omit the overline notation for rational constants. Finally, the sort S_R includes a finite set of constant symbols C_R to denote finite sets of roles.

Before we proceed with the introduction of the L_{BC} syntax, it is important to remark two questions with respect to our notation. On the one hand, note that the symbols $x, x_1, x_2, \dots$ are for variables of sort S_F in general, while the symbols $\lceil\varphi\rceil$ are constants of sort S_F that denote only formulas of the language L_{basic}. On the other hand, given a finite set of roles $\delta = \{R_1, \dots, R_1\}$, we have in C_R a constant, say c, denoting this set of roles. When we introduce the axiomatization of the logic, we use the notation $\mathcal{E}(c)$ to refer to the set δ denoted by constant c. For the sake of clarity we use sometimes the set of roles instead of the constant in some axioms. For instance, if the constant c denotes the set of roles

$$\{seller(quality), seller(dTime)\}$$

we will write the latter instead of c.

Now we specify the predicate symbols corresponding to various sorts. In the notation introduced below, the predicate symbol B, for instance, is written $B(S_A, S_F, S_P, S_R)$. This means that B is a predicate symbol of arity 4, with first argument in S_A, second argument in S_F, third argument in S_R and fourth argument in S_P. The language L_{BC} contains the following predicate symbols:

- Belief Predicate: $B(S_A, S_F, S_P, S_R)$.
- Image Predicate: $E(S_A, S_F, S_P, S_R)$.
- Reputation Predicate: $S(S_A, S_F, S_P, S_R)$.

L_{BC} contains various function symbols, that allow us to deal with parts of the agents' formulas and to express the reasoning of the agents. The functions applied to the sort S_F are one unary function $neg : S_F \to S_F$ for the negation of formulas, and the binary functions $con : S_F \times S_F \to S_F$ for conjunctions and $imp : S_F \times S_F \to S_F$ for implications.

For instance, if $\ulcorner\varphi\urcorner, \ulcorner\phi\urcorner \in C_F$ then $imp(\ulcorner\varphi\urcorner, \ulcorner\phi\urcorner)$ is interpreted as $\ulcorner\varphi \to \phi\urcorner$, $con(\ulcorner\varphi\urcorner, \ulcorner\phi\urcorner)$ as $\ulcorner\varphi \wedge \phi\urcorner$, and $neg(\ulcorner\phi\urcorner)$ as $\ulcorner\neg\phi\urcorner$. The expression $or(x, y)$ stands for $\neg(con(\neg(x), \neg(y)))$. At first sight, all these functions can be regarded as purely syntactic transformations, but they are important in our construction because they allow us to write sentences that talk about parts of the formulas of L_{basic}.

The semantics of L_{BC} is the usual for a first-order many-sorted language. In this section we have presented only a few definitions and notation. A detailed introduction to the syntax and semantics of first-order many-sorted logics can be found in [Enderton, 1972].

4.2.3 The Basic Axioms

In this section we define a theory Γ over L_{BC}, i.e. the axioms that agents use to reason. The theory contains the minimal formulas to describe the behavior of the predicates introduced above. We assume that the function product, sum and subtraction are defined for rational constants: $* : S_P \times S_P \to S_P$ where $* \in \{\cdot, +, -\}$. Remark that we are not giving an axiomatization of the logic, but only a set of axioms for a theory (a set of sentences in this first-order language closed under the logical consequence relation). For that reason we do not need to introduce inference rules. We assume that the deductive system is given (for instance, as defined in [Enderton, 1972]).

RA: Conjunction

$$\forall a x_1 x_2 p_1 p_2 r_1 r_2 (B(a, x_1, p_1, r_1) \wedge B(a, x_2, p_2, r_2) \to B(a, con(x_1, x_2), p_1 \cdot p_2, r_3))$$

when $\mathcal{E}(r_1) \cap \mathcal{E}(r_2) = \emptyset$ and r_3 denotes the union of the two sets of roles $\mathcal{E}(r_1) \cap \mathcal{E}(r_2)$. Intuitively speaking, the axiom indicates that when two formulas talk about independent distributions (so, disjoint set of roles) we can ensure that the joint probability is the product.

MP: Modus ponens

$$\forall a x_1 x_2 r (B(a, x_1, 1, r) \wedge B_i(a, imp(x_1, x_2), 1, r) \to B_i(a, x_2, 1, r))$$

Note that this axiom (formulated as in [Grant et al., 2000]) indicates that agents use modus ponens when reasoning with formulas of the object language L_{basic}. It is an axiom of a theory, not of the logic. We assume an standard axiomatization of many-sorted first-order logic (cf. [Enderton, 1972]) in which the modus ponens rule holds for every formula of the language.

NE: Necessity axiom for actions

$$\forall axr(B(\iota, x, 1, r) \rightarrow B(a, x, 1, r))$$

The axiom ensures that when the agent believes that a formula is true with a probability 1, after whichever action is performed, the formula will be also true.

CO: Completeness of probability

$$\forall ax(B(a, x, p, r) \rightarrow B(a, neg(x), 1 - p, r))$$

This ensures that when an agent knows the probability of a formula, also knows its complementary. The axiom is interesting because it states that a formula and its complementary cover all the probabilistic space.

Moreover, note that given $e \in C_A$ a constant denoting an action, $\lceil\varphi_1\rceil, \lceil\varphi_2\rceil \in C_F$, $d_1, d_2 \in C_P$ denoting rational numbers and $c_1, c_2 \in C_R$ denoting sets of roles, when $B(e, \lceil\varphi_1\rceil, d_1, c_1)$ and $B(e, \lceil\varphi_2\rceil, d_2, c_2)$ hold, $\mathcal{E}(c_1) \cap \mathcal{E}(c_2) = \emptyset$, and $c_3 \in C_R$ is a constant denoting the union of sets of roles $\mathcal{E}(c_1) \cup \mathcal{E}(c_2)$, the previous axiomatization accomplishes the additive property of probabilistic spaces[1]:

$$\begin{array}{l} B(e, con(\lceil\varphi_1\rceil, \lceil\varphi_2\rceil), d_1, c_3) \wedge \\ B(e, con(\lceil\varphi_1\rceil, neg(\lceil\varphi_2\rceil), d_2, c_3)) \rightarrow B(e, \lceil\varphi_1\rceil, d_1 + d_2, c_1) \end{array}$$

Also, under the same condition the disjunction of independent formulas ensures the standard calculus of probabilities:

$$\begin{array}{l} B(e, \lceil\varphi_1\rceil, d_1, c_1) \wedge \\ B(e, \lceil\varphi_2\rceil, d_2, c_2) \wedge \\ B(e, con(\lceil\varphi_1\rceil, \lceil\varphi_2\rceil), d_3, c_3) \rightarrow B(e, or(\lceil\varphi_1\rceil, \lceil\varphi_2\rceil), d_1 + d_2 - d_3, c_3) \end{array}$$

GBEL: Ground Beliefs

$$B(e_1, \lceil\varphi_1\rceil, 1, c_\emptyset)$$

$$\vdots$$

$$B(e_n, \lceil\varphi_n\rceil, 1, c_\emptyset)$$

Those are the beliefs that describe the general knowledge of the agent. Each a_k is an action (possibly also the empty action ι), and each $\lceil\varphi_k\rceil$ denotes a proposition, a conjunction of propositions or a rule of the form $(\varphi_1 \wedge \varphi_m) \rightarrow \varphi$ from L_{basic}. The probabilistic distributions are given by the predicates E and

[1] The standard formulation in probability theory is $Pr(A \cap B) + Pr(A \cap \overline{B}) = Pr(A)$.

S interpreted by the Repage system, which ensures that the distributions are correct. To avoid inconsistencies we require that all the propositions are positive. $c_\emptyset$ denotes the empty set of roles.

GI: Ground Images

Let $\varphi_1 \dots \varphi_n$ be formulas of L_{basic} that completely define the space of a distribution corresponding to role R. Let also $e \in C_A$ be a constant denoting an action and $d_1, \dots, d_4 \in C_P$. Then the following formulas are in the theory:

$$\begin{gathered}
E(e, \lceil \varphi_1 \rceil, d_1, \{R\}) \\
E(e, \lceil \varphi_2 \rceil, d_2, \{R\}) \\
\vdots \\
E(e, \lceil \varphi_n \rceil, d_n, \{R\}) \\
E(e, \lceil \varphi_1 \vee \varphi_2 \rceil, d_1 + d_2, \{R\}) \\
E(e, \lceil \varphi_1 \vee \varphi_3 \rceil, d_1 + d_3, \{R\}) \\
\vdots \\
E(e, \lceil \varphi_2 \vee \varphi_3 \rceil, d_2 + d_3, \{R\}) \\
E(e, \lceil \varphi_2 \vee \varphi_4 \rceil, d_1 + d_4, \{R\}) \\
\vdots \\
E(e, \lceil \varphi_1 \vee \varphi_2 \vee \varphi_3 \rceil, d_1 + d_2 + d_3, \{R\}) \\
E(e, \lceil \varphi_1 \vee \varphi_2 \vee \varphi_4 \rceil, d_1 + d_2 + d_4, \{R\}) \\
\vdots \\
E(e, \lceil \varphi_1 \vee \varphi_2 \vee \dots \varphi_n \rceil, 1, \{R\})
\end{gathered}$$

They describe the full probabilistic space with the constraint that the disjunction of all the propositions belonging to the distribution corresponding to role R covers the complete space. For the kind of reasoning we want to perform, this is enough.

GR: Ground Reputations

Let $\varphi_1 \dots \varphi_n$ be formulas of L_{basic} that completely define the space of a distribution corresponding to role R. Let also $e \in C_A$ be a constant denoting an action and $d_1, \dots, d_4 \in C_P$. Then the following formulas are in the theory:

$$
\begin{gathered}
S(e,\lceil\varphi_1\rceil,d_1,\{R\})\\
S(e,\lceil\varphi_2\rceil,d_2,\{R\})\\
\vdots\\
S(e,\lceil\varphi_n\rceil,d_n,\{R\})\\
S(e,\lceil\varphi_1\vee\varphi_2\rceil,d_1+d_2,\{R\})\\
S(e,\lceil\varphi_1\vee\varphi_3\rceil,d_1+d_3,\{R\})\\
\vdots\\
S(e,\lceil\varphi_2\vee\varphi_3\rceil,d_2+d_3,\{R\})\\
S(e,\lceil\varphi_2\vee\varphi_4\rceil,d_1+d_4,\{R\})\\
\vdots\\
S(e,\lceil\varphi_1\vee\varphi_2\vee\varphi_3\rceil,d_1+d_2+d_3,\{R\})\\
S(e,\lceil\varphi_1\vee\varphi_2\vee\varphi_4\rceil,d_1+d_2+d_4,\{R\})\\
\vdots\\
S(e,\lceil\varphi_1\vee\varphi_2\vee\ldots\varphi_n\rceil,1,\{R\})
\end{gathered}
$$

IRB: Image-Reputation-Belief

Finally, the following axiom scheme combines E and S predicates over the same action, formula and distribution to generate beliefs. Depending of how we define the axioms, we can model different kinds of agents. The most general case is:

$$\forall axp_1p_2r(E(a,x,p_1,r)\wedge S(a,x,p_2,r))\rightarrow B(a,x,h(p_1,p_2),r)$$

where $h : [0,1]\cap\mathbb{Q}\times[0,1]\cap\mathbb{Q}\rightarrow[0,1]\cap\mathbb{Q}$ is a function that combines the probabilities and preserves the probability distribution properties. An example of such a function could be the average, or weighted average function in order to give more importance to image or reputation information. Next section discusses it in more detail.

Equality Predicate

For all formulas φ,ϕ of L_{basic} the theory contains the following:

$$neg(\lceil\varphi\rceil)=\lceil\neg\varphi\rceil$$

$$imp(\lceil\varphi\rceil,\lceil\phi\rceil)=\lceil\varphi\rightarrow\phi\rceil$$

$$con(\lceil\varphi\rceil,\lceil\phi\rceil)=\lceil\varphi\wedge\phi\rceil$$

We must include them to ensure the completeness with respect to our intended semantics.

4.2.4 The Basic Semantics

In this subsection we show that the set of axioms presented above defines a first-order theory (say Γ) that is consistent. We do it by showing that the theory has, at least, a model that contains a set of positive atoms that exist in the model. Such model represents the reasoning process that the agent follows to deduce belief predicates. Following a similar approach than [Grant et al., 2000], we consider only models that contain ground terms of the language, so Herbrand models.

Proposition The theory Γ has a minimal model $\mathcal{M}$ for any underlying language L_{basic}.

Proof To prove it, we construct $\mathcal{M}$ by induction following a *stratification* construction of the model. The main idea is to add the minimal number of atoms that accomplish the axioms, starting from the atoms that must be present in all the models, i.e. GBEL (ground beliefs), GI (ground images), GR (ground reputation) and the equality predicates for terms and rational numbers, and continuing by induction. Like in the construction of models used in logical programming, the strata k ($k \geq 1$) of the model includes all the generated atoms that require the application of at least k axioms to be created. Thus, the ground atoms generated from the axioms GBEL, GI, GR and equality predicates are in the first strata, and belong to the model $\mathcal{M}$ (note that they are all positive), becoming the starting point of the construction. In the induction step we assume that $\mathcal{M}$ already contains the atoms until the strata k. The generation of the strata $k+1$ is done by applying any relevant axiom to the atoms already in $\mathcal{M}$.

The *application* of axioms in the induction step implies to add the minimal number of atoms that satisfy each axiom. We do not show the details on how each axiom creates and add new atoms. However, we illustrate it with the axiom RA (conjunction). Let us assume that the following ground atoms are already in the model.

$$B(buy(john), VeryGood\_Q, 0.9, \{seller(quality)\})$$

$$B(buy(john), dTime \leq 5, 0.5, \{seller(dtime)\})$$

Then, to preserve the consistency of the model, the axiom RA is applied, and then, the following atom must be included into the model:

$$B(buy(john), VeryGood\_Q \wedge dTime, 0.45, \{seller(quality), seller(dTime))\}$$

Under the assumption that GI and GR are well-constructed, so, they define correct probabilistic distributions, and that all GBEL axioms contain positive propositions, the construction of the model can be done for any underlying L_{basic} without falling into inconsistencies. $\Box$

Given ground beliefs (GBEL), ground images (GI) and ground reputations (GR), the construction of the model $\mathcal{M}$ gives us the belief formulas that the agent holds. Note that one and only one model exists, because all the axioms are universally quantified and do not contain disjunctions.

Also, note that under the assumption that GI and GR define correct probabilistic distributions, the axiomatization models the behavior of probability spaces for each role, and the combination of them when they are independent (different roles are involved). This is what the axiom IRB (image-reputation-belief) ensures.

4.2.5 Related Work

Some current state-of-the-art logics inspired us for defining the logic. The probabilistic and dynamic notions have been mostly treated in epistemic logic ([Kooi, 2003], [Fagin and Halpern, 1994]), and in a simpler way in belief logic [Casali et al., 2004]. Propositional probabilistic variants of dynamic logic have been studied with the goal of analyzing probabilistic programs (for instance [Kozen, 1983]).

Furthermore, some formalizations of trust using belief logic have been done [Liau, 2003], where trust is related to information acquisition in multi-agent systems, but in a crisp way. Similar to this, in [Demolombe and Lorini, 2008], modal logic is used to formalize trust in information sources, also with crisp predicates. Here, actions and communicated formulas are also used.

Regarding fuzzy reasoning on trust issues, in [Flaminio et al., 2008] a trust management system is defined in a many-valued logic framework where beliefs are graded. Also, in [Demolombe and Liau, 2001] it is proposed a logic that integrates reasoning about graded trust (on information sources) and belief fusion in multi-agent systems. Our logic does not use graded beliefs. Instead, we use the notion of beliefs on probability sentences, since Repage social evaluations describe probabilities on the outcomes of future direct experiences.

Finally, in [Pinyol et al., 2008] a probabilistic dynamic belief logic is defined for dealing also with image and reputation notions. In this logic, beliefs and actions are considered normal modalities while probability predicates are considered non-standard modalities. In [Pinyol et al., 2008] only the expressiveness of the logic is explored.

Notice that we could have extended any of the previous logics (or other formalisms such as Gabbay's labelled deductive systems) to fulfill our original necessities. However, we wanted a very flexible logical framework with a very clear orientation towards possible implementations. Even when first-order logic is semi-decidable and it is not possible to guarantee very low complexities, it is indisputable that restricting the logic to Horn clauses together with other minimal assumptions, would ensure an easy adaptation to logic programming platforms.

4.3 Grounding Image and Reputation to L_{BC}

In this section we show how L_{BC} is capable of capturing image and reputation predicates from Repage, and how such information is transformed into the beliefs of the agent.

4.3.1 Image and Reputation Predicates

As stated, image and reputation predicates computed from Repage are captured by the following expressions

- $Img(j, r, [V_{w_1}, \ldots, V_{w_m}])$

- $Rep(j, r, [V_{w_1}, \ldots, V_{w_m}])$

corresponding to the Image and Reputation of agent j playing the role r, from the point of view of the evaluator. We mention that the original implementation of Repage considers a tuple of 5 elements to represent the value of the evaluations. However, we generalize it, considering m ($m \geq 2$) elements. When in Repage the role and its labeled weights are defined, the role uniquely identifies an interaction model with two participants, and each w_k identifies a predicate, a formula from L_{basic}. To simplify, we can assume that the interaction model identified by a role is summarized in a single action[2]. Thus, we presuppose the definition of a mapping $\mathcal{R}_{r,j}$ between a given role r and agent j to an action. In a similar way, we assume a mapping $\mathcal{T}_{r,w_k}$ between each role r and label w_k to a formula written in L_{basic}.

We illustrate this with an example: In a typical market, the transaction of buying a certain product involves two agents, one playing the role of buyer (the evaluator) and the other playing the role of seller (j). From the point of view of the buyer, if she wants to evaluate other agents that play the role of seller, she knows that the associated action is *buy* at agent j. So, $\mathcal{R}_{seller,j}$ maps to $buy(j)$. In the same way, the agent must know the meaning of each label w_k of Repage. Then, we can define that $\mathcal{T}_{seller,w_1}$ is $veryBadProduct$, $\mathcal{T}_{seller,w_2}$ is $okProduct$, etc.

In this mapping, the Repage predicate $Img(j, seller, [0.2, 0.3, \ldots])$ indicates that the buyer believes that there is a probability of 0.2 that after executing the action $\mathcal{R}_{seller,j}$ (corresponding to the action $buy(j)$), she will obtain a $\mathcal{T}_{seller,w_1}(veryBadProduct)$; with 0.3 that she will obtain $\mathcal{T}_{seller,w_2}(OKproduct)$, etc. With reputation predicates the structure is similar, but the concept is different. In this case it indicates that the buyer believes that the corresponding evaluation is said by the agents in the group.

Following these indications, the representation of both predicates in L_{BC} is quite simple. Let j be an agent identifiers and r a role, then

$$\frac{Img(j, r, [V_{w_1}, V_{w_2}, \ldots])}{\begin{array}{c} E(\mathcal{R}_{rj}, \mathcal{T}_{r,w_1}, V_{w_1}, \{r\}) \\ E(\mathcal{R}_{rj}, \mathcal{T}_{r,w_2}, V_{w_2}, \{r\}) \\ \cdots \end{array}} \qquad \frac{Rep(j, r, [V_{w_1}, V_{w_2}, \ldots])}{\begin{array}{c} S(\mathcal{R}_{rj}, \mathcal{T}_{r,w_1}, V_{w_1}, \{r\}) \\ S(\mathcal{R}_{rj}, \mathcal{T}_{r,w_2}, V_{w_2}, \{r\}) \\ \cdots \end{array}}$$

Repage ensures a correct probabilistic information in terms of a probabilistic distribution, and from these assignments it is easy to calculate the remaining disjunction probabilities necessary for the logical theory.

As a matter of example and following the scenario above, let j_1, j_2 be agents, if Repage has generated the following predicates:

$$Img(j_1, seller, [.1, .1, .1, .2, .5])$$
$$Rep(j_2, seller, [.6, .1, .1, .1, .1])$$

[2] An interaction model can be seen as a set of actions to be performed by the agents.

The logical theory should include regarding j_1

$$\begin{array}{l} E(buy(j_1), VBadProduct, 0.1, \{seller\}) \\ E(buy(j_1), BadProduct, 0.1, , \{seller\}) \\ E(buy(j_1), OKProduct, 0.1, , \{seller\}) \\ E(buy(j_1), GoodProduct, 0.2, , \{seller\}) \\ E(buy(j_1), BadProduct, 0.5, , \{seller\}) \end{array}$$

And regarding j_2:

$$\begin{array}{l} S(buy(j_2), VBadProduct, 0.6, \{seller\}) \\ S(buy(j_2), BadProduct, 0.1, \{seller\}) \\ S(buy(j_2), OKProduct, 0.1, \{seller\}) \\ S(buy(j_2), GoodProduct, 0.1, \{seller\}) \\ S(buy(j_2), BadProduct, 0.1, \{seller\}) \end{array}$$

4.3.2 Relationship between Image and Reputation

One of the key points of Repage and the cognitive theory of reputation that underlies it [Conte and Paolucci, 2002] is the relationship between image and reputation. The theory states that both are social evaluations but distinct objects. With the representation we give for image and reputation in the L_{BC} and the axiomatization (the theory Γ), the difference depends on the relationship between the predicate E and the predicate S.

Regarding the key question: *How does reputation influence image?*, Conte and Paolucci in [Conte and Paolucci, 2002] state that the relation is established basically at the pragmatic-strategic level of the agent. At this level, agents must decide which source of information to use. Typically, reputation information is used only if image information is not present, but from this perspective, reputation cannot influence the inner beliefs of the agent. However, from our logical perspective, this relationship seems closer and is defined by the axiom IRB (Image-Reputation-Belief):

$$\forall a x p_1 p_2 r (E(a, x, p_1, r) \wedge S(a, x, p_2, r)) \rightarrow B(a, x, h(p_1, p_2), r)$$

Different functions $h : [0,1] \cap \mathbb{Q} \times [0,1] \cap \mathbb{Q} \rightarrow [0,1] \cap \mathbb{Q}$ model different behaviors. We only require that h preserves the probability distribution properties. Some elaborated aggregation functions can be found in [Sabater-Mir and Paolucci, 2007], but basically, they are based on weighted averages. Thus, a family of functions is determined by the expression:

$$h(p_E, p_S) = \frac{\delta_E \cdot p_E + \delta_S \cdot p_S}{\delta_E + \delta_S}$$

where $\delta_E, \delta_S \in \mathbb{Q}_{\geq}$. Table 4.1 summarizes the behavior of a family of agents depending on the values of δ_E and δ_S. Note that h can be defined globally, as it is in the axiomatization, but we can have different functions for different distributions (roles). For instance, following the example above,

Class	**Condition**	**Description**
$\mathcal{H}_1$	$\delta_E \neq 0$, $\delta_S = 0$	Only image - The agent does not trust in reputation information.
$\mathcal{H}_2$	$\delta_E \neq 0$, $\delta_S = 0$	Only reputation - The agent does not trust in image information
$\mathcal{H}_3$	$\delta_E = \delta_S \neq 0$	The agent considers that both sources of information have the same importance.
$\mathcal{H}_4$	$\delta_E > \delta_S$	Image is more important than reputation
$\mathcal{H}_5$	$\delta_E < \delta_S$	Reputation is more important than image

Table 4.1: Different h function classes when it is based on a weighted average: $h(p_E, p_S) = \frac{\delta_E \cdot p_E + \delta_S \cdot p_S}{\delta_E + \delta_S}$

$$\begin{array}{ll} \forall axp_1p_2 & (E(a,x,p_1,\{Seller(Q)\})\wedge \\ & S(a,x,p_2,\{Seller(Q)\})) \rightarrow B(a,x,h_q(p_1,p_2),\{Seller(Q)\}) \end{array}$$

$$\begin{array}{ll} \forall axp_1p_2 & (E(a,x,p_2,\{Seller(dTime)\})\wedge \\ & S(a,x,p_2,\{Seller(dTime)\})) \rightarrow B(a,x,h_t(p_1,p_2),\{Seller(dTime)\}) \end{array}$$

where $h_q \in \mathcal{H}_2$ and $h_t \in \mathcal{H}_4$ (see table 4.1 for a description of $\mathcal{H}_2$ and $\mathcal{H}_4$). This indicates that the evaluator does not trust its own experiences regarding the quality of the product and relies on reputation. Instead, regarding the delivery time the agent gives more importance to its own direct experiences. This configuration may look strange, but let us consider for instance an agent that is aware of its limitations regarding certain skills, or a robot agent that is aware that its sensors do not work well. In general, to establish this function on design time is quite difficult, because it requires precise knowledge of the society. Ideally, one can design metareasoning processes to establish the *best* function when the system is running in a real scenario. In fact, simple q-learning techniques suffice to some extend for this purpose (See the appendix B).

4.4 Conclusions

This chapter defines the language used in our BDI+Repage model to express and reason about the domain knowledge of the agent, and in particular, the social evaluations coming from Repage. We already exposed the related work that we originally explored for this enterprise and the reasons why we decided to define a new logic.

The logic allows the necessary probabilistic reasoning to capture the information computed from Repage. It this sense, the Repage model computes social evaluations (image and reputation) as probability distributions over the possible outcomes that each role may achieve. Under the assumption that the roles are independent, their associated distributions are stochastically independent as well. Current state-of-the-art probabilistic logics do not take advantage of such knowledge and use Lukasiewicz-like axiomatizations to model probabilistic inference, which does not consider the knowledge of independent distributions.

Then, our L_{BC} logic permits such inferences, and captures the *ground* meaning of image and reputation from Repage to beliefs. In our settings, such relationship is established by the axiom IRB of the L_{BC} theory. Different IRB axioms model different kinds of agents. In particular, we have defined five families of agents, all of them based on the combination of image and reputation through weighted averages. In this chapter we do not intend to provide a complete categorization of such families of agents, but show a representative set of simple combinations that substantially alterate the reasoning process of the agent. A prove of that can be found in the appendix, where we show how the selection of the IRB axiom at run-time should be performed to allow certain level of adaptation. In the next chapter, we put in context the logic within a BDI agent architecture.

Chapter 5

Reasoning Using Social Evaluations

5.1 Introduction

In the previous sections we have defined the language L_{BC} and a theory written in that language that expresses the reasoning process of the agent. We have also shown how the theory captures the semantics of image and reputation predicates coming from Repage, and how such information is combined to finally generate beliefs.

In this section, we propose a possible integration of Repage in a BDI agent[1]. The underlying idea is to define a BDI agent, specified as a multi-context system, that uses the logic presented in Section 4.2 to describe the belief base of the agent. Then, such information would be combined with the desires of the agent and other functional components to generate intentions, which in turn would end up generating proper actions. In the first part of the section, we briefly introduce the notion of multi-context system and some of the related work regarding existent multi-context BDI specifications. The second part relies on the explanation of each element that compounds our BDI+Repage architecture.

5.1.1 Multi-context Systems

Multi-context systems (MCS) provide a framework to allow several distinct theoretical components to be specified together, with a mechanism to relate these components [Giunchiglia and Serafini, 1994]. These systems are composed of a set of contexts (or units), and a set of bridge rules. Each context can be seen as a logic and a set of formulas written in that logic. Bridge rules are the mechanisms to infer information from one context to another.

[1] A preliminary version of the model described in this section was originally published at [Pinyol and Sabater-Mir, 2009a].

Giunchiglia and Serafini [Giunchiglia and Serafini, 1994] proposed the following formalization of MCS: Let I be the set of context names, a MCS is formalized as $\langle \{C_i\}_{i \in I}, \triangle_{br} \rangle$:

- $C_i = \langle L_i, A_i, \triangle_i \rangle$, where L_i is a formal language with its syntax and semantics, A_i is a set of axioms and $\triangle_i$ the set of inference rules. Thus, L_i and A_i define an axiomatic formal system, a logic for the context C_i. Beside axioms, it is possible to include a theory T_i as predefined knowledge. All A_i, $\triangle_i$ and T_i are written in the language L_i.

- $\triangle_{br}$ is a set of bridge rules.

Bridge rules can be seen as inference rules among contexts. Each one has a set of antecedents (or preconditions) and a consequent (or postcondition). Then, when each formula in the antecedent is true in its respective context, the consequent becomes true as well (also in its context). A bridge rule is represented as follows:

$$\frac{C_{i_1} : \varphi_1, \ldots, C_{i_n} : \varphi_n}{C_{i_x} : \varphi_x}$$

where $C_{i_k} : \varphi_k$ indicates that formula φ_k belongs to the context C_{i_k}, formulas $\varphi_1 \ldots \varphi_n$ are the antecedents and φ_x is the consequent. Each φ_i is a formula that belongs to its respective context, and written in its own language. So, when the formulas $\varphi_1, \ldots \varphi_n$ hold in their contexts, the formula φ_x is generated in the context C_{i_x}. However, we extend this approach by allowing in preconditions, comparisons between rational numbers. For this, the antecedent may include a set $Q_1, \ldots, Q_n$ (where $n \geq 0$) of extra conditions that must be evaluated as true to make the bridge rule applicable. Each Q_i has the form $r_1 \leq r_2$ where $r_1, r_2 \in \mathbb{Q}$ and $\leq$ corresponds to the standard boolean comparison on rational numbers.

5.1.2 MCS and BDI Agents

The use of MCS offers several advantages when specifying and modeling agent architectures [Sabater-Mir et al., 2002]. From a software engineering perspective, MCS supports modular architectures and encapsulation. From a logical modeling perspective, it allows the construction of agents with different and well-defined logics, keeping all formulas of the same logic in their corresponding context. This increases considerably the representation power of logical agents, and at the same time, simplifies their conceptualization.

Also, the use of MCS to specify BDI is not new. The BDI architecture defined in [Parsons et al., 1998] uses one context for each attitude; there is the belief context (B), the desire context (D) and the intention context (I). Each of them is equipped with a logic that corresponds to the premises that Rao and Georgeff [Rao and Georgeff, 1991] stated. Bridge rules among contexts determine the relationship between the attitudes and the type of agent: strong realism, realism and weak realism [Rao and Georgeff, 1991]. A communication context (C) is also included.

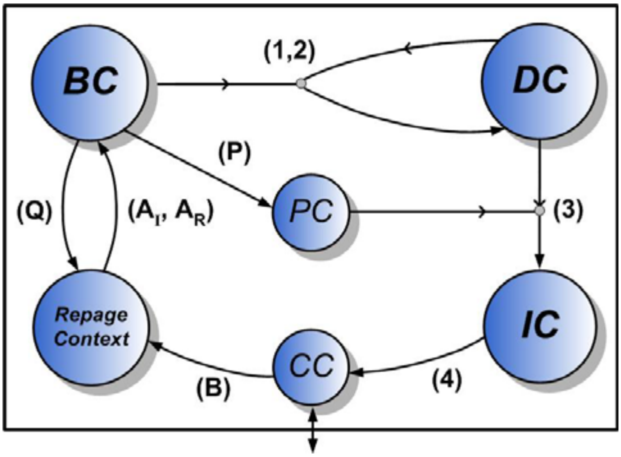

Figure 5.1: The Repage context embedded in a multi-context BDI agent. Circles and arrows represent contexts and bridge rules respectively.

In [Gaertner et al., 2006], this specification is extended by means of a new commitment context, equipped with a deontic logic, creating then a new attitude of obligation. In [Casali et al., 2004] a multi-context BDI agent is specified and its attitudes are graded. Therefore, beliefs, desires and intentions are multi-valued with grades from 0 to 1. For our BDI model, we take the logic defined for desires and intentions described in [Casali et al., 2004] and [Casali et al., 2008].

5.2 The Multi-context BDI Model

The specification of our BDI agent as a multi-context system is formalized with the tuple $Ag = \langle\{BC, DC, IC, PC, CC, RC\}, \triangle_{br}\rangle$. These correspond to Belief, Desire, Intention, Planner, Communication and Repage contexts respectively. The set of bridge rules $\triangle_{br}$ incorporates the rules $1, 2, 3, 4, P, Q$ and B (shown in Figure 5.3) and the bridge rules A_I and A_R (shown in Figure 5.2). Figure 5.1 shows a graphical representation of this multi-context specification. In the next sections we briefly explain each context and bridge rule.

5.2.1 Belief Context (BC)

This context contains the beliefs of the agent. Hence, we use the logic introduced in Section 4.2, to integrate the knowledge coming from the reputation model Repage and other knowledge gathered by the agent. Since L_{BC} is a many-sorted first-order logic, the inference rules in this context are those from first-order logic. Thus, BC-context becomes an inference system that incorporate the theory defined in section 4.2.

5.2.2 Desire Context (DC)

This context deals with the desires of the agent. Like the BDI model described by Rao and Georgeff in [Rao and Georgeff, 1991], they are attitudes that are explicitly represented and that reflect the general objectives of the agent. We consider that desires are graded, and for that, we use the multi-valued logic (DC-logic) based on the Lukasiewicz logic and described in [Casali, 2008]. The motivation for this decision arises when considering that reputation information has already a graded nature, in our case, represented as probabilities. Like in decision theory where agents manage expected utilities, we consider that from one side we obtain the probabilities, and from the other the strength of the desires. Combining them, we implement the idea of expected utility.

DC-language is built as an extension of a propositional language (in our case we use L_{basic}, the object language), by adding two fuzzy modal operators: D^+ and D^-. The intended meaning of $D^+\varphi$ is that the formula φ is desired by the agent holding it, and its truth degree, from 0 (minimum) to 1 (maximum), represents the level of satisfaction if φ holds. The intended meaning of $D^-\varphi$ is that φ is negatively desired, and the truth degree represents the level of disgust if φ holds. Also, DC-logic includes truth constants $\overline{r}$ where $r \in [0,1] \bigcap \mathbb{Q}$, and the connectives & and $\Rightarrow$ corresponding to the Lukasiewicz conjunction and implication respectively. In our architecture, agents' preferences are expressed by a set of desire expressions (both positive and negative) defining a theory.

We differentiate generic from concrete desires. Generic desires define the general preferences of the agent, and are formulas like $D^*\phi$, where $*$ stands from $+$ or $-$ and ϕ does not contain any action. Concrete desires are formulas like $D^*_\alpha\phi$ and define the desire to satisfy ϕ by executing action α. The original DC-logic from [Casali, 2008] does not consider subindex for the actions. However it uses this notation for the intentions (see next subsection). With this we indicate that a concrete desire takes into account the action to achieve the content. In this case, the grade represents the *expected* satisfaction level (or disgust if it is a negative desire) if the action is executed, implementing an equivalent *expected utility* from decision theory. Also, it serves to indicate that in the framework, actions do not behave as in dynamic logic. In our model, concrete desires are generated from generic desires and beliefs through bridge rules 1 and 2 (see section 5.2.7).

Because in Lukasiewicz logic the formula $\phi \Rightarrow \varphi$ is 1-true iff the truth value of φ is greater or equal to that of ϕ, and the truth value of $\overline{r}$ is exactly r, formulas like $\overline{r} \Rightarrow D^+\varphi$ in the theory of an agent i indicate that the level of *satisfaction* of agent i is at least r if φ holds. The same with negative desires and the level of *disgust*. From now on we will write these formulas as $(D^+\varphi, r)$ and $(D^-\varphi, r)$. The semantics is given in terms of a positive and negative preference distributions over the possible worlds. The axiomatization includes the classical logic axiom of propositional logic for non-modal formulas, plus the axioms of Lukasiewicz [Hájek et al., 1995]. It is important to remark that the author defines the semantic condition that a world that is negatively desired to some extend cannot be positively desired. In terms of the axiomatization, this implies that the same formula cannot be both negatively and positively desired.

Note though that the inclusion of $D^{-}\varphi$ and $D^{+}\neg\varphi$ is completely valid. $D^{-}\varphi$ points out to the worlds that the agent does not want to reach, but this does not mean that he will try actively to avoid it. Instead, when we include $D^{+}\neg\varphi$ in the theory the agent will try to reach worlds where $\neg\varphi$ holds. We refer to [Casali, 2008] for technical details and proof of completeness of the logic.

5.2.3 Intention Context (IC)

This context describes the intentions of the agent. Like in the Rao and Georgeff's BDI model [Rao and Georgeff, 1991], intentions are explicitly represented, but in our case generated from beliefs and desires. Also, we consider that intentions are graded, and for this we use the IC-logic defined in [Casali et al., 2004].

Similar to DC-logic, IC-logic is built on the top of a propositional language (in our case, the L_{basic}) defining a fuzzy modal operator to express formulas like $I_{\alpha}\varphi$. It indicates that the agent has the intention to achieve φ through the action α, and its truth degree (from 0 to 1) represents a measure of the trade-off between the benefit and counter-effects of achieving φ through α. Moreover, IC-logic is defined in terms of a Lukasiewicz logic in the same way as DC-logic. Also, formulas like $\overline{r} \Rightarrow I\varphi$ will be written as $(I\varphi, r)$. For the technical details and the proof of completeness we refer to [Casali, 2008].

Our system generates intentions through the bridge rule 3, from a positive concrete desire and the set of negative desires that may be achieved through the same action.

5.2.4 Planner Context (PC) and Communication Context (CC):

The logic in the Planner context is a first-order logic restricted to Horn clauses. In this first approach, this context only holds the special predicate *action*, which defines a primitive action together with its precondition. We look forward to introducing plans as a set of actions in the future. Communication context is a functional context as well, and its logic is also a first-order logic restricted to Horn clauses with the special predicates *does* to perform actions, and $rec_j\varphi$ to indicate that the agent has received the communication φ from agent j. They are first order predicates, not modalities.

5.2.5 Repage Context (RC)

The Repage context contains the Repage model. We capture the information that the model computes with the predicates $Img(j, r, [V_{w_1}, V_{w_2}, \ldots])$ and $Rep(j, r, [V_{w_1}, V_{w_2}, \ldots])$, corresponding to the Image and Reputation of agent j playing the role r. See chapter 4 for the details.

$$\mathbf{A}_I\textbf{:} \quad \frac{RC : Img(j, r, [V_{w_1}, V_{w_2}, \ldots])}{\begin{array}{c} BC : E(\mathcal{R}_{rj}, \mathcal{T}_{r,w_1}, V_{w_1}, \{r\}) \\ BC : E(\mathcal{R}_{rj}, \mathcal{T}_{r,w_2}, V_{w_2}, \{r\}) \\ \ldots \end{array}}$$

$$\mathbf{A}_R\textbf{:} \quad \frac{RC : Rep(j, r, [V_{w_1}, V_{w_2}, \ldots])}{\begin{array}{c} BC : S(\mathcal{R}_{rj}, \mathcal{T}_{r,w_1}, V_{w_1}, \{r\}) \\ BC : S(\mathcal{R}_{rj}, \mathcal{T}_{r,w_2}, V_{w_2}, \{r\}) \\ \ldots \end{array}}$$

Figure 5.2: The bridge rules A_I and A_R (see Figure 5.1). They translate Image and Reputation predicates respectively into the belief context.

5.2.6 Bridge Rules

Bridge Rules $\mathbf{A}_I$ and $\mathbf{A}_R$

Bridge rules A_I and A_R (see Figure 5.2) are in charge of generating the corresponding E and S predicates from images and reputations respectively, as explained in section 4.3. The key idea in this interface is that if the image or reputation information changes in Repage, the previously generated E and S predicates will not have the *support* to be valid any more, and thus, they must be out withdrawn from the theory (together with all the inferences performed so far from these predicates), placing the new ones instead. In this way, the theory is always consistent with the information that Repage computes.

5.2.7 Bridge Rules 1, 2, 3, 4

Bridge rules 1 and 2 (see Figure 5.3) transform generic desires to more concrete and realistic desires. To do this, these bridge rules merge generic desires from DC (with absolute values of satisfaction or disgust) with the information contained in BC, which includes the probability to achieve the desire by executing certain action. The result is a desire whose gradation has changed, becoming more realistic. This is calculated by the function g. If we define it as the product of both values, we obtain an expected level of satisfaction/disgust[2].

Bridge rule 3 generates intentions. It takes into account both the expected level of satisfaction and the cost of the action. At the same time, executing an action to achieve certain formula can generate undesirable counter-effects. Thus, bridge rule 3 also takes into account the possible negative desires that can be reached by executing this action. In this bridge rule, for each positive realistic desire (D^+), we must include all negative desires (D^-) that can result from the same action. In this way we have the value of the positive desire (δ^+) and the sum of all negative desires (δ^-) that can be achieved by executing the same

[2] When g is defined as the product, the outcome is very similar to the notion of expected utility used in decision theory.

$$\textbf{1:}\quad \frac{\begin{array}{c} DC : (D^{+}\varphi, d_{\varphi}) \\ BC : B(\alpha, \varphi, p_{\psi}, Q) \end{array}}{DC : (D^{+}_{\alpha}\varphi, g(d_{\varphi}, p_{\psi}))}$$

$$\textbf{2:}\quad \frac{\begin{array}{c} DC : (D^{-}\varphi, d_{\varphi}) \\ BC : B(\alpha, \varphi, p_{\psi}, Q) \end{array}}{DC : (D^{-}_{\alpha}\varphi, g(d_{\varphi}, p_{\psi}))}$$

$$\textbf{3:}\quad \frac{\begin{array}{c} DC : (D^{+}_{\alpha}\varphi, \delta), PC : action(\alpha, P), PC : P \\ DC : (D^{-}_{\alpha}\psi_1, \delta_{\psi_1}), \ldots, (D^{-}_{\alpha}\psi_n, \delta_{\psi_n}) \\ \delta - \sum_{k=1}^{n} \delta_{\psi_k} \geq 0 \end{array}}{IC : (I_{\alpha}\varphi, f(\delta, \sum_{k=1}^{n} \delta_{\psi_k}))}$$

$$\textbf{4:}\quad \frac{IC : (I_{\alpha}\varphi, \epsilon_{max})}{CC : does(\alpha)}$$

$$P, Q, B\textbf{:}\quad \frac{BC : B\varphi}{PC : \varphi}\ , \quad \frac{BC : B\varphi}{RC : \varphi}\ , \quad \frac{CC : rec_j\varphi}{RC : rec_j\varphi}$$

Figure 5.3: The bridge rules 1, 2, 3, 4, P, Q and B (see Figure 5.1).

action. The strength of the intention that is created is defined by a function f. Different f functions would model different behaviors. In our examples we use the following definition: $f(\delta^{+}, \delta^{-}) = max(0, \delta^{+} - \delta^{-})$.

Finally, bridge rule 4 instantiates a unique intention (the one with maximum degree) and generates the corresponding action in the communication context.

5.2.8 Bridge Rules P, Q and B

Bridge rules P and Q allow the planner and Repage context respectively to be aware of the beliefs of the agent. The planner context uses this information to build plans, actions and their preconditions. Repage uses the information to configure the mappings $\mathcal{R}$ and $\mathcal{T}$.

Rule B reflects the reaction of the communication context once it receives communicated images, communicated reputation, third party images from other agents and fulfillment predicates. The content of these communications is directly introduced in Repage, which will update its information.

5.3 An Example

In this section we analyze the reasoning processes performed by an executable version of the model presenting an example.

The base scenario we use involves a BDI agent that, as a manager of a small restaurant, needs to periodically order wine to refill the stock. In this scenario, several providers are available. The information our agent wants to capture about them includes reliable information, for instance the price she will have to pay, but also uncertain information such as the delivery time of the orders and the quality of the wine. While reliable information is introduced as beliefs of probability 1, uncertain information will result in beliefs of lower probability values.

This situation can be formalized in multiple ways. We can define four possible pairwise disjoint predicates for the quality of the wine: *poorWine, averageWine, goodWine, excellentWine* (pW, aW, gW and eW from now on) and five pairwise disjoint predicates for the delivery time: $days(0,1)$*,* $days(2,3)$*,* $days(4,5)$*,* $days(6,10)$*,* $days(11,\infty)$ indicating respectively a delivery time up to 1 day, between 2 and 3 days etc. Also we define the predicates *paid(X), paidLess(X), paidMore(X)* to indicate that the agent has paid X, less than X and more than X respectively,and the implication relation $paid(X) \rightarrow paidLess(Y)$ when $X < Y$, and $paid(X) \rightarrow paidMore(Y)$ when $X > Y$. The predicate *budget(X)* indicates that the money she has in the budget is X. This knowledge and the implication among predicates must be introduced also as beliefs.

The interaction model defining the purchase of wine indicates that providers act as *wineSeller*s, but agent i wants to evaluate them in the two independent dimensions: the quality of the wine and the delivery time. Thus, Repage uses the roles $wineSeller(quality)$ and $wineSeller(dTime)$. The mapping $\mathcal{R}$ (see section 4.3.1) of these two roles points to the same action *buyWine* (*buy* from now on), which then summarizes the entire interaction model. The mapping $\mathcal{T}$ of the role $wineSeller(quality)$ relates w_1 to $poorWine$, w_2 to $averageWine$ etc, and the mapping $\mathcal{T}$ of the role $wineSeller(dTime)$ relates w_1 with $days(0,1)$, w_2 with $days\ (2,3)$, etc.

5.3.1 The Initial Knowledge

In this world, our agent knows the existence of four providers represented by *alice, bob, charlie* and *debra* respectively. Our agent is aware of their prices, and so this knowledge is introduced as beliefs:

$$\begin{gathered} B(buy(alice), hasWine \wedge paid(1000), 1, e_\emptyset) \\ B(buy(bob), hasWine \wedge paid(900), 1, e_\emptyset) \\ B(buy(charlie), hasWine \wedge paid(400), 1, e_\emptyset) \\ B(buy(debra), hasWine \wedge paid(1300), 1, e_\emptyset) \end{gathered} \tag{5.1}$$

Bridge rule P introduces the information above into the planner context in order to generate the corresponding plans (simple actions in this case). It follows then, that in PC we find

$$action(buy(alice), hasMoreMoney(1000))$$

indicating that the action of buying wine from *alice* is preconditioned on the budget having more than 1000.

5.3.2 Study Cases

Exploring the Space: Case 1

Our agent is new to the business and only *trusts* her own direct experiences. It means that axiom IRB uses a function h of the class $\mathcal{H}_1$. Since she is just starting the business, she is mostly concerned about the quality of the wine rather than the delivery time. She has a budget of 1350 ($budget(1350)$) for the purchase. Regarding her desires, she would be satisfied with paying up to 1350 for an excellent wine. With the same strength she would be satisfied paying up to 800 for a good wine. In any case, she needs the wine. What she does not want is a poor or average wine. Lower on her priority list is obtaining the wine quickly, but still a long delivery time is not desired. These preferences can be formalized as desires in the DC as follows:

$$\begin{gathered}(D^+(hasWine \land paidLess(1350) \land eW), .9)\\(D^+(hasWine \land paidLess(800) \land gW), .9)\\(D^+ hasWine, .7)\\(D^- pW, 1)\\(D^- aW, .8)\\(D^- days(11, \infty), .5)\\(D^- days(6, 10), .4)\end{gathered} \tag{5.2}$$

Since she does not have any information about the providers, Repage predicates contain the maximum possible uncertainty. For instance, the corresponding image predicates for *charlie* are:

$$\begin{gathered}Img(charlie, wineSeller(quality), [.25, .25, .25, .25])\\Img(charlie, wineSeller(time), [.2, .2, .2, .2, .2])\end{gathered} \tag{5.3}$$

Under these conditions the reasoning process leads to a random choice between three agents (*charlie*,*bob* and *alice*) to achieve the desire $hasWine$. In the following lines we briefly explain the most relevant steps.

Bridge rule A_I generates beliefs in the BC from images. As said before, the epistemic decision is not done at this rule but inside Repage, which computes image and reputation. In the case of *charlie* this rule is activated regarding the role $wineSeller(quality)$ as:

$$\frac{RC : Img(charlie, wineSeller(quality), [.25, .25, .25, .25])}{\begin{gathered}BC : E(buy(charlie), pW, .25, \{wineSeller(quality)\})\\BC : E(buy(charlie), aW, .25, \{wineSeller(quality)\})\\BC : E(buy(charlie), gW, .25, \{wineSeller(quality)\})\\BC : E(buy(charlie), eW, .25, \{wineSeller(quality)\})\\\dots\end{gathered}} \tag{5.4}$$

All possible outcomes after buying from *charlie* have the same probability. This rule also generates the probabilities of disjoint formulas:

$BC : E(buy(charlie), pW \vee aW, .50, \{wineSeller(quality)\})$
$BC : E(buy(charlie), pW \vee gW, .50, \{wineSeller(quality)\})$
$BC : E(buy(charlie), pW \vee eW, .50, \{wineSeller(quality)\})$
$BC : E(buy(charlie), aW \vee gW, .50, \{wineSeller(quality)\})$
$BC : E(buy(charlie), aW \vee eW, .50, \{wineSeller(quality)\})$
$BC : E(buy(charlie), gW \vee eW, .50, \{wineSeller(quality)\})$
$BC : E(buy(charlie), pW \vee aW \vee gW, .75, \{wineSeller(quality)\})$
$BC : E(buy(charlie), pW \vee gW \vee eW, .75, \{wineSeller(quality)\})$
$BC : E(buy(charlie), aW \vee gW \vee eW, .75, \{wineSeller(quality)\})$
$BC : E(buy(charlie), pW \vee aW \vee gW \vee eW, 1, \{wineSeller(quality)\})$

The previous E predicates are directly transformed to B predicates through the axiom IRB, which uses a h function belonging to $\mathcal{H}_1$ (only images are taken into account). In BC, because of the assumption that the quality and delivery time dimensions are stochastically independent, probabilistic inference rules of the L_{BC} theory are applied. For example, from $B(buy(charlie), eW, .25, \{wineSeller(quality)\})$ and $B(buy(charlie), days(0,1), .2, \{wineSeller(time)\})$ can be deduced

$$B(buy(charlie), eW \wedge days(0,1), .05, \{wineSeller(quality), wineSeller(time)\})$$

, where .05 is the product of .25 and .2. In particular, and for the interest of our example, the following belief is also generated:

$B(buy(charlie), hasWine \wedge paid(400) \wedge eW, .25,$
$\{wineSeller(quality), wineSeller(time)\})$

Bridge rules 1 and 2 are executed for each generic positive and negative desire respectively. For instance, rule 1 is fired for the first desire as follows:

$$\frac{\begin{array}{c} DC : (D^{+}(hasWine \wedge paidLess(1350) \wedge eW), .9) \\ BC : B(buy(charlie), hasWine \wedge paidLess(1350) \wedge eW, .25, Q) \end{array}}{(D^{+}_{buy(charlie)}(hasWine \wedge paidLess(1350) \wedge eW), g(.9, .25))} \quad (5.5)$$

If we consider that $g(p, q) = p \cdot q$, the resulting grade of the positive concrete desire is .225. It indicates that performing the action of buying from *charlie* to obtain an excellent wine and paying less than 1350 has an expected level of satisfaction of .225. Of course, for the same desire bridge rule 1 can be executed several times because different actions can lead to the same desire. Negative desires fire bridge rule 2 generating concrete negative desires. They indicate the expected level of disgust if the action is executed.

These negative desires are used in bridge rule 3 to take into account possible counter-effects of satisfying certain desire. Rule 3 is executed only one time for each positive concrete desire. For example, considering the desire above:

$$\begin{array}{c}
DC : (D^{+}_{buy(charlie)}(hasWine \wedge paidLess(1350) \wedge eW), .225) \\
DC : (D^{-}_{buy(charlie)}days(11, \infty), .08) \\
DC : (D^{-}_{buy(charlie)}days(6, 10), .08) \\
DC : (D^{-}_{buy(charlie)}aW, .2) \\
DC : (D^{-}_{buy(charlie)}pW, .25) \\
PC : action(buy(charlie), budgetMore(400)) \\
PC : budget(1100) \rightarrow budgetMore(400) \\
\hline
IC : (I_{buy(charlie)}(hasWine \wedge paidLess(1350) \wedge eW), f(.225, .61))
\end{array} \tag{5.6}$$

In this case, notice that the expected level of satisfaction of achieving the desire by buying from *charlie* is .225 but its counter-effects bring an expected level of disgust of .61. Taking $f(\delta^{+}, \delta^{-}) = max(0, \, \delta^{+} - \, \delta^{-})$, this intention has a grade of 0. Why would we perform an action if we expected from it to obtain more disgust than benefit?.

If the intention had the maximum degree, bridge rule 4 would generate the corresponding action. In our example, after calculation, the intentions with a grade higher than 0 result to be:

$$\begin{array}{c}
(I_{buy(charlie)}hasWine, .14) \\
(I_{buy(bob)}hasWine, .14) \\
(I_{buy(alice)}hasWine.14)
\end{array} \tag{5.7}$$

As expected, since Repage does not have any information and our agent needs to buy wine, a random choice can be made among these possibilities. Buying from *debra* is not considered because in rule 3 the precondition of having a budget greater than 1300 does not hold (see the action definition in the planner context). Assuming that she picks ($I_{buy(charlie)}$,$hasWine$,.14), bridge rule 4 is fired executing the action $buy(charlie)$.

The result of this transaction fulfills the agent's desires in terms of delivery time and quality. This information is inserted into Repage by means of the bridge rule B. Repage evaluates the outcomes and updates the values of image and reputation. In the next reasoning process, this information will be introduced as beliefs by bridge rule A_I and A_R, as we have shown at the beginning of this case.

Continuing with our example, we suppose that *charlie* delivers the wine quite fast, in less than one day, but the quality of the wine is not very good. This makes Repage update image predicates as

$$\begin{array}{c}
Img(charlie, wineSeller(quality), [.4, .4, .1, .1]) \\
Img(charlie, wineSeller(time), [.45, .25, .1, .1, .1])
\end{array} \tag{5.8}$$

We recall here that $w_1, w_2, \ldots$ in the role $wineSeller(quality)$ correspond to pW, aW,... meanwhile in the role $wineSeller(\ time)$ they correspond to $days(0, 1)$, $days(2, 3)$, ... respectively.

Receiving Reputation Information: Case 2

After a while, our agent needs to buy more wine. She has exactly the same desires as before and the same budget, so she is mainly interested in the quality of the wine rather than delivery time. But this time, her image information about *charlie* has changed. Furthermore, we assume that she has received several reputation communications, about both *charlie* and *alice*. This information makes Repage generate the following reputation predicates:

$$\begin{array}{c} Rep(charlie, wineSeller(quality), [.5, .3, .1, .1]) \\ Rep(alice, wineSeller(quality), [.1, .2, .2, .4]) \end{array} \tag{5.9}$$

The reputation information regarding *charlie* coincides more or less with the image our agent has about him. This is not the case with *alice*. Through bridge rule A_R these predicates generate beliefs into BC. For *charlie*:

$$\frac{RC : Rep(charlie, wineSeller(quality), [.5, .3, .1, .1])}{\begin{array}{c} BC : S(buy(charlie), pW, .5, \{wineSeller(quality)\}) \\ BC : S(buy(charlie), aW, .3, \{wineSeller(quality)\}) \\ BC : S(buy(charlie), gW, .1, \{wineSeller(quality)\}) \\ BC : S(buy(charlie), eW, .1, \{wineSeller(quality)\}) \end{array}} \tag{5.10}$$

Note that these beliefs refer to what others say, not what our agent really believes. Since our agent only *trusts* herself, she does not take into account these predicates. In terms of the *BC*-logic it indicates that there is no relationship between operator S and operator B_i so far. This situation is also common: we can accept that a given person has a bad reputation, that most people *say* this, even when we believe the opposite [Conte and Paolucci, 2002].

Under these conditions, the reasoning process is similar to the previous case. This time though, *charlie* is no longer a possible choice, since the last experience with him was bad regarding the quality of the wine. Bridge rule 3 generates the intention to buy from *charlie* with a very low grade, in fact zero, since it is likely a poor or average wine would be delivered. In this case, the generated intentions are

$$\begin{array}{c} (I_{buy(bob)} hasWine, .14) \\ (I_{buy(alice)} hasWine, .14) \end{array} \tag{5.11}$$

Our agent chooses *alice*. This time we suppose the result is in tune with the expectations of our agent; she obtains a good wine, even though the delivery time is not very fast. Repage updates image predicates regarding *alice* as follows:

$$\begin{array}{c} Img(alice, wineSeller(quality), [0, 0, .15, .85]) \\ Img(alice, wineSeller(time), [0, 0, 0, .1, .9]) \end{array} \tag{5.12}$$

Keeping the Same Desires: Case 3

Maintaining the exact same desires as case 1 and 2, the next time that our agent wants to buy wine, she has the following intentions whose grade is higher than

0:

$$\begin{array}{c}(I_{buyWine(bob)}hasWine, .14),\\ (I_{buyWine(alice)}hasWine, .35)\\ (I_{buyWine(alice)}(hasWine \wedge paidLess(1350) \wedge eW), .365)\end{array} \tag{5.13}$$

Since *alice* provided wine that was mostly excellent, and this is the main concern of our agent, she chooses again to buy from *alice*, but to satisfy the desire $hasWine \wedge paidLess(1350) \wedge eW$. The option to buy from *bob* appears due to the uncertainty around his performance. We suppose that the resulting transaction confirms the same results as the previous case: an excellent wine but a long delivery time.

Changing Desires: Case 4

This time our agent accepts the suddenly request to host a big birthday banquet that will take place in less than 12 days. Her cellar is not prepared for this event, so, she needs to order more wine. In this situation, her desires are different, since delivery time is now a key issue while the quality of the wine drops in importance:

$$\begin{array}{c}(D^{+}hasWine \wedge paidLess(1350) \wedge days(0,1), .9)\\ (D^{+}hasWine \wedge paidLess(800) \wedge days(2,3), .7)\\ (D^{-}pW, .2)\\ (D^{-}aW, .2)\\ (D^{-}days(11,\infty), .8)\\ (D^{-}days(6,10), .7)\end{array} \tag{5.14}$$

Thanks to her previous interactions with the providers our agent already has some information about their performance. In this case, the only intention with a degree higher that 0 is

$$(I_{buyWine(charlie)}(hasWine \wedge paidLess(1350) \wedge days(0,1)), .095)$$

She picks *charlie*, and the results are like the first time she bought from him in case 1: a short delivery time but a low quality.

Using Reputation Information: Case 5

Several weeks after the successful banquet, our agent recuperates her initial desires and needs to order wine again. During this time she has heard about both *bob* and *debra*'s reputations which indicates that both offer excellent wines and that furthermore *debra* is capable to deliver the order in a day. This is not the case with *bob*:

$$\begin{array}{c}Rep(bob, wineSeller(quality), [0, 0, .05, .95])\\ Rep(bob, wineSeller(time), [.1, .2, .3, .3, .1])\\ Rep(debra, wineSeller(quality), [0, 0, 0, 1])\\ Rep(debra, wineSeller(time), [1, 0, 0, 0, 0])\end{array} \tag{5.15}$$

This information is introduced through rule A_R as S predicates. Unfortunately for our agent, *alice* notifies that she will not be available this time because she will be on holidays. Because of that, and because the reputation information she received in case 2 was in concordance with what she really believed, our agent starts *trusting* what others gossip. In this new scenario, the IRB axiom is set to use a h function belonging to $\mathcal{H}_2$ (only reputation is taken into acount). Thus, the axiom IRB states the following:

$$\forall a x p_1 p_2 r (I(a, x, p_1, r) \wedge S(a, x, p_2, r)) \rightarrow B(a, x, p_2, r)$$

It means that reputation predicates from Repage, once they have been inserted into the BC-context as S predicates, they become belief predicates. For instance, regarding *bob* in the role of $wineSeller(quality)$, rule A_R generates, among others, the following predicate: $S(buy(bob)$, eW, .95, $\{wineSeller(quality)\})$, meaning that people is gossiping that with is a probability of .95, the wine will be excellent when buying from *bob*. Since our agent *believes* what it gossiped due to axiom IRB, it can be deduced that $B(buy(bob)$, eW, .95, $\{wineSeller(quality)\})$. In this case, the only non-zero graded intention generated is

$$(I_{buy(bob)}(hasWine \wedge paidLess(1350) \wedge eW), .565)$$

From the activation of bridge rule 3 as follows:

$$\frac{\begin{array}{c} DC : (D^{+}_{buy(bob)}(hasWine \wedge paidLess(1350) \wedge eW), .0.855) \\ DC : (D^{-}_{buy(bob)} days(11, \infty), .08) \\ DC : (D^{-}_{buy(charlie)} days(6, 10), .21) \\ PC : action(buy(bob), budgetMore(900)) \\ PC : budget(1100) \rightarrow budgetMore(900) \end{array}}{IC : (I_{buy(bob)}(hasWine \wedge paidLess(1350) \wedge eW), f(.855, .29))} \tag{5.16}$$

We suppose in this situation that the results are not as the agent expects, obtaining an average wine. Thus, Repage image predicates are updated as:

$$\begin{array}{c} Img(bob, wineSeller(quality), [.3, .4, .2, .1]) \\ Img(bob, wineSeller(time), [.1, .2, .3, .3, .1]) \end{array} \tag{5.17}$$

Image and Reputation Interference: Case 6

Note that in the previous situation, the image about *bob* in the role $wineSeller(quality)$ contradicts *bob*'s reputation in the same role. This has already happened in case 2 with *alice*, but axiom IRB was only taking into account image information. in this new case, we assume that the IRB uses a h function from the class $\mathcal{H}_4$, where both image and reputation are taken into account but image is more important. As a matter of example, we set function h as

$$h(p_E, p_S) = \frac{7 \cdot p_E + 3 \cdot p_S}{10}$$

We show how the reasoning process proceeds. Regarding the role $wineSeller(quality)$, through bridge rule A_I the following E predicates are generated into the belief context:

$$\begin{array}{l} E(buy(bob), pW, 0.3, \{wineSeller(quality)\}) \\ E(buy(bob), aW, 0.4, \{wineSeller(quality)\}) \\ E(buy(bob), gW, 0.2, \{wineSeller(quality)\}) \\ E(buy(bob), eW, 0.1, \{wineSeller(quality)\}) \\ \ldots \end{array}$$

and through bridge rule A_R the following:

$$\begin{array}{l} S(buy(bob), pW, 0, \{wineSeller(quality)\}) \\ S(buy(bob), aW, 0, \{wineSeller(quality)\}) \\ S(buy(bob), gW, 0.05, \{wineSeller(quality)\}) \\ S(buy(bob), eW, 0.95, \{wineSeller(quality)\}) \\ \ldots \end{array}$$

Then, the presence of axiom IRB with the h function defined above combines both predicates generating a new probability distribution. In this case:

$$\begin{array}{l} B(buy(bob), pW, 0.21, \{wineSeller(quality)\}) \\ B(buy(bob), aW, 0.28, \{wineSeller(quality)\}) \\ B(buy(bob), gW, 0.155, \{wineSeller(quality)\}) \\ B(buy(bob), eW, 0.355, \{wineSeller(quality)\}) \\ \ldots \end{array}$$

In this way we preserve the properties of probability distributions, reflecting in the resulting beliefs a combination of the both source of information: image and reputation from Repage.

Turning again to the example above, note that the resulting beliefs for *bob* presents a distribution that model an almost uncertain distribution, here values are close to 0.25. This make sense since image and reputation information regarding *bob* where quite contradictory. In this situation, our agent picks *alice*.

Increasing the Budget: Case 7

To conclude, we want to show the effect of a simple environment change. In this case, our agent decides to increase the wine budget to 2000. With exactly the same desires and the same reputation and image information as before, the reasoning process generates the maximum intention to buy from *debra*. This provider was always filtered out at bridge rule 3 because the precondition of buying from *debra* (to have more than 1300) was never fulfilled. Thus, the intention to buy from *debra* is only slightly higher than buying from *alice*.

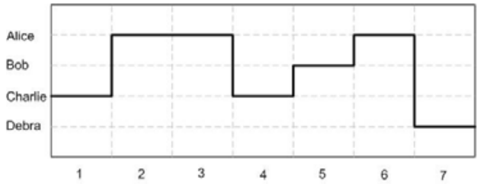

Figure 5.4: The choices of the agent throughout the situations explained in this section

5.3.3 Implementation Details

The scenario and each one of the situations have been implemented in Prolog[3]. An implementation of logical systems usually entails the simplification or limitation of some aspects of the logic. In our case, we assume that each logical formula is expressed as a Horn clause and that modal operators are first-order predicates. Also, we do not accept logically omniscient agents that use a forward-reasoning engine, even when some implementations of multi-context systems use this approach [Sabater-Mir et al., 2002]. Instead, we take advantage of the backward-reasoning engine of Prolog.

Note that the multi-context system specification of our BDI agent models an agent whose purpose is to execute a single action. This action is generated through rule 4 by choosing the intention of maximum grade. For this choice the agent must generate all possible intentions, which are created through rule 3 from desires, and so on. This schema follows a backward-reasoning algorithm that can be implemented in Prolog.

Thus, considering predefined knowledge as Prolog predicates, and inference rules and bridge rules as Prolog rules, the agent's reasoning can be started by asking Prolog to satisfy the predicate $does(A)$. While this is an oversimplification of what should be understood as multi-context systems, for simple examples the results are coherent and useful. We plan to study implementation issues in the future, an the effects of the simplifications in the desirable properties of the system.

5.4 Extending the BDI+Repage Architecture : The Norm Context

In this section we show how the organizational mechanism describes in [Centeno et al., 2009a] is integrated into the BDI+Repage defined above, showing the flexibility of our model. In this case, we assume that agents are aware of the norms of the society, and due to that, are capable of evaluating other agents' behaviour according to them. In our integrated model, such evaluations are computed by the reputation model Repage, and through the appropriate

[3] The source code can be download at
`http://www.iiia.csic.es/∼ipinyol/sourceJAAMAS09.zip`.

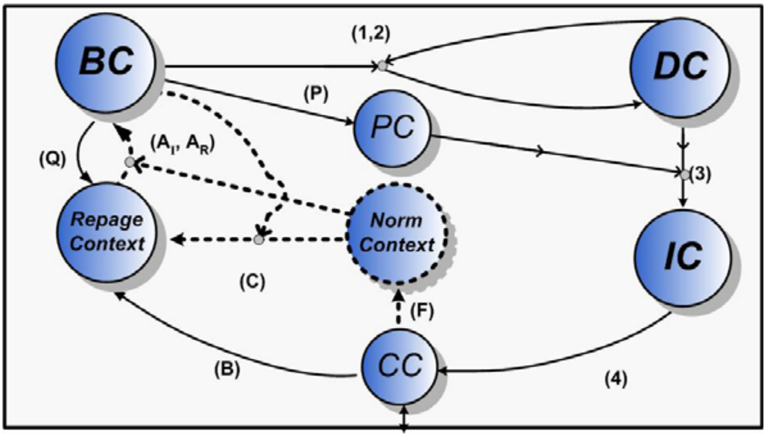

Figure 5.5: A graphical representation of the BDIRepage+Norm model. Elements with dot lines are the new elements introduced in this section.

bridge rules, the knowledge is introduced into the belief base of the agent. This information then can be used in the normal reasoning BDI process of the agent as shown earlier

5.4.1 Preliminaries

In this extension, we deal with the notion of personal and organizational norms, that was introduced in [Centeno et al., 2009a]. Such organization model formalizes a particular type of organized multiagent system - from now on *organization* - following the framework proposed in [Centeno et al., 2009a] that provides a minimum set of mechanisms to regulate agents' interactions: $\mathcal{R}^{om}$ and $\mathcal{ON}^{om}$. A $\mathcal{R}^{om}$ is an organizational mechanism based on roles that defines the positions agents may enact in the organization.

Formally, an organization is defined as a tuple $\langle \mathcal{A}g, \mathcal{A}, \mathcal{X}, \phi, x_0, \varphi, \{\mathcal{ON}^{om}, \mathcal{R}^{om}\}\rangle$ where $\mathcal{A}g$ represents the set of agents participating within the organization; $\mathcal{A}$ is the set of actions agents can perform; $\mathcal{X}$ stands for the environmental states space; ϕ is a function describing how the system evolves as a result of agents actions; x_0 represents the initial state of the system; φ is the agents' capability function describing the actions agents are able to perform in a given state of the environment; $\mathcal{ON}^{om}$ is an organizational mechanism based on organizational norms; and $\mathcal{R}^{om}$ is an organizational mechanism based on roles that defines the positions agents may enact in the organization (see [Centeno et al., 2009b] for more details).

A $\mathcal{R}^{om}$ is an organizational mechanism [Centeno et al., 2009a], that attempts to regulate agents' interactions by providing different positions to agents. Role characteristics are: *i*: provides a first-order block to build organized multiagent systems; *ii*: encapsulates a set of functionalities an agent playing such a role is permitted to perform in the system; and *iii*: informs about an expected behavior that an agent playing such a role should show.

Agents participating in an organization are involved in different *situations* through the time. Situations represent an agent – in $\mathcal{A}g$ – playing a role, defined by $\mathcal{R}^{om}$, performing an action in the system – in $\mathcal{A}$. A situation is defined as a tuple $\langle \mathcal{A}g, \mathcal{R}, \mathcal{A}, T \rangle$, that is, it defines an agent ($\mathcal{A}g$) playing a role ($\mathcal{R}$) in certain action ($\mathcal{A}$) during a time period (T).

A $\mathcal{ON}^{om}$ is an organizational mechanism that regulates participants' behavior by using norms, and it is the part of the organizational mechanism that is relevant in this section. An organizational norm is defined as a tuple $\langle deon, Sit, Org \rangle$, where *deon* is a deontic concept in the set {PROHIBITION, OBLIGATION, PERMISSION} representing the different constraining possibilities over the situation *Sit* (where an agent is playing a role and executing an action) within the organization *Org*.

Agents in an organization are supposed to have their own preferences and goals. In [Centeno et al., 2009b] the concept of *personal norm* is proposed to represent agent's preferences over different situations in which other agents may be involved. Thus, a personal norm models how an agent wants the others to behave when interacting with it. A personal norm is defined as a tuple $\langle \mathcal{A}g, deon, Sit \rangle$, where $\mathcal{A}g$ is the owner of the norm, *deon* is a deontic concept in the set {PROHIBITION, OBLIGATION, PERMISSION} representing the preferences of agent $\mathcal{A}g$ over the situation *Sit*.

5.4.2 Norms and the BDI+Repage Model: An example

Let us consider a supply chain (SC) formed by beverage/food providers and pubs. Pubs contact the beverage and food providers with the aim of buying the goods that they later will sell to their customers. The following roles participate in such SC:

Providers	sell their goods to the *Pubs*.
Pubs	buy beverages and foods from *Providers* and sell them to *Customers*.
Customers	buy the beverages and foods sold by *Pubs*.

For our example we stress on the relationships between pubs and providers. Those relationships are regulated by some market rules, that all participants must fulfil. In the scenario we take the perspective of a BDI agent (from now on *our* agent, or agent *i*) that represents a pub owner. This agent needs very often to place orders to refill the stock. *Our* agent has a set of possible providers to choose from, and makes the selection following certain criteria (monetary cost, delivery time, quality of the product, etc.). One of these criteria is the observance of norms. For instance, one of the organizational norms that rules our scenario is:

- **ON** - *Orders must not be delivered later than 7 days after the date they* *were placed.*

1. Norm **ON** is evaluated after the action *placeOrder* is performed by an agent playing the role *pub*.
2. This evaluation can be done because after the action, a *fulfilment* indicates that the number of days for the delivery was exactly $dTime$.
3. If $dTime < 7$ the norm is fulfilled while if $dTime \geq 7$ the norm is violated. In both cases, this information is taken into account by the reputation model for future interactions.

Notice that it is not the same to deliver the product in 8 days than in 20. For this, we introduce the concept of *evaluative patterns* of a norm, which enriches the reasoning capabilities of the agent. Following the example, we consider four evaluative patterns for ON: $dTime < 7$, $7 \leq dTime < 9$, $9 \leq dTime < 15$, $15 \leq dTime$. After a transaction, the fulfillment of the norm regarding $dTime$ is classified in one and only one of the previous evaluative patterns. This information is introduced into the Repage context. Then Repage computes a probabilistic distribution over the four possible patterns that estimates the potential behavior of the agent playing the role seller.

As we will see in section 5.4.5, two bridge rules introduces such evaluations as beliefs. Once this step is performed, desires start playing an important role for the practical reasoning process. On the one hand, Repage information provides for each agent evaluations according to the evaluative patterns. On the other side, the desires of our agent determine a preference between each one of the situations. For instance, our agent i can have the following desire: $(D^{+}dTime < 7, 1)$ indicating that i wants to achieve a $dTime$ lower than 7 days with a strength of 1. So, she wants the norm completely fulfilled. However, in another situation we could have: $(D^{+}dTime < 7, 1), (D^{+}7 \leq dTime < 10, 0.7), (D^{-}10 \leq dTime, 1)$. In this case, agent i wants with maximum strength a delivery time below 7 days, but also would consider a delivery time between 7 and 10 days, with less strength (0.7). What agent i rejects with maximum strength is a delivery time higher or equal than 10.

We argue that the separation between an objective evaluation and the desired behaviour is crucial for real autonomous entities. Then, an agent can change the desires but keeping and using the same evaluations. In the following sections we formally describe the new BDI+Repage+Norm model.

5.4.3 The Norm Context (NC)

The new BDI+Repage+Norm multicontext model is represented with the tuple $Ag = \langle\{BC, DC, IC, PC, CC, RC, NC\}, \triangle_{br}\rangle$. These correspond to Belief, Desire, Intention, Planner, Communication and Repage contexts, respectively, plus a new NC (norm context). The set of bridge rules $\triangle_{br}$ incorporates the original rules $1, 2, 3, 4, P, Q$ and B, shown in Figure 5.3, plus the modified rules A_I and A_R (section 5.4.5), and rules F, R and C (section 5.4.4) that are new and related to the norm context. Figure 5.5 shows a graphical representation of this multicontext specification.

To specify NC we define the language L_{norm} as a first-order language with the special predicates $F(\cdot)$ and $N(\cdot)$ to model fulfilments and evaluative patterns respectively. We restrict the language to a conjunction of such predicates. It is important to remark that the language is used to describe how the norms are evaluated. Thus, there is no reference to the deontic concepts of the norm, which are implicit in the description and in the desires of the agent.

The syntax of L_{norm}

The two special predicates in L_{norm} are identified by their sorts. The sorts that L_{norm} includes are a finite set of agent identifiers $\mathcal{A}$, a finite set of role identifiers $\mathcal{R}$, the finite set $\mathcal{I} \subset \mathbb{N}$ of indexes to identify each evaluative pattern of a norm and a countable set of time instants $\mathcal{T}$ to represent the time that fulfillments are produced. To express the content of the normative patterns and fulfillments we need an object language that *talks* about the domain and that must be the same used in the beliefs, desires and intentions. Such language is L_{basic}, defined in chapter 4 and used earlier in this chapter. Again, we introduce each φ belonging to the set of well-formed formula of L_{basic} ($wff(L_{basic})$) as a constant $\ulcorner\varphi\urcorner$ of L_{norm}. In the examples, we omit the quote $\ulcorner\cdot\urcorner$.

Let $\varphi, \phi \in wff(L_{basic})$, $j \in \mathcal{A}$, $r \in \mathcal{R}$, $n \in \mathcal{I}$ and $t \in \mathcal{T}$, the predicates of the language are:

- $N(n, r, \varphi)$: It describes an evaluative pattern for a given role. For instance, the previous example that has four evaluative patterns for the norm ON can be represented as

  $$\begin{array}{l} N(1, provider(ON), dTime < 7) \\ N(2, provider(ON), 7 \leq dTime < 9) \\ N(3, provider(ON), 9 \leq dTime < 15) \\ N(4, provider(ON), 15 \leq dTime) \end{array}$$

  Since each role can be evaluated by different norms, we consider evaluative patterns for each role × norm, as shown in the example ($provider(ON)$).

- $F(j, r, \phi, t)$: It indicates that after an interaction with agent j playing the role r at time t, ϕ holds. For instance, the formula

  $$F(j, provider(ON), dTime = 6, 2)$$

  indicates that the result of the interaction with agent j playing the role *provider* at time 2 has been a delivery time of 6 days. Again, we write $F_i(j, r, \varphi, t)$ to indicate that agent i is the holder of the predicate.

For a consistent interpretation of the norm context, we require that L_{basic} predicates involved in evaluative patterns of the same role are pairwise disjoints. Formally, let us consider the set of evaluative patters over the role r: $N(1, r, \varphi_1), N(2, r, \varphi_2), \ldots, N(p, r, \varphi_p)$. Then, we must guarantee that for each m, n such that $m \neq n$ and $1 \leq m, n \leq p$, it happens that $\varphi_n \wedge \varphi_m \vdash_{basic} \perp$, where $\vdash_{basic}$is a classical logical consequence defined over L_{basic}. This ensures that two or more evaluative patterns do not cover the same space.

Intuitively, the evaluative patterns classify the possible results that the agent wants to evaluate, providing semantics to the evaluation of norms. After each transaction, the fulfillment is captured by $F(\cdot)$ predicates in the NC-context. Through the appropriate bridge rules, the information is introduced into the Repage context as outcomes. This mechanism is explained in the next subsection.

5.4.4 Rules F and C

On the one hand, rule F is in charge of introducing fulfillments into the norm context, in the form of $F(\cdot)$ predicates . We assume that the communication context is able to capture the fulfillment of the transactions and generate such predicates (it is domain-dependent).

On the other hand, Rule C is in charge of generating outcome predicates to feed the Repage model. It is defined as:

$$\textbf{C:} \qquad \frac{\begin{array}{c} NC : N(n, r, \varphi) \\ NC : F(j, r, \phi, t) \\ BC : B(\iota, \phi \rightarrow \varphi, 1, e_{\emptyset}) \end{array}}{RC : Outcome(j, r, n, t)}$$

Again, following the example, if agent i after interacting with j generates through rule F the predicate $F(j, seller, dTime = 8, t)$, rule C would fire as

$$\textbf{C:} \qquad \frac{\begin{array}{c} NC : N(2, provider(ON), 7 \leq dTime \leq 10) \\ NC : F(j, provider(ON), dTime = 7, t) \\ BC : B(\iota, (dTime = 8) \rightarrow (7 \leq dTime \leq 10), 1, e_{\emptyset}) \end{array}}{RC : Outcome(j, provider(ON), 2, t)}$$

Under the assumption that the norm context is consistent as defined above, rule C is only fired one time for each fulfillment. With outcome predicates, Repage is able to calculate a probability distribution for each agent and role over the defined evaluative patterns.

From outcomes and communications, Repage generates image and reputation predicates, and through rules A_I and A_R the knowledge is introduced into the belief context.

5.4.5 Rules A_I and A_R

In the original BDI-Repage model these rules are in charge of updating the beliefs of the agent with the information coming from the reputation model. In the extended model we have modified the original rules to take into account the information contained in the norm context:

$$
\begin{array}{cc}
\mathbf{A}_I\textbf{:} & \mathbf{A}_R\textbf{:} \\
\\
\dfrac{\begin{array}{c} RC : img(j, r, [V_1, V_2, \ldots]) \\ NC : N(1, r, \varphi_1) \\ NC : N(2, r, \varphi_2) \\ \ldots \end{array}}{\begin{array}{c} BC : E(\mathcal{R}_{rj}, \varphi_1, V_1, \{r\}) \\ BC : E(\mathcal{R}_{rj}, \varphi_2, V_2, \{r\}) \\ \ldots \end{array}}
&
\dfrac{\begin{array}{c} RC : rep(j, r, [V_1, V_2, \ldots]) \\ NC : N(1, r, \varphi_1) \\ NC : N(2, r, \varphi_2) \\ \ldots \end{array}}{\begin{array}{c} BC : S(\mathcal{R}_{rj}, \varphi_1, V_1, \{r\}) \\ BC : S(\mathcal{R}_{rj}, \varphi_1, V_2, \{r\}) \\ \ldots \end{array}}
\end{array}
$$

The key idea is that each linguistic label of the probability distribution provided by Repage and a role r refers to a unique evaluative pattern, i.e. a single predicate N (notice that this mechanism implements the mapping $\mathcal{T}$ introduced in the previous chapter). Also, since an agent j in a role r determines a concrete interaction model (the mapping $\mathcal{R}$ introduced also in the previous chapter), the agent can infer the probability to achieve certain results after interacting with j in the role r.

To illustrate this, imagine that agent i has interacted with j as *provider* several times, and that most of the times the delivery time was below 7 days ($dTime < 7$). Assuming the evaluative patterns for norm ON in the example of section 5.4.3, Repage may have generated the following image predicate $img(j, provider(ON), [0.8, 0.1, 0.1, 0])$. In this situation, rule A_I is fired instantiated as follows, assuming that $\mathcal{R}_{provider,j}$ is the action $order(j)$

$$
\mathbf{A}_I\textbf{:} \quad \dfrac{\begin{array}{c} RC : img(j, provider(ON), [.8, .1, .1, 0]) \\ NC : N(1, provider(ON), dTime < 7) \\ NC : N(2, provider(ON), 7 \leq dTime < 9) \\ NC : N(3, provider(ON), 9 \leq dTime < 15) \\ NC : N(4, provider(ON), 15 \leq dTime) \end{array}}{\begin{array}{c} BC : E(order(j), dTime < 7, .8, \{provider(ON)\}) \\ BC : E(order(j), 7 \leq dTime < 9, .1, \{provider(ON)\}) \\ BC : E(order(j), 9 \leq dTime < 15, .1, \{provider(ON)\}) \\ BC : E(order(j), 15 \leq dTime, 0, \{provider(ON)\}) \end{array}}
$$

5.4.6 An Example

Let us consider a supply chain (SC) formed by beverage and food providers, and pubs. Pubs contact the beverage and food providers with the aim of buying the goods they sell to their customers afterwards. The following roles participate in such SC:

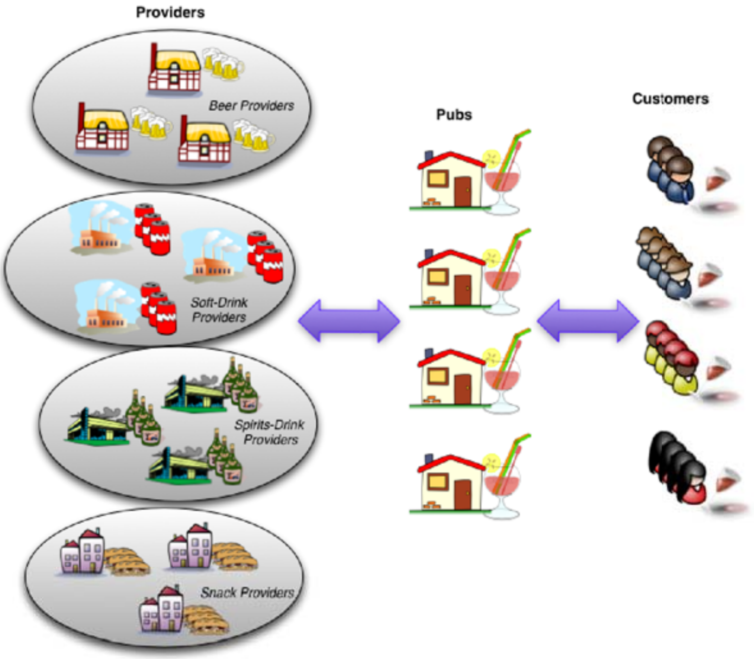

Figure 5.6: Example of Bar Supply Scenario

Providers	sell their goods to the *Pubs*.
Pubs	buy beverages and foods from *Providers* and sell them to *Customers*.
Customers	buy the beverages and foods sold by *Pubs*.

The scenario we present stresses on the relationships between pubs and providers. Those relationships are regulated by some market rules, that all participants must fulfill. For instance, providers must not change the agreed price for an order or, a customer cannot pay less than the price fixed by a pub for a drink. Furthermore, each participant could have their own preferences. For example, pubs would like to have extra batches of goods when the orders they place are significantly expensive.

Codifying the Initial Knowledge

In the scenario we take the perspective of a BDI agent (from now on *our* agent, or agent i) that represents a pub owner. This agent needs very often to place orders to refill the stock. We describe the following two cases: *(i)* how the agent evaluates others' behavior regarding both organizational and personal norms, and *(ii)* how the agent reasons using this knowledge to select appropriate actions.

Summarizing:

1. In the first case we show how the agent incorporates the information related to others' behavior regarding organizational and personal norms. That is, the dynamics of the architecture from the perceived facts to the evaluation using Repage.

2. The second case represents our agent's reasoning using the acquired knowledge to reach concrete intentions, and thus, the best *reasonable* action.

For the sake of clarity, predicates of the shape $Pred(X)$ where $X \in I\!R$ will be written as $Pred = X$. In the same way, predicates of the shape $Pred(X) \wedge X \odot Y$ where $\odot$ stands for whatever boolean binary function over the set of real numbers, will be written as $Pred \odot Y$. For instance, $order \leq 100$ stands for $order(X) \wedge X \leq 100$, and $agreedPaid < paid$ stands for $agreedPaid(X) \wedge paid(Y) \wedge X < Y$, where X and Y are variables of the logical language.

Organizational Norms: The following are the organizational norms that rule our scenario. These norms are known by all participants in the system at start-up. We assume they already exist in the Normative Context (NC) in our agent's BDI architecture:

- **ON** - *Orders must not be delivered later than 7 days after the date they were placed.*

This norm was introduced above, in section 5.4.3. Assuming the agent wants to capture different grades in the possible violation of the norm, the set of evaluative patterns for the role *Provider(ON)* is specified as:

$$N(1, Provider(ON), dTime \leq 7)$$
$$N(2, Provider(ON), 7 < dTime \leq 9)$$
$$N(3, Provider(ON), 9 < dTime \leq 15)$$
$$N(4, Provider(ON), 15 < dTime)$$

Personal Norms: Our agent holds the following personal norm:

- **PN** - *Providers should give away some extra chips units when the order exceeds 100 units.*

Notice that this norm is only applicable when certain condition (order exceeds 100 units) occurs. Since we have an underlying logical language this is easy to model by adding logical conjunctions:

$$N(1, Provider(PN), order \geq 100 \wedge extraChips = 0)$$
$$N(2, Provider(PN), order \geq 100 \wedge 0 < extraChips \leq 15)$$
$$N(3, Provider(PN), order \geq 100 \wedge 15 < extraChips \leq 30)$$
$$N(4, Provider(PN), order \geq 100 \wedge 30 < extraChips)$$

From now on we will write xCh for $extraChips$, dT for $dTime$ and odr for $order$.

Predefined knowledge: Firstly, our agent is interested in buying chips batches. She previously knows three different snack providers, namely: *Mr. Potato*, *Fringles* and *Truffles*. Our agent needs to select one of these providers to place an order. As a first approach the agent asked those providers for their offers to sell their goods. So we consider our agent already knows the price per batch. Thus she has already introduced this information as beliefs in BC for all possible X:

$$B(buy(Mr.Potato, X), odr = X \wedge agreedToPay(10 \cdot X), 1, e)$$
$$B(buy(Fringles, X), odr = X \wedge agreedToPay(7 \cdot X), 1, e)$$
$$B(buy(Truffles, X), odr = X \wedge agreedToPay(11 \cdot X), 1, e)$$

In this example, the action *buy* can be instantiated with the number of chips batches (X). Then, when an agent performs the action $buy(Mr.Potato, 230)$ the predicates $odr = 230$ and $agreedToPay(2300)$ hold in the belief context to support the action.

5.4.7 Norms Evaluation: Example

Following the example introduced in the previous section, and considering the organizational norm ON2 and the personal norm PN, we show as a matter of example, how such norms are evaluated. We show how our agent, playing the role Pub, evaluates the behavior of three different providers regarding such two norms that regulate the role Provider (we assume that the action associated with the role Provider is $buy(j, X)$, where j is the agent playing the role Provider and X the quantity of chips unit). The process involves the rules F, C. Aforesaid, after an action is performed, a *fulfillment* predicate is introduced into the system by means of bridge rule F. Once the fulfillment is introduced into the Norm Context, rule C can be fired. We illustrate this with an example: let $buy(Mr.Potato, 230)$ be the action that our agent has just performed. Due to this, the following predicate is generated:

$$F(Mr.Potato, provider, order = 230 \ \wedge \ xCh = 35 \ \wedge \ dTime = 8, t)$$

For norm ON, rule C is instantiated as (we only put the relevant information of the fulfillment)

$$\frac{\begin{array}{c} NC : N(2, provider(ON), 7 < dTime \leq 9) \\ NC : F(Mr.Potato, provider(ON), dTime = 8, t) \\ BC : B(\iota, imply(dTime = 8, 7 < dTime \leq 9), 1, e_{\emptyset}) \end{array}}{RC : Outcome(Mr.Potato, provider(ON), w_{2_{ON}}, t)}$$

For personal norm PN, rule C is fired as

$$\frac{\begin{array}{c} NC: N(4, provider(PN), order \geq 100 \ \wedge\ 30 < xCh) \\ NC: F(Mr.Potato, provider(PN), xCh = 35, t) \\ BC: B(\iota, (order = 230 \ \wedge\ xCh = 35) \rightarrow (order \geq 100 \ \wedge\ 30 < xCh), 1, e_{\emptyset}) \end{array}}{RC: Outcome(Mr.Potato, provider(PN), w_{4_{PN}}, t)}$$

In the example, $w_{2_{ON}}$ and $w_{4_{PN}}$ correspond to the linguistic labels used to evaluate the outcomes of the transaction for each norm.

Repage context receives the outcome predicates and aggregates them, updating the final image predicates. These predicates provide a probabilistic distribution over the possible outcomes. For instance, the predicate $Img(Mr.Potato,$ $provider(ON)$, $[0.5, 0.4, 0.1, 0.0]$) indicates that when agent i interacts with $Mr.$ $Potato$, who plays the role of $provider$, $dTime < 7$ occurs with a probability of 0.5, while $7 \leq dTime < 9$, $9 \leq dTime < 15$ and $15 \leq dTime$ occurs with a probability of 0.4, 0.1 and 0.0, respectively. Through rules A_I and A_R image and reputation predicates are introduced into the belief context. Next subsection illustrate the whole reasoning process.

5.4.8 Reasoning Using Norms Evaluation: Example

In this part we show the agent's reasoning process, in which knowledge acquired through the evaluation of others w.r.t. different norms is used.

Let us assume that our agent wants to order 230 units of chips batches. She needs them in a week, but she may consider few days of delay. What she does not want at all is a provider who makes her paying more than the quantity they agreed. Furthermore, as the order exceeds 100 units, she would like to receive some extra units. One possible theory in the Desire context could be:

$$\begin{array}{l} (D^{+} order = 230 \ \wedge\ dTime \leq 7 \ \wedge\ xCh > 30, 1) \\ (D^{+} order = 230 \ \wedge\ 7 < dTime \leq 9 \ \wedge\ xCh > 30, .8) \\ (D^{-} order = 230 \ \wedge\ 15 \leq dTime, 1) \end{array}$$

On the other side, the Belief Context has the means by which the agent can achieve the desires. In particular, notice that both organizational norms (ON) and personal norms (PN), are somehow present in the desires. The performance of the agents regarding these norms is computed by the Repage Context which provides the information in terms of Image and Reputation predicates. In the example we only consider the Former. To show the reasoning process we assume that the agent has already been interacting with the providers, generating the following Images:

$$\begin{array}{c} Img(Mr.Potato, provider(ON), [.8, .1, .1, 0]) \\ Img(Fringles, provider(ON), [.6, .2, .2, 0]) \\ Img(Truffles, provider(ON), [.2, .5, .2, .1]) \end{array}$$

$$\begin{array}{c} Img(Mr.Potato, provider(PN), [0, .1, .2, .7]) \\ Img(Fringles, provider(PN), [.3, .5, .1, .1]) \\ Img(Truffles, provider(PN), [.5, .4, .1, 0]) \end{array}$$

These predicates instantiate rule A_I generating beliefs. For instance, regarding $MrPotato$:

$$\begin{array}{c} NC : N(1, provider(ON), dTime \leq 7) \\ NC : N(2, provider(ON), 7 < dTime \leq 9) \\ NC : N(3, provider(ON), 9 < dTime \leq 15) \\ NC : N(4, provider(ON), 15 < dTime) \\ RC : Img(Mr.Potato, provider(ON), [0.8, 0.1, 0.1, 0]) \\ \hline BC : B(buy(MrPotato, X), dTime \leq 7, 0.8, \{provider(ON)\}) \\ BC : B(buy(MrPotato, X), < dTime \leq 9, 0.1, \{provider(ON)\}) \\ BC : B(buy(MrPotato, X) < dTime \leq 15, 0.1, \{provider(ON)\}) \\ BC : B(buy(MrPotato, X)15 < dTime, 0, \{provider(ON)\}) \end{array}$$

$$\begin{array}{c} NC : N(1, provider(PN), order \geq 100 \ \wedge \ xCh = 0) \\ NC : N(2, provider(PN), order \geq 100 \ \wedge \ 0 < xCh \leq 15) \\ NC : N(3, provider(PN), order \geq 100 \ \wedge \ 15 < xCh \leq 30) \\ NC : N(4, provider(PN), order \geq 100 \ \wedge \ 30 < xCh) \\ RC : Img(Mr.Potato, provider(PN), [0, .1, .2, .7]) \\ \hline BC : B(buy(MrPotato, X), order \geq 100 \ \wedge \ xCh = 0, 0, \{provider(PN)\}) \\ BC : B(buy(MrPotato, X), order \geq 100 \ \wedge \ 0 < xCh \leq 15, .1, \{provider(PN)\}) \\ BC : B(buy(MrPotato, X), order \geq 100 \ \wedge \ 15 < xCh \leq 30, .2, \{provider(PN)\}) \\ BC : B(buy(MrPotato, X), order \geq 100 \ \wedge \ 30 < xCh, .7, \{provider(PN)\}) \end{array}$$

Also, assuming independence between ON and PN, by simple logical deduction, these predicates can be combined by multiplying their probabilities (see chapter 4). For instance:

$$\begin{array}{c} B(buy(MrPotato, X), dTime \leq 7, 0.8, \{provider(ON)\}) \\ B(buy(MrPotato, X), order \geq 100 \ \wedge \ 30 < xCh, .7, \{provider(PN)\}) \\ \hline B(buy(MrPotato, X)dTime \leq 7 \ \wedge \ order \geq 100 \wedge \\ \wedge \, 30 < xCh, .56, \{provider(ON), provider(PN)\}) \end{array}$$

Bridges rule 1 and 2 are executed for each generic positive and negative desires, respectively. For the positive desire in our example:

$$\frac{\begin{array}{c} DC : (D^{+} order = 230 \ \wedge \ dTime \leq 7 \ \wedge \ xCh > 30, 1) \\ BC : B(buy(MrPotato, 230), dTime \leq 7 \ \wedge \ order \geq 100 \wedge \\ 30 < xChips \ \wedge \ order = 230, .56, \{provider(ON), provider(PN)\}) \\ BC : B(\iota, order = 230 \rightarrow order \geq 100, 1, e_{\emptyset}) \end{array}}{\begin{array}{c} DC : (D^{+}_{buy(MrPotato,230)} order = 230 \wedge \\ order \leq 7 \ \wedge \ xCh > 30, g(1, .56)) \end{array}}$$

where $g(x, y)$ represents the grade of the positive desire. We will consider $g(x, y) = x \cdot y$, resulting a value of .56. This value indicates an expected level of satisfaction for our agent if she places the order to $MrPotato$, regarding the positive desire. Positive desires indicate the grade of satisfaction for our agent w.r.t. a concrete action (in this case $buy(MrPotato, X)$). Analogously, if we apply bridge rule 1 and 2 for the remaining desires we obtain the following new concrete desires:

$$DC : (D^{+}_{buy(MrPotato,230)} order = 230 \ \wedge \ 7 < dTime \leq 9 \ \wedge xCh > 30, .07)$$
$$DC : (D^{-}_{buy(MrPotato,230)} order = 230 \ \wedge \ 15 \leq dTime, 0)$$

Thus, using positive desires and taking into account negative desires, bridge rule 3 generates intentions:

$$\frac{\begin{array}{c} DC : (D^{+}_{buy(MrPotato,230)} order = 230 \wedge \\ dTime \leq 7 \ \wedge \ xCh > 30, .56) \\ DC : (D^{-}_{buy(MrPotato,230)} order = 230 \ \wedge \ 15 \leq dTime, 0) \\ PC : action(buy(MrPotato, 230), \top) \end{array}}{\begin{array}{c} IC : (I_{buy(MrPotato,230)}(order = 230 \ \wedge \ dTime \leq 7 \wedge \\ \wedge \, xCh > 30), f(.56, .0)) \end{array}}$$

In this case, the expected level of satisfaction of achieving the desire by buying from $MrPotato$ is .56 and there are not counter-effects entailing an expected level of disgust, since no negative desires affect this action. Taking $f(\delta_{+}, \delta_{-}) = max(0, \delta_{+} - \delta_{-})$, the generated intention would have a grade of .056. Applying the same to the other positive desire we obtain:
$IC : (I_{buy(MrPotato,230)}(order = 230 \ \wedge \ 7 < dTime \leq 9 \ \wedge \ xCh > 30), f(.07, 0))$

If we apply the same process for the rest of providers we would obtain the following intentions (we have omitted the intermediate process):

$$(I_{buy(Fringles,230)}(order = 230 \ \wedge \ dTime \leq 7 \ \wedge \ xCh > 30), .06)$$
$$(I_{buy(Fringles,230)}(order = 230 \ \wedge \ 7 < dTime \leq 9 \ \wedge \ xCh > 30), .02)$$
$$(I_{buy(Truffles,230)}(order = 230 \ \wedge \ dTime \leq 7 \ \wedge \ xCh > 30), 0)$$
$$(I_{buy(Truffles,230)}(order = 230 \ \wedge \ 7 < dTime \leq 9 \ \wedge \ xCh > 30), 0)$$

After calculating all possible intentions, bridge rule 4 would generate the action $CC : does(buy(MrPotato, 230))$, since the intention for the action

$$buy(MrPotato, 230)$$

has the maximum grade of satisfaction. Consequently, our agent will select $MrPotato$ as a seller to place her order to.

5.5 Conclusions

The chapter introduces the BDI+Repage model, one of the main contributions of this work. We define an agent architecture, a *belief-desire-intention* (BDI) architecture, that integrates image and reputation information calculated from Repage [Sabater-Mir et al., 2006] into the practical reasoning process of the agent. Even when in the introduction of this book we already state the main features of the model, we would like to enhance the following ones:

- It is defined as a multi-context system (MCS) [Giunchiglia and Serafini, 1994]. From a software engineering perspective it supports modular architectures and encapsulation. From a logical modeling perspective, it allows the construction of agents with different and well-defined logics, keeping all formulas of the same logic in their corresponding context. This increases considerably the representation power of logical agents, and at the same time, simplifies their conceptualization.
- It is based on solid logical frameworks. We use an existing complete logic of preferences based on Lukasiewicz [Casali, 2008] to model the desires and intentions, and we use the logic defined in chapter 4 to model the beliefs.
- It handles *image* and *reputation*. The Repage model is based on a cognitive theory of reputation that states a main difference between image and reputation.
- It is generic. The model is not attached to any specific domain ontology nor network typology, and inherits the properties and characteristics of the underling reputation model. We use Repage as a paradigmatic example, but any model whose information can be captured by the reputation language L_{rep} could be placed into the system.

Moreover, we introduce the BDI+Repage+Norm model, an extension of the BDI+Repage model that deals with norms. For developing such extension, we introduce a new context (normative context) which includes a first-order language to describe how norms are evaluated. This new context is endowed with a set of bridge rules that feed the other context of the model.

The extension demonstrates the flexibility of the BDI+Repage model and illustrate how other possible extensions could be done.

Chapter 6

Arguing about Social Evaluations

6.1 Introduction and Motivation

Reputation and trust models provide social evaluations of the potential performance of agents in the society regarding a specific context, using the history of interactions and its results, and third-party communications as a main source to compute them. Also, some of them include a reliability measure attached to the social evaluations, indicating how *confident* the agent is about the evaluation (see [Sabater and Sierra, 2005] for a review).

The latter though carries out a big problem when such evaluations are communicated. Due to the subjectivity of reputation information, a social evaluation totally reliable by an agent A may not be reliable for B, because the bases under which A has inferred the social evaluation cannot be accepted by B. This can happen because agents have different inference rules, have had different experiences, have different goals, etc. Usually, reputation models that manage reliability measures consider a threshold below which communicated social evaluation are not taken into account. Yet, since the source agent calculates the reliability measure and it is a subjective matter, the acceptance/rejection of communicated social evaluations according to this criteria may produce noise for the recipient agent.

This paper offers a possible solution that can complement already existing methods. We suggest that, in communicated social evaluations, the reliability measure cannot be dependent on the source agent, but must be fully evaluated by the recipient agent according to its own knowledge. In our approach, rather than allow only single communications, we allow agents to participate in argumentation-based dialogs regarding reputation elements in order to decide on the reliability (and thus acceptance) of a communicated social evaluation. Our approach differs from others in that it is the recipient agent, not the source agent, who decides about the reliability of a communicated evaluation.

We develop an argumentation-based dialog protocol for the exchange of reputation-related information. Due to the subjectivity of reputation information, a social evaluation totally reliable by an agent A may not be reliable for B, because the bases under which A has inferred the social evaluation cannot be accepted by B. This can happen because agents have different inference rules, have had different experiences, have different goals, etc. When such information is communicated this can become very problematic, specially if the reputation model assigns a reliability measure to the communicated information, because of the reasons above.

The main characteristics of the system are:

- Only the recipient agent decides about the reliability of a communicated evaluation. This differs from other approaches in which the source agent attaches a reliability measure to the communicated social evaluation. This makes more difficult for dishonest agents to intentionally send fraudulent information, because they must be aware of the knowledge of the recipient and justify the *lie* accordingly.

- It uses argumentation frameworks to give semantics to the dialog. We exploit the L_{rep} language (a many-sorted first-order language to express reputation related concepts) to completely define how arguments are constructed and how arguments influence one another. We instantiate a weighted abstract argument framework to define the acceptability semantics of a communicated social evaluation.

- It handles quantitative and qualitative graded information. One of the main characteristics of reputation information is that it is graded. Nowadays it is strange to find a model that provides crisp evaluations of the agents. For instance, an agent A may be *bad*, *very bad* or *very good* etc. as a car driver, and this has to be taken into account when arguing about evaluations.

- It permits dialogs between parties that use different reputation models. Even when we assume that agents use the same language to talk and reason about reputation information (L_{rep} language), we suppose that they can use different inference rules (different reputation models) without having to exchange the exact rules that each agent uses for the inferences.

6.2 Communicated Social Evaluations and their Reliability

6.2.1 Preliminaries

The problematic regarding communicated social evaluations that the subjective notion of reputation brings, it is the same as for any rhetorical construct that depends on internal elements that are private. Let us consider a very simple example:

i :	*How is John as a car driver?*
j :	*He is a very good driver*
i :	*Why?*
j :	*Well, Emma told me that, and she is a good informer*
i :	*Oh! for me, Emma is very bad as informer!*

In the previous example, should agent i consider the information sent by j saying that $John$ is a *very good* driver? Notice that j is justifying her opinion with a previous communication from $Emma$, which she thinks is a good informer. But it contradicts an information that i considers valid. For i, the information is not reliable, even when j may be totally honest.

When talking about social evaluations and reputation models, usually the model already handles possible inconsistent knowledge. Different opinions referring to the same target agent may be totally contradictory, and the agent integrates and aggregates the information in order to achieve a consistent mental state. Determining whether a piece of information is acceptable in a possibly inconsistent knowledge base has been faced in argumentation theory. In this field, each piece of information is justified by the elementary elements from which it has been inferred, the so called arguments. Then, two arguments can *attack* each other, indicating that the information supporting them would be inconsistent if they are both accepted at the same time.

6.2.2 Characterizing the Problems behind Reliability Measures

We start this subsection by recalling the notion of reputation theory explained in chapter 3 and how we characterized the reputation information that agents hold.

Definition (*Reputation Theory*) Let $\Delta \subset wff(L_{rep})$, we say that Δ is a reputation theory when $\forall \alpha \in \Delta$, α is a ground element. Then, letting $d \in wff(L_{rep})$, we write $\Delta \vdash d$ to indicate that from the reputation theory Δ, it can be deduced d via $\vdash$. Ground elements are *communications* and *direct experiences* (*comm* and DE respectively), while non-ground elements are *images*, *reputations*, *shared voices* and *shared evaluations*.

Having introduced the reputation language, we can illustrate more precisely the kind of problems we deal with in this chapter, and the characteristics of the proposed system. We start with a very simple example. Let i, j be two agents with their respective reputation theories and reputation models $\langle \Delta_i, \vdash_i \rangle$ and $\langle \Delta_j, \vdash_j \rangle$. Let us consider that agent i has a VG image of $John$ as a *car_seller* (with a maximum reliability), so $\Delta_i \vdash_i Img_i(John, car\_seller, VG)$. When i communicates such information to j at time t, j updates its reputation theory with a new communication:

$$\Delta_j \cup \{Comm_j(i, \lceil Img(John, car\_seller, VG)\rceil, t)\}$$

Let us assume that i inferred the image of $John$ as a car_seller from (1) a communication from $Alice$ and (2) the very good reputation (according to i) of $Alice$ as $informer$:

$$(1) Comm_i(Alice, \lceil Img_{Alice}(John, car\_seller, VG)\rceil, t)$$
$$(2) Rep_i(Alice, informer, VG)$$

Also, assume that j has a very different opinion about the reputation of $Alice$ as $infomer$, a very bad reputation indeed. Specifically, $\Delta_j \vdash_j Rep_j$ ($Alice$,$informer$,VB). With this scenario, at least one question arises. Should j update its reputation theory with the original communication from i?

We argue that the communicated information from i is not reliable for j in this example. Without the analysis of the internal elements, such a situation is impossible to detect, and the effects of including i's communication in j's reputation theory can be devastating for j. Agents use social evaluations to decide what to do. It may happen that i's communication helps j choose $John$ as a car_seller when j wants to buy a car. If the direct interaction with $John$ does not go well, several things may occur:

1. Direct experiences are costly. Probably j has bought a car before noticing that it was not good.

2. j may generate a bad image of i as informer, which can lead to j not considering anymore future communications from i, even when i, according to j's knowledge, was honest.

3. Also j may spread bad reputation of i as informer, and thus collide with the opinion of other members of the society that are aligned with i. Consequently such members may retaliate against j [Conte and Paolucci, 2002].

All the previous situations can be avoided if j has the capability to decide whether the piece of information is reliable enough, not based on the reliability measure that i assigns, but on the internal elements that i uses to justify the communicated social evaluation and that j can check. Furthermore, our approach makes more difficult to intentionally lie, since a potential liar should know beforehand what the recipient knows, and build the argument accordingly to it. In current approaches, a liar agent can put a very high reliability value in the communicated social evaluation to introduce noise in the recipient agent.

To allow agents to analyze the justifications, we propose a protocol that performs a dialectical process between the agents. Intuitively, both the source and the recipient agents, following a well-defined protocol, can exchange at each turn a justified social evaluation (argument) that counterargues (attacks) some of the arguments uttered by the other agent. At the end of the process, the recipient agent holds a tree of arguments that can be used to decide whether the original communication from the source agent is reliable, and update its reputation theory accordingly. The technical details to design such protocol and the posterior analysis are taken from the field of computation argumentation, which has proposed frameworks and methods to deal with similar situations. We have taken some of these concepts and tools and adapted them to confront the peculiarities that reputation information and our scenarios have. We highlight just two:

The attacks are graded: In the previous example, j holds a very different opinion of the reputation of *Alice* as $informer$ than i has, *very bad* (VB) against *very good* (VG) respectively. However, this would note the case if j thinks that the reputation of *alice* as $informer$ is *good* (G), so $\Delta_j \vdash_j Rep(Alice, informer, G)$. The *attack* should be considered weaker in the latter case. Our framework handles graded attacks by assuming that each agent has a distance function $\ominus : G \times G \to \mathbb{Q}$ over the totally order set $M = \langle G, \leq \rangle$ which is used to represent the values of the social evaluations (see section 3.3.1).

Heterogeneity of the agents: Even when agents use the same language to talk and reason about reputation, they may use different reputation models. Usually, an argument is defined as a pair composed of a conclusion and a set of elements that have been used to infer such conclusion (supporting set). The conclusion is the element that is being justified by the supporting set. If agents use different inference rules, the supporting set must include enough information to reconstruct the reasoning path followed by the agent that has built the argument. This could also be *easily* done by sending the exact inference rules of the reputation model in the arguments, but it would violate the privacy of the agents and therefore is not an option. Instead, our framework provides an intermediate solution. We define a very simple inference consequence relation $\vdash_{arg}$ that all agents must know, and specify a transformation that agent should use to build arguments using $\vdash_{arg}$. From $\langle \Delta_i, \vdash_i \rangle$ and $\langle \Delta_j, \vdash_j \rangle$, we move to $\langle \Gamma_i, \vdash_{arg} \rangle$ and $\langle \Gamma_j, \vdash_{arg} \rangle$, where Γ_i and Γ_j are argumentative theories built from their respective reputation theories and reputation models. Argumentative theories contain all the elements from their respective reputation theories, and simple implication rules that simulate inference steps performed by their respective reputation model, without indicating how they were performed internally.

The protocol allows agents to construct trees of arguments with their respective attacks. We provide then an acceptability semantics, a mechanism for deciding whether the information from the source agent can be considered reliable enough for the recipient. We can do that because the argumentation framework we instantiate [Dunne et al., 2009] introduces the concept of *inconsistency budgets*. Intuitively, inconsistency budgets indicate the *amount* of inconsistency that an agent can (wants to) tolerate. For instance, in the previous example where $\Delta_j \vdash_j Rep(Alice, informer, G)$, agent j may consider that the difference between G and VG is small enough to accept that they are not contradictory, even when that might not be the case for another agent. Agents *autonomously* decide the strength of a given attack according to their own distance function and therefore to which extent they can accept inconsistencies.

The next section formally describes: (1) how agents build arguments; (2) how agents construct an argumentative theory from a reputation theory; (3) how such arguments influence each other and with which strength; and (4) how the recipient agent can decide whether a piece of communicated information is reliable or not.

6.3 The Reputation Argumentation Framework

Our approach suggests that agents use argumentation techniques to decide whether a piece of information can be considered reliable or not. For this, we need to define an argumentation framework for reputation-related concepts. First, we specify the notion of argument, the construct of arguments, and how they influence each other. Second, we define L_{arg}, a language based on L_{Rep} to write argument sentences, and the consequence relation $\vdash_{arg}$ associated with the language and used to build arguments. We also give an acceptability semantics, indicating under which conditions, an agent would *accept* a given communicated social evaluation as reliable.

Definition (*Argument*) A formula $(\Phi{:}\alpha) \in wff(L_{arg})$ when $\alpha \in wff(L_{Rep})$ and $\Phi \subseteq wff(L_{Rep})$. Intuitively, we say that the set Φ is the supporting set of the argument, and α its conclusion. It indicates that α has been deduced from the elements in Φ.

The validity of a given well-formed argument must be contextualized in an argumentation theory, a set of elementary argumentative formulas, called *basic declarative units* (bdu). We adapt the following definition from [Chesevar and Simari, 2007]:

Definition (*Argumentative Theory*) A *basic declarative unit* (bdu) is a formula $(\{\alpha\}{:}\alpha) \in wff(L_{arg})$. Then, a finite set $\Gamma = \{\gamma_1, \ldots, \gamma_n\}$ is an argumentative theory iff each γ_i is a bdu.

From an argumentative theory Γ, we can now define how arguments are constructed. For this we use the inference relation $\vdash_{arg}$, characterized by the deduction rules *Intro-BDU*, *Intro-AND* and *Elim-IMP* (figure 6.1). Rule *Intro-BDU* allows the introduction of a basic declarative unit from the argumentative theory. Rule *Intro-AND* permits the introduction of conjunctions. Finally, rule *Elim-IMP* performs the traditional *modus ponents.*

Definition (*Valid Argument and Subargument*) Let $(\Phi{:}\alpha) \in wff(L_{arg})$ and let Γ be an argumentative theory. We say that $(\Phi{:}\alpha)$ is a valid argument in the bases of Γ iff $\Gamma \vdash_{arg} (\Phi{:}\alpha)$. Also, we say that a valid argument $(\Phi_2{:}\alpha_2)$ is a subargument of $(\Phi{:}\alpha)$ iff $\Phi_2 \subset \Phi$.

As mentioned earlier, each agent i has to construct its argumentative theory Γ_i in order to build arguments. This argumentative theory is based on the reputation information that i has, characterized with the tuple $\langle \Delta_i, \vdash_i \rangle$. Assuming that $\vdash_i$ is defined by a finite set of natural deduction rules $\{\vdash_{i_1}, \ldots, \vdash_{i_m}\}$,

$$\text{Intro-BDU:} \quad \frac{}{(\{\alpha\}{:}\alpha)} \qquad \text{Intro-AND:} \quad \frac{(\Phi_1{:}\alpha_1),\ \ldots\ (\Phi_n{:}\alpha_n)}{(\bigcup_{i=1}^{n} \Phi_i{:}\alpha_1, \ldots, \alpha_n)}$$

$$\text{Elim-IMP:} \quad \frac{(\Phi_1{:}\alpha_1, \ldots, \alpha_n \to \beta) \qquad (\Phi_2{:}\alpha_1, \ldots, \alpha_n)}{(\Phi_1 \cup \Phi_2 : \beta)}$$

Figure 6.1: Deductive rules for the consequence relation $\vdash_{arg}$.

- For all $\alpha \in \Delta_i$ then $(\{\alpha\}{:}\alpha) \in \Gamma_i$. That is, all ground elements from the reputation theory are bdu in the argumentative theory.
- For all $\alpha_1, \ldots, \alpha_n$ s.t. $\Delta_i \vdash_i \alpha_k$ where $1 \leq k \leq n$, if there exists m s.t. $\alpha_1, \ldots, \alpha_n \vdash_{i_m} \beta$, then $(\{\alpha_1, \ldots, \alpha_n \to \beta\}{:}\alpha_1, \ldots, \alpha_n \to \beta) \in \Gamma_i$. This construct introduces every instantiated deductive step as a rule in the form of a basic declarative unit. For instance, if $\alpha, \beta \vdash_{i_2} \gamma$, the argumentative theory will include the bdu formula $(\{\alpha, \beta \to \gamma\} : \alpha, \beta \to \gamma)$.

The following proposition is easy to prove.

Proposition Let $\langle \Delta_i, \vdash_i \rangle$ be the reputation information associated with agent i, and Γ_i its argumentative theory. If $\Delta_i \vdash_i \alpha$, then there exists an argument $(\Phi : \alpha)$ such that $\Gamma_i \vdash_{arg} (\Phi : \alpha)$.

6.3.1 Argument Interactions

We have explain how agents construct their argumentative theory from their reputation information, and how from such theory they can build arguments using $\vdash_{arg}$. In this subsection we detail how arguments generated from different agents influence one other (attack). Differently from argumentation systems used as theoretical reasoning processes to analyze the possible inconsistencies that a single agent may hold, our framework is designed to be part of a dialectical process, where attacks are produced only from arguments sent by other agents.

To specify the *attack* relationship among arguments, we define first the binary relation $\cong$ between L_{rep} predicates. Let α, β be well-formed non-ground formulas from L_{Rep}. Then, $\alpha \cong \beta$ iff $\mathit{type}(\alpha) = \mathit{type}(\beta)$, $\alpha.\mathit{target} = \beta.\mathit{target}$, $\alpha.\mathit{context} = \beta.\mathit{context}$ and $\alpha.\mathit{value} \neq \beta.\mathit{value}$. We can see that $\cong$ is symmetric but not reflexive nor transitive. For instance, $\mathit{Rep}(i, \mathit{seller}, VB) \cong \mathit{Rep}(i, \mathit{seller}, G)$, but $\mathit{Rep}(i, \mathit{seller}, VB) \ncong \mathit{Img}(i, \mathit{seller}, G)$ and $\mathit{Rep}(i, \mathit{seller}, VB) \ncong \mathit{Rep}(i, \mathit{buyer}, VG)$.

Definition (*Attack between Arguments*) Let $(\Phi_1{:}\alpha_1)$, $(\Phi_2{:}\alpha_2)$ be valid arguments in the bases of Γ. We say that $(\Phi_1{:}\alpha_1)$ *attacks* $(\Phi_2{:}\alpha_2)$ iff $\exists(\Phi_3{:}\alpha_3)$ subargument of $(\Phi_2{:}\alpha_2)$ s.t. $(\alpha_1 \cong \alpha_3)$.

We want also to quantify the strength of the attack. Let $a = (\Phi_1{:}\alpha_1)$ be an argument that attacks $b = (\Phi_2{:}\alpha_2)$. Then, by definition, a $(\Phi_3{:}\alpha_3)$ subargument of (Φ_2:α_2) s.t. ($\alpha_1 \cong \alpha_3$) exists. The strength of the attack is calculated through the function w as $w(a,b) = \alpha_1.value \ominus \alpha_3.value$, where $\ominus$ is a binary function defined over the domain of the representation values used to quantify the evaluations (the total ordered set $M = \langle G, \leq \rangle$). For instance, if $M = \langle [0,1] \cap \mathbb{Q}, \leq \rangle$, we can define $\ominus(x,y) = |x-y|$. In this case, 1 is the strongest attack. If $M = \langle \{VB, B, N, G, VG\}, \leq_s \rangle$, we could first assign each label a number: $f(VB) = 0$, $f(B) = 1$, $\ldots$, and then, $\ominus(x,y) = |f(x) - f(y)|$. In this case, the strongest attack is quantified with 4. $\ominus$ implements a *difference* function among the possible values.

The previous attack definition does not consider attacks between direct experiences nor communications. This means that discrepancies at this level cannot be argued, even when they are completely contradictory. Yet, this is justified by the fact that, in our framework, ground elements are not generated from any other piece of information. Thus, a communicated ground element should be introduced directly into the reputation theory. Obviously, the language could be extended to capture the elementary elements that compose direct experiences (contracts, fulfillments etc.). Again though, we think that sharing this low level information would violate the privacy of the agents.

6.3.2 Deciding about the Reliability

At this point, agents can build arguments, determine when their arguments attack arguments from other agents (and vise versa), and assign a strength to these attacks. However, we are still missing how agents can decide when to accept a given argument, considering that they will have a weighted tree of arguments where each node is an argument and each edge represents the strength of the attack. For this, we instantiate a weighted version of the classic Dung abstract argumentation framework [Dung, 1995], and use an acceptability semantics defined for this framework.

Dung's framework is defined as follows:

Definition (*Abstract Argumentation Framework*) An abstract argument system (or argumentation framework) is a tuple $AF = \langle A, R \rangle$ where A is a set of arguments and $R \subseteq A \times A$ is an attack relation. Given $a, b \in A$, if $(a,b) \in R$ (or aRb), we say that a attacks b. Let $S \subseteq A$, and $a, b \in A$ then

- S is *conflict-free* iff $\nexists a, b \in S$ s.t. aRb.
- An argument b is *acceptable* w.r.t. the set S iff $\forall a \in A$, if aRb then $\exists c \in S$ s.t. cRa.
- S is *admissible* if it is conflict-free, and each argument in S is acceptable w.r.t. the set S. Also, S is a preferred extension iff it is maximal w.r.t. the set inclusion.
- An argument b is *credulously accepted* iff it belongs to at least one preferred extension.

This abstract framework does not consider strength in the attacks. Recent work from Dunne *et al.* [Dunne et al., 2009] extends Dung's framework with weights.

Definition (*Weighted Abstract Argumentation Framework*) A weighted argument system is a triple $AF_w = \langle A, R, w\rangle$ where $\langle A, R\rangle$ corresponds to a Dung's argumentation framework, and $w : R \to I\!R_>$ is a function that assigns weights to each attack relation[1].

The semantics of w gives a pre-order between possible inconsistencies. Let $a_1, b_1, a_2, b_2 \in A$ where $a_1 R b_1$ and $a_2 R b_2$, if $w((a_1, b_1)) < w((a_2, b_2))$ means that accepting both a_1 and b_1 is *more consistent* than accepting both a_2 and b_2. This leads to the definition of inconsistency budgets and $\beta-$solutions (β s.t. $\beta \in I\!R_\geq$). Intuitively, a β-solution is a solution of the unweighted Dung's framework in which the *amount* of inconsistency (calculated through the sum of the weights of the attacks) is lower or equal to β. Formally:

Definition (β*-solutions* [Dunne et al., 2009]) Given $AF_w = \langle A, R, w\rangle$, a solution $S \subseteq A$ is a $\beta-$solution if $\exists T \in sub(R, w, \beta)$ s.t. S is a solution of the unweighed system $AF = \langle A, R\backslash T\rangle$. Function *sub* returns a set of subsets of R in which the weights sum up to a maximum of β: $sub(R, w, \beta) = \{T | T \subseteq R$ and $(\sum_{r\in T} w(r)) \leq \beta\}$

We theorize that a credulous semantics for the acceptance of reliable information is enough. In the weighted version, we can define that, given $AF_w = \langle A, R, w\rangle$, an argument $b \in A$ is credulously accepted if it belongs to at least one β-preferred extension, so, if $\exists T \in sub(R, w, \beta)$ s.t. $b \in S$, and S is a preferred extension of the Dung's framework $AF = \langle A, R\backslash T\rangle$.

We can instantiate now the weighted argument system by using the constructs defined in this section. Let Γ be an argumentative theory as defined in this section. We define:

- $C(\Gamma) = \{(\Phi : \alpha) | \Gamma \vdash_{arg} (\Phi : \alpha)\}$, the set of all valid arguments that can be deduced from Γ.

- $R(\Gamma) = \{((\Phi_1 : \alpha_1), (\Phi_2 : \alpha_2)) | (\Phi_1 : \alpha_1)$ attacks $(\Phi_2 : \alpha_2)$ and $(\Phi_1 : \alpha_1) \in C(\Gamma)$ and $(\Phi_2 : \alpha_2) \in C(\Gamma)\}$, the set of all possible attack relations between the arguments in $C(\Gamma)$.

Then, we can describe the instantiation:

Definition (*Reputation Argument Framework*) The reputation argument system for the argumentative theory Γ is defined as $AF_\Gamma = \langle C(\Gamma), R(\Gamma), w\rangle$, where $w : R(\Gamma) \to I\!R$ is the strength function as defined above using the $\ominus$ difference function.

[1] Following the notation in [Dunne et al., 2009], we write $I\!R_>$ and $I\!R_\geq$ to refer to the set of real numbers greater than 0 and greater or equal to 0 respectively.

$\text{counter}^k_{PRO}(b)$	**Precondition** (1) k is even, $b \in C(\Gamma_{PRO} \cup X^{k-1}_{OPP})$ and b has not been issued yet (2) $\exists r \in I\!N$ s.t. $1 \leq r < \lvert S^{k-1} \rvert$, r is odd and $(b, S^{k-1}_r) \in R(\Gamma_{PRO} \cup X^{k-1}_{OPP})$ (3) $\nexists \gamma \in C(\Gamma_{PRO} \cup X^{k-1}_{OPP})$ s.t. $(\gamma, S^{k-1}_t) \in R(\Gamma_{PRO} \cup X^{k-1}_{OPP})$ where $r + 1 \leq t < \lvert S^{k-1} \rvert$ and t is odd **Postcondition** (i)$X^k_{PRO} = X^{k-1}_{PRO} \cup BDU(supp(b))$ (ii)$X^k_{OPP} = X^{k-1}_{OPP}$ (iii)$S^k = \langle S^{k-1}_0, \ldots, S^{k-1}_r, b \rangle$
$\text{counter}^k_{OPP}(b)$	**Precondition** (1)k is odd, $b \in C(\Gamma_{OPP} \cup X^{k-1}_{PRO})$ and b has not been issued yet (2)$\exists r \in I\!N$ s.t. $0 \leq r < \lvert S^{k-1} \rvert$, r is even and $(b, S^{k-1}_r) \in R(\Gamma_{OPP} \cup X^{k-1}_{PRO})$ (3)$\nexists \gamma \in C(\Gamma_{OPP} \cup X^{k-1}_{PRO})$ s.t. $(\gamma, S^{k-1}_t) \in R(\Gamma_{OPP} \cup X^{k-1}_{PRO})$ where $r + 1 \leq t < \lvert S^{k-1} \rvert$ and t is even **Postcondition** (i)$X^k_{PRO} = X^{k-1}_{PRO}$ (ii)$X^k_{OPP} = X^{k-1}_{OPP} \cup BDU(supp(b))$ (iii)$S^k = \langle S^{k-1}_0, \ldots, S^{k-1}_r, \beta \rangle$ (or $\langle S^{k-1}_0, b \rangle$ if $r = 0$)

Table 6.1: Possible movements of the dialog game at turn k. The function *supp*(b) returns the supporting set of b.

This finishes the definition of the reputation argument system. The idea is that each agent will be equipped with its own argumentation reputation system, and will add incrementally the arguments issued by the other agent. Intuitively, if the argument that justifies the original communicated social evaluation belongs to a preferred extension of the recipient agent, the latter will introduce the social evaluation into its reputation theory.

6.3.3 The Dialog Protocol

A dialog between two parties that can be seen as a game in which each agent has an objective and a set of legal movements (illocutions) to perform at each turn. Walton *et al.* in [Walton and Krabbe, 1995] state several types of dialogs depending on the participants' goals. In our case, we model a special kind of *information-seeking* dialog. The goal of the game then is to see whether the opponent (OPP) can *accept* reasonably the inquiring information from the proponent (PRO).

We use the argumentation framework defined in the previous sections to give semantics to the dialogs. The key is that each agent participating in the dialog will use its own argument framework to deal with possible inconsistencies. It is important to notice that agents do not have access to the set of arguments of the other agents. They incorporate such knowledge from the exchange of illocutions uttered in the dialog.

Let PRO and OPP be the proponent and the opponent agents engaged in the dialog respectively. Following a similar approach used in [Amgoud et al., 2000], both agents are equipped with a reputation argument system:

$$AF_{PRO} = \langle C(\Gamma_{PRO} \cup X_{OPP}), R(\Gamma_{PRO} \cup X_{OPP}), w_{PRO}\rangle$$
$$AF_{OPP} = \langle C(\Gamma_{OPP} \cup X_{PRO}), R(\Gamma_{OPP} \cup X_{PRO}), w_{OPP}\rangle$$

where $\Gamma_{PRO}, \Gamma_{OPP}$ are the argumentative theories of agents PRO and OPP, which are private. w_{PRO} and w_{OPP} are the weight functions of agents PRO and OPP. Finally, X_{PRO} (X_{OPP}) is the set of bdu from the arguments that results from the proponent's (opponent's) issued arguments. Both X_{PRO} and X_{OPP} are public and are the result of the exchange of arguments. This allows the agents to recognize and reconstruct arguments from the other agent. As for the state of our dialog protocol, we give a definition inspired in [Dunne and Bench-Capon, 2003]:

Definition (*State of the Dialog*) A state of a dialog at the k-th turn (where $k \geq 0$) is characterized by the tuple $\langle S^k, X^k_{PRO}, X^k_{OPP}\rangle^k$ where $S^k = \langle S^k_1, \ldots, S^k_t\rangle$ is the ordered set of arguments that represents a single dispute line. A dispute line is a finite sequence of arguments $a_1, \ldots, a_n$ where $\forall l$ s.t. $1 \leq l < n$, a_{l+1} attacks a_l. X^k_{PRO}, X^k_{OPP} are the public sets of BDU formulas of the proponent and the opponent respectively at turn k, incrementally built after each argument exchange and that are public.

The proponent is the initiator of the dialog and issues the argument $a = (\Phi{:}\alpha)$. The initial state at turn 0 is then characterized by $\langle\langle a\rangle, BDU(\Phi), \{\}\rangle^0$. The function $BDU(X)$ returns the set of elements from X as a BDU formula. So, if $\alpha \in X$, then $\{\alpha\}{:}\alpha \in BDU(X)$. The possible types of movements are summarized in figure 6.1, where we include preconditions and postconditions:

The proponent can perform the movement $counter^k_{PRO}(b)$ when the turn k is even (1). Of course, b should be a valid argument built from its argumentative theory and the bdu from the previous exchange of arguments ($C(\Gamma_{PRO} \cup X^{k-1}_{OPP})$) (1). We also require that b attacks some of the arguments of the current dispute line that the opponent has issued (so, in an odd position)(2). When this occurs, we also want to ensure that the proponent cannot attack any other argument issued by the opponent later than the one being attacked (3). Once the illocution is submitted, the effects in the dialog state are also described in figure 6.1. First, the set X_{PRO} is updated with the supporting set of the argument b (i). Notice that in the way we define the construction of arguments (see section 6.3) the supporting set only contains bdu. Thus, since this set is added to the argumentative theory of the opponent, it is able also to recognize the argument and attack it if necessary. Moreover, when an argument of the dispute line is attacked at point r of the dispute line, the dialog starts a new dispute line from that point(iii).

The opponent can submit counterarguments by sending $counter^k_{OPP}(b)$ with symmetric effects as explained in the previous paragraph. In this case, k must be odd. A dialog finishes when there are no possible movements.

The winner is the last participant who makes a move. Hence, if the number of moves is even, the winner is the proponent. If the number of moves is odd, the opponent wins. This protocol is a simplification of a TPI-Dispute

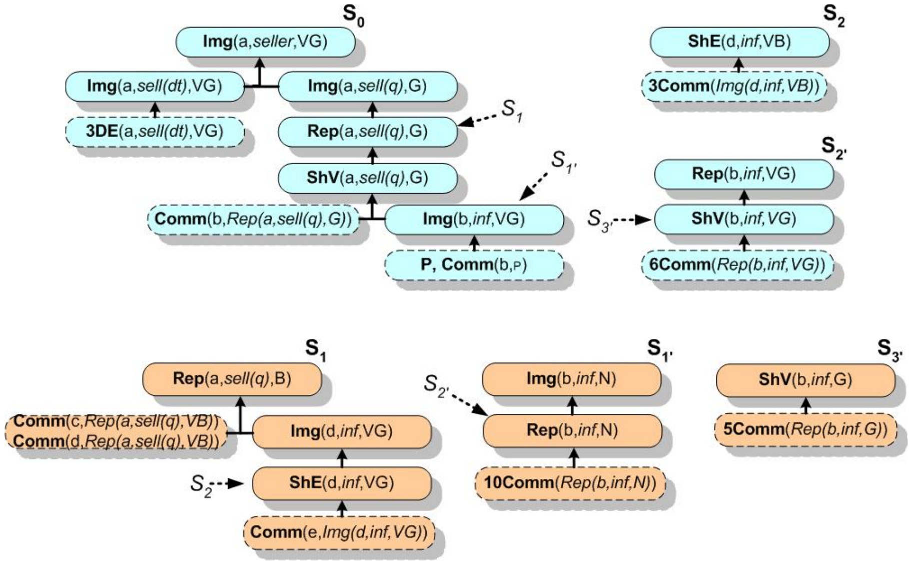

Figure 6.2: Arguments uttered by the proponent (S_0,S_2,$S_{2'}$) and the opponent (S_1,$S_{1'}$,$S_{3'}$) in the example respectively. Dot arrows indicate attack. For the sake of clarity, we omit the notation $\lceil\cdot\rceil$. $nComm(\cdot)$ and $nDE(\cdot)$ indicates that the agent holds n communications and n direct experiences respectively.

(*two-party immediate response dispute*) and instantiates a protocol described in [Amgoud et al., 2000]. From there, the following proposition can be deduced:

Proposition Let AF_{PRO} and AF_{OPP} be the argument frameworks of the participants of a dialog. When the game is finished and the proponent is the winner, the original argument $a = (\Phi{:}\alpha)$ belongs to a 0-preferred extensions of AF_{OPP}.

This means that the argument a is credulously accepted by the opponent. Therefore, the conclusion α can be introduced into the reputation theory of the opponent.

If OPP wins, OPP cannot find a 0-preferred extension that includes the argument a. In this case, OPP could choose not to update its reputation theory. However, depending on its tolerance to inconsistencies, OPP can find a 1-preferred extension that includes argument a, or even a 2-preferred extension. By increasing the inconsistency budget, the original argument may become acceptable, and thus the communicated social evaluation considered reliable. This might be seen equivalent to the threshold that some reputation models that manage reliability measures use to accept communicated information. The difference is that contrary to the measure being calculated by the source agent, in our approach, the reliability is computed by the recipient, who assigns strengths that can be different from the source. Algorithm 1 formalizes the procedure we have just described. In the next subsection we provide an example that shows the use of the protocol and the inconsistency budgets.

Algorithm 1: Reputation Theory Update Algorithm (for the agent j)

```
Data: Agent $i, j$
Data: Argument $\Phi$:$\alpha$ (sent by $i$)
Data: Reputation Information $\langle \Delta, \vdash_R \rangle$
Data: Inconsistency Budget $b$
Result: $\Delta_{res}$
(Reputation Theory Updated)
$\Gamma_j \leftarrow$ Argumentative Theory from $\langle \Delta, \vdash_R \rangle$;
$AF_j \leftarrow \langle C(\Gamma_j), R(\Gamma_j), w_j \rangle$ /*The argument framework of $j$*/;
$\langle winner, X_i \rangle \leftarrow$ dialogGame /*$AF_j$,$i$, $\langle \langle \alpha \rangle, \Phi, \{\} \rangle^0$*/;
if $winner = i$ then
    $\Delta_{res} \leftarrow \Delta \cup \{Comm(i, \alpha)\}$ /*$i$ wins, then $j$ updates its reputation
    theory*/;
else
    if $\Phi$:$\alpha$ is acceptable w.r.t. $\langle C(\Gamma_j \cup X_i), R_j, w_j \rangle$ and budget $b$ then
        $\Delta_{res} \leftarrow \Delta \cup \{Comm(i, \alpha)\}$ /*With inconsistency budget $b$, $j$
        accepts also the argument*/;
    else
        $\Delta_{res} \leftarrow \Delta$ /*Agent $j$ rejects the argument*/;
    end
end
```

6.3.4 An Example

We want to finish this section by showing a simple example. Here, agent i (the proponent) sends the first communication to j (the opponent). The arguments they build are shown in figure 6.2. In the domain, we have the context *seller*, composed of *sell(q)* (quality dimension of the sold products) and *sell(dt)* (delivery time of the product). We use the L_{rep} language taking M as $\langle \{VB, B, N, G, VG\}, \leq_s \rangle$ (see chapter 3). Also, the context *Inf* is used and stands for *informant.* For instance, the argument S_0 (figure 6.2) indicates that the agent has a VG (very good) image of a as a seller, because of the images it has about a taking into account the quality of the products (sell(q)) and the delivery time (sell(dt)) are G and VG respectively. The latter is justified because it had three direct experiences with a resulting in a very good delivery time (VG), and so on. In the figure, elements in dot lines belong to the ground elements of the argumentation theory of i. Arrows represent implication relation which are also in the argumentative theory. For instance, there is an implication relation in the theory that says:

$$Img(a, sell(dt), VG), Img(a, sell(q), G) \rightarrow Img(a, seller, VG)$$

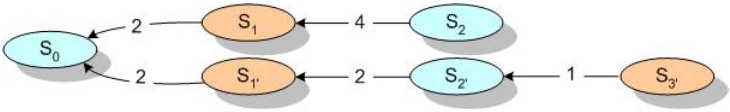

Figure 6.3: The dialectical tree of agent j after the game. Arrows represent attack relation and the labels indicate the strength of the attacks.

and another saying that

$$Rep(a, sell(q), G) \to Img(a, sell(q), G)$$

In figure 6.2 we show arguments and sub-arguments already instantiated to facilitate the reading.

The next table shows the illocutions that the agents exchange. The column *Dispute Line* shows the state of the dispute line.

Action	**Dispute Line**
-	$S^0 = \{S_0\}$
counter$^1_{OPP}(S_1)$	$S^1 = \{S_0, S_1\}$
counter$^2_{PRO}(S_2)$	$S^2 = \{S_0, S_1, S_2\}$
counter$^3_{OPP}(S_{1'})$	$S^3 = \{S_0, S_{1'}\}$
counter$^4_{PRO}(S_{2'})$	$S^4 = \{S_0, S_{1'}, S_{2'}\}$
counter$^5_{OPP}(S_{3'})$	$S^5 = \{S_0, S_{1'}, S_{2'}, S_{3'}\}$

In the first move, the opponent (OPP) utters the argument S_1 which attacks the original S_0. S_1 has the conclusion formula $Rep(a, sell(q), B)$ and attacks the subargument of S_0 that has as a conclusion $Rep(a, sell(q), G)$. The strength is calculated applying the function $\ominus$ on the values of the predicates. In this case, $\ominus(B, G) = 2$. In the next move, the proponent (PRO) attacks S_1 by sending S_2 (strength = 4). At this point, we assume that OPP cannot attack S_2, but it can attack again the original S_0. In movement 3, OPP sends the argument $S_{1'}$ to attack S_0 (strength = 2). Notice that the dispute line has changed. Then, the proponent counterargues $S_{1'}$ by sending $S_{2'}$ (strength = 2). Finally, OPP finishes the game at movement 5 by issuing $S_{3'}$, which attacks $S_{2'}$ (strength = 1).

The opponent wins the game. This means that OPP considers unreliable the initial information from PRO. The dialectical tree after the game is shown in figure 6.3. With this tree, OPP cannot construct an admissible set that includes S_0, and thus cannot accept it. But this is only true when OPP takes an inconsistency budget of 0. As soon as it tolerates a budget of 1, the result changes. Now, the set $\{S_0, S_2, S_{2'}, S_{3'}\}$ is a 1-preferred extension and S_0 becomes acceptable. At this point OPP, could update its reputation theory, considering that the information is reliable enough.

6.4 Related Work and Discussion

A review on reputation and trust models that use reliability measures calculated from the source can be found in [Sabater and Sierra, 2005]. In this related work though, we provide an overview of the work that takes advantage of the constituent elements of reputation-related concepts. For instance, models like [Huynh et al., 2006b] and [Maximilien and Singh, 2002] use *certified reputation*, in which the same target agent is able to justify its own reputation by presenting references (like reference letters when applying for a job). However, neither dialogs nor specific acceptability semantics is provided. Work presented in [Heras et al., 2009] explicitly uses argumentation techniques to handle recommendations. Its focus is bounded to peer-to-peer networks and recommendation systems. In a similar approach, in [Bentahar et al., 2007] reputation values are justified by the history of interactions and social network analysis. In this approach, argumentation is used as a theoretical reasoning process, instead of a dialectical procedure.

The work presented in [Pinyol and Sabater-Mir, 2007] analyses reputation-related concepts in terms of the internal elements used to infer them. However, it does not provide any formal protocol nor any acceptability semantics. More pragmatic approaches provide agent architectures for fuzzy argumentation on trust and reputation, but they lack formal definitions of acceptability [Stranders et al., 2008].

A promising research line that can be complementary to our approach comes from the solution of the trust alignment problem [Koster et al., 2009]. Their approach suggests that with the exchange of ground elements to justify trust values (they consider only interactions, composed of contracts and fulfillments), it is possible to model other agents' inferences through inductive logic algorithms. This approach requires though a very stable social groups where agents can gather a lot of shared interactions and relatively simple reputation models. Also, the exchanged of this kind of information can be seen as a violation of agents' privacy.

Finally, we do not want to forget the incursion of argumentation-based negotiation in the reputation and trust field. For instance, the work presented in [Morge, 2008] acknowledge the notion of trust as a multi-faced holistic construct, based on evaluable elements that can be used to argue and that lead the decision making. We can say that the approaches are somehow complementary. While our work focuses on the analysis of the internal elements of reputation-related components, contributing to the field of computational reputation models, negotiation approaches try to integrate it in argumentation-based negotiation processes.

6.5 Experimental Results

We have presented so far the theoretical development of the reputation argumentation framework and have discussed the reasons why we need such a system, and under which theoretical conditions the framework should help improving the performance of the agents. However, there are questions that are difficult (or impossible) to answer without experimental simulations. In particular, the most relevant question that arises from the previous development is very clear:

When is it worth it to use the reputation argumentation framework?

Firstly, there is an obvious trade off between time and cost, which in our framework is strictly related to *the number of exchanged messages* and the *achieved accuracy* respectively. In this sense, (1) if the cost of achieving a bad interaction is higher than the cost of messaging (or waiting time), the accuracy becomes a crucial issue, and argumentation can help. On the opposite, (2) when the cost of messaging dominates the potential failure of an interaction, for sure argumentation is not a good solution. Of course, we focus on (1), and assume that the cost of messaging is not relevant in comparison to the cost of a bad interaction [2].

Also, notice that the reputation argumentation framework is nothing else but a complement attached to an existing reputation model, which in general is already pretty accurate. In fact, there is no guarantee that the use of our argumentation protocol significantly improves the accuracy of the agents. In this section we present the results of a set of simulations that explore some parameters that we consider crucial, and that empirically validate that the argumentation-based protocol for social evaluation exchange significantly improves (statistically) the accuracy of the agents when modeling the behavior of others.

The simulations should be considered only a proof-of-concept environment that proves that in the scenario described below the use of argumentation is useful. Even when we would require a more complete set of experiments to completely validate the utility of the argumentation-based protocol, interesting conclusions can be extracted, and of course can be extrapolated to other scenarios.

6.5.1 Description of the Simulations

As in the simulations presented in chapter 2 we consider an scenario with buyers and sellers. In this scenario, sellers offer products with constant quality (q) and deliver them with a constant delivery time (dt). Buyers are endowed with a reputation model and evaluate sellers in the role *seller*, which is composed of *sell(q)* (quality dimension of the sold products) and *sell(dt)* (delivery time). Also, the context *Inf* stands for *informant*.

To enhance the utility of the protocol and simulate a difference between reputation models, we consider that different buyers can evaluate sellers in different ways (they have different goals). To simplify, some buyers only take into account good sellers accordingly to the quality of the products (QBuyers), while others,

[2] Even when the cost of messaging is null, if the cost of interacting is very low there is no motivation for the agents to exchange information. An experimental evidence of this can be found at [Sabater-Mir, 2003] (chapter 7)

only accordingly to the delivery time (DTBuyers). The crucial point is that initially, buyers communicate social evaluations only regarding the role *seller*. Because of that, a *good seller* for agent A is not necessary good for agent B.

Our main hypothesis is that when no argumentation is present, the introduction of information from unaligned buyers may bias the final accuracy of the model, while when using argumentation, such information can be filtered, and thus, the accuracy should improve. This evidence clashes with the idea that to argue, both the source and the recipient agents must have some knowledge about the environment, but not *too much*. If agents do not have any information (or few), no argumentation is possible. On the opposite, if agents have already a lot of information that includes a high number of direct trades, agents will not be able to respond, since direct experiences cannot be attacked. To parametrize the former situation we include a bootstrap phase, where agents explore the environment without arguing. To deal with the latter, we do not let agents to trade directly with all the agents, only with a subset of them. In concrete, our simulations have the following phases:

- **Bootstrap phase:** It is used to endow the buyers with some knowledge about the environment (that is, other sellers and buyers). At each turn, each buyer perform two task: (1) chooses randomly a seller, buying from it, and (2)sends a communication of an image predicate regarding a random seller or buyer (in this case as informant) to a random buyer agent. No argumentation is present in this phase. The number of direct trades and messages send at each turn by each buyer agent can be parameterized. In our simulations we allow to each agent one direct trade and one message per turn.

  To parametrize the fact that agents do not have too much information in terms of direct trades, a percentage of the buyers (*pctBuyers-Bootstrap*) can only perform direct trades with a percentage of sellers (*pctSellers-Bootstrap*). The other buyers can trade directly to any seller. Such special buyers will have to model the reminding sellers only using third-party information. Figure 6.4 illustrates the scenario in this phase.

- **Experimental phase:** After the bootstrapping phase we create a single Q-buyer agent (our subject of study) that wants to model the behavior of a set of sellers (which correspond to a 100 - pctSellers-Bootstrap% of the sellers [3]) before trading with one of them. As discussed earlier, both the source and the recipient of the communication must have some knowledge before arguing, and because of that, our subject of study needs to go also though a bootstrap phase.

  An intuitive example that fits into the structure of this phase is a situation in which a human buyer navigates though on-line forums starting new traces before making the decision of buying an expensive good (like a laptop, a car, etc.).

[3] We use the same *pctSellers-Bootstrap* value, but the set of sellers is not necessary the same as in the bootstrap phase

Once the subject of study finishes the bootstrap, the simulation proceeds. As said before, the subject of study wants to buy a good, and for this, he needs to model the behavior of the unknown sellers. It receives a single message from each buyer agent about each of the unknown sellers. Depending on the experimental condition, the studied agent will aggregate the communication directly to its reputation theory without any argumentation phase, or will argue and decide whether to accept or not the message. See figure 6.5 for an illustration of this phase.

Thus, the two experimental conditions are:

- **NO-ARG**: The studied agent does not use argumentation when receives the messages. This means that the reputation model is the only mechanism to avoid information from bad informants. Agents use an extension of the Repage system (see chapter 2) that contemplates an ontological dimension. In any case, the reputation model is able to detect bad informants comparing what they said with what they experienced. We recall here that we have two groups of buyer agents (QBuyer and DTBuyer) and that have different perspectives of what a *good* seller is.

- **ARG**: The studied agent and the source buyer agent (informant) engage in a dialog following the protocol described in the chapter. In this experimental condition the β parameter plays a crucial role. It is easy to see that the higher the β parameter, the closer the performance results to be to the NO-ARG condition, since when β is big enough, the argument is always accepted [Dunne et al., 2009].

The main parameters that we manage in the simulations are:

Parameter	**Description**
#sellers	Number of sellers (40)
#buyers	Number of buyers (20)
pctGoodQuality	Percentage of sellers that offer good quality (25%)
pctGoodDTime	Percentage of sellers that offer good delivery time (25%)
pctQBuyers	Percentage of QBuyers. 100 - **pctQBuyers** are DTBuyer (20%, 50%, 80%)
pctBuyers-bootstrap	Percentage of buyers that during the bootstrap phase can only trade with a subset of randomly selected sellers that represent the percentage pctSellers-bootstrap
pctSellers-bootstrap	It also determines the percentage of sellers that during the experimental phase the studied agent does bootstrap with. Then, the rest of sellers are those that the studied agent must discover only through messages (20%).
turnsBootstrap	Turns in the bootstrap phase. We use such parameter to control the amount of initial information
β	Inconsistency budget (0)

For the simulations, agents use the L_{rep} language taking M as $\langle\{VB, B, N, G, VG\}, \leq_s\rangle$, as in the example shown in this chapter. The performance of an execution computes how well the studied agent is able to model the unknown sellers. We compare the best possible evaluation as a seller (according to the parameters of the seller and the goals of the studied agent), with the actual evaluation. For instance, given a seller who offers a bad quality and a very good delivery time, the best theoretically evaluation for an agent that is only interested in the quality dimension should be B (bad). In the case that our studied agent has evaluated such seller as G, the difference between both evaluations gives us a measure of the achieved accuracy.

We make use of the difference function $\ominus$ defined over the ordered set M, in which we consider a mapping $f : \{VB, B, N, G, VG\} \rightarrow [0, 4] \cap I\!N$ where $f(VB) = 0$, $f(B) = 1$, $f(N) = 2$, $f(N) = 3$, $f(N) = 4$, and define $\ominus(X, Y) = |f(X) - f(Y)|$. Then, 0 is the minimum difference, when both evaluations have the exact same value, and 4 is the maximum, occurring when one evaluation is VG and the other VB.

In our simulations we define the accuracy as the percentage of improvement with respect to the expected difference of two random evaluations. It is easy to compute that given two random evaluations, their expected difference is exactly $\frac{40}{25} = 1.6$. The computation is summarized in the following table:

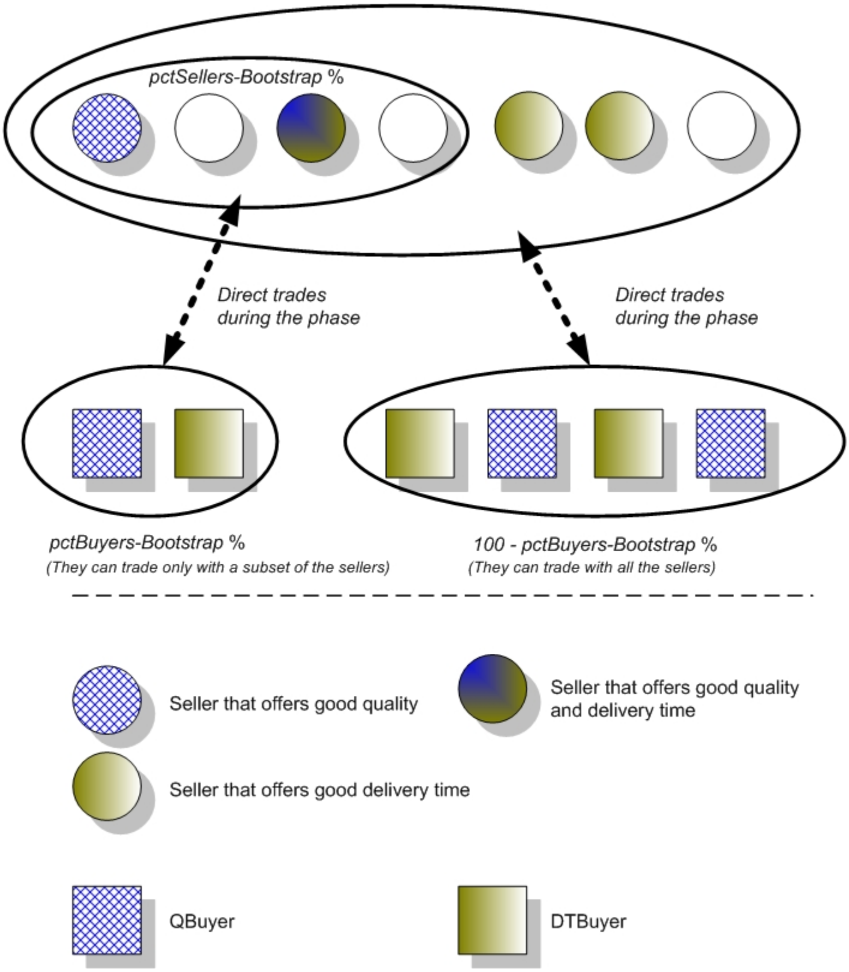
pctSellers-Bootstrap %
Direct trades during the phase
Direct trades during the phase
pctBuyers-Bootstrap %
(They can trade only with a subset of the sellers)
100 - pctBuyers-Bootstrap %
(They can trade with all the sellers)
Seller that offers good quality
Seller that offers good quality and delivery time
Seller that offers good delivery time
QBuyer
DTBuyer

Figure 6.4: A possible scenario during the bootstrap phase

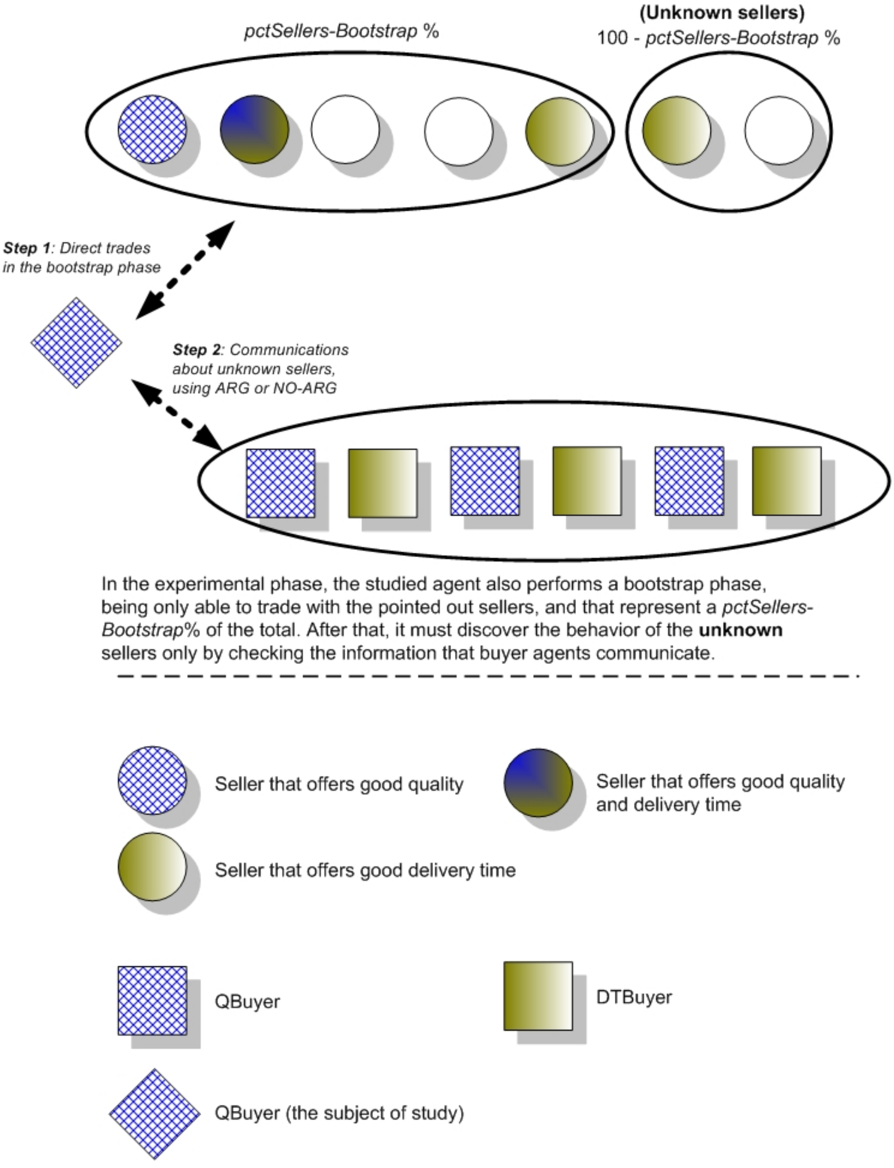
pctSellers-Bootstrap %
(Unknown sellers)
100 - pctSellers-Bootstrap %
Step 1: Direct trades in the bootstrap phase
Step 2: Communications about unknown sellers, using ARG or NO-ARG
In the experimental phase, the studied agent also performs a bootstrap phase, being only able to trade with the pointed out sellers, and that represent a pctSellers-Bootstrap% of the total. After that, it must discover the behavior of the unknown sellers only by checking the information that buyer agents communicate.
Seller that offers good quality
Seller that offers good quality and delivery time
Seller that offers good delivery time
QBuyer
DTBuyer
QBuyer (the subject of study)

Figure 6.5: A possible scenario during the experimental phase

Difference	**Possible Values**	**Prob.**	**Expected**
4	(VB, VG)	$2 \cdot \frac{1}{5} \cdot \frac{1}{5}$	0.32
3	(VB, G),(B, VG)	$2 \cdot 2 \cdot \frac{1}{5} \cdot \frac{1}{5}$	0.48
2	(VB, N),(B, G),(N, VG)	$2 \cdot 3 \cdot \frac{1}{5} \cdot \frac{1}{5}$	0.48
1	(VB, B),(B, N),(N, G),(G, VG)	$2 \cdot 4 \cdot \frac{1}{5} \cdot \frac{1}{5}$	0.32
		Total	**1.6**

For instance, to compute the partial expectation of obtaining a difference of 2, we have to realize that there are only three situations in which this occurs: (VB, N), (B, G), (N, VG). Therefore, the probability of archiving a situation in which the difference is two is $3 \cdot \frac{1}{5} \cdot \frac{1}{5}$. Since we also consider the symmetric situation (so, (N, VB), (G, B) and (VG, N)), the probability is in fact $2 \cdot 3 \cdot \frac{1}{5} \cdot \frac{1}{5} = 0.24$. Thus, the partial expected value is $0.24 \cdot 2 = 0.48$.

One expects that the reputation model improves such value, so, that the difference decreases to some extend from 1.6. For this, we calculate the average difference of all the unknown sellers and compute the percentage with respect to 1.6. For instance, an average difference of 0.5 improves 68.75% ($68.75 = (1.6 - 0.5) \cdot 100/1.6$), while an average difference of 0.3 improves 81.25% the random expected difference. Using this measure we can compare the experimental conditions ARG and NO-ARG.

6.5.2 Simulation Results

As said above, our main concern is to validate that the use of our argumentation mechanism improves significantly the accuracy of the agents in some conditions, and characterize them to some extend. Concretely, and in the terms used in the description of the experiment, the hypothesis is:

H: The experimental condition *ARG* achieves a higher improvement than the condition *NO-ARG*

We analyze such statement by the parameters *pctBuyers-Bootstrap* and *turn-Bootstrap*

pctBuyers-Bootstrap

The parameter models the number of buyer agents that in the bootstrap phase cannot interact with all the sellers, only with a subset of them. Then, when the parameter is high it indicates that most of the buyers are not able to explore directly some sellers, and when it is low, that most of the buyers are able to explore all the sellers. In our setting this parameter is an indicator of how well the set of buyers is informed. We theorized that too few information, as well as too much can be critical in the use of argumentation. The simulation results are in tune with this idea.

Figures 6.6, 6.7 and 6.8 show the performance of *ARG* and *NO-ARG* when varying *pctBuyers-Bootstrap* from 5% to 95% (setting *turnsBootstrap* to 20) with pctQBuyers=80%, pctQBuyers=50% and pctQBuyers=20% respectively . The

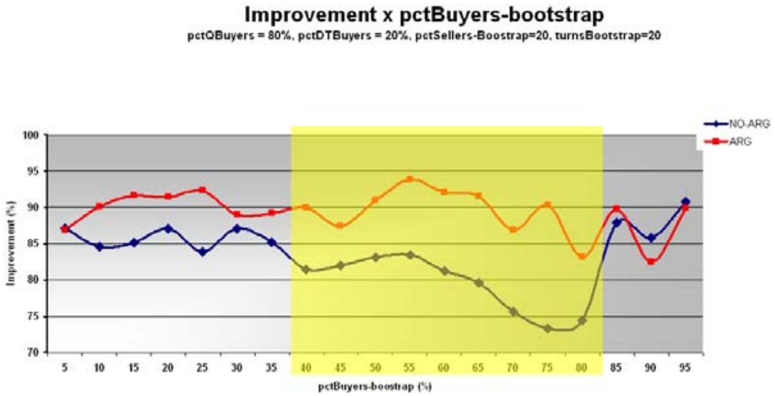

Figure 6.6: The performance of both experimental conditions varying the *pctBuyers-Boostrap*, with pctQBuyers=80% and pctDTBuyers = 20%

results confirm the hypothesis for most of the points in the graph[4] (we highlight such intervals in the figures). It is interesting to observe that all three graphs show a range of *pctBuyers-Bootstrap* in which the hypothesis is always confirmed with a p_value ≤ 0.01. The following table summarizes them:

pctQBuyers	**Intervals (%)**
80% (fig. 6.6)	40 - 80
50% (fig. 6.7)	35 - 70
20% (fig. 6.8)	5 - 70

The results indicate that when *pctBuyers-Bootstrap* is high (higher than 80%), ARG does not improve significantly NO-ARG. Those are the cases where there is not enough ground information to make useful the argumentation process. Also, when *pctBuyers-Bootstrap* is low, ARG does not necessary improve significantly NO-ARG. This is when the agents have already too much ground information that the studied agent cannot reject any argument.

It is also interesting to observe that as *pctBuyers-Bootstrap* increases, the accuracy of both ARG and NO-ARG decreases a bit. This is because direct trades offer always a better way to discover sellers than just communications. Then, when *pctBuyers-Bootstrap* is high, less buyers can directly interact with all the sellers.

[4] We performed t-test statistical analysis to validate whether *ARG* significantly improves *NO-ARG* with a *p_value* $\leq$ 0.01. When this is the case, we say that H is confirmed. For the statistical analysis, each simulation is repeated 20 times. To give arguments in favor of assuming normality on the distributions, we applied the Jarque-Bera (JB) test for normality, and we could not reject the null hypothesis, which assumes that the distribution follows a normal distribution.

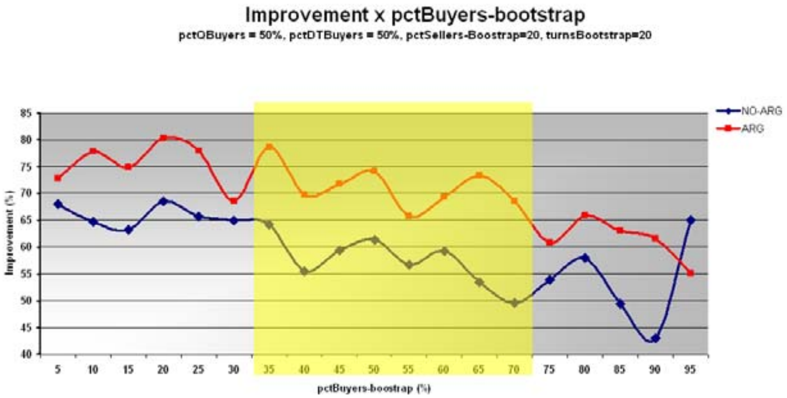

Figure 6.7: The performance of both experimental conditions varying the *pctBuyersBoostrap*, with pctQBuyers=50% and pctDTBuyers = 50%

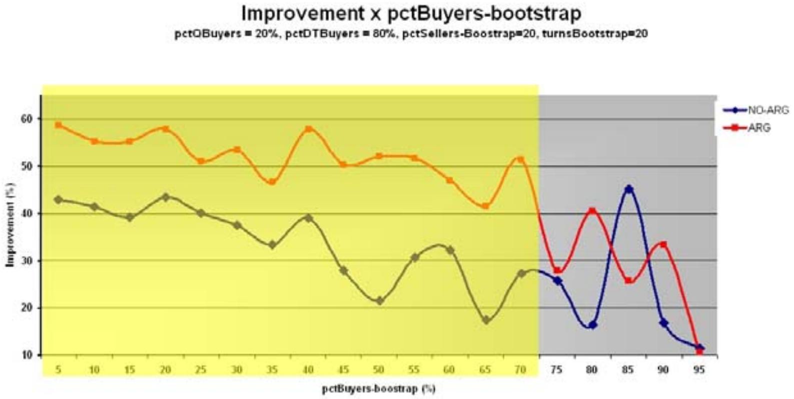

Figure 6.8: The performance of both experimental conditions varying the *pctBuyers-Boostrap*, with pctQBuyers=20% and pctDTBuyers = 80%

turnBootstrap

Related to the previous parameter, we want to study the amount of information that is needed to actually achieve an improvement by using argumentation. For this, we set the parameter *pctBuyers-Bootstrap* to 75% and vary the number of turns that agents spend in the bootstrap phase. The higher *turnsBootstrap* is, the higher the amount of information the agents will have about the sellers when the experimental phase starts.

Figures 6.9, 6.10 and 6.11 illustrate the obtained results with pctQBuyers=80%, pctQBuyers=50% and pctQBuyers=20% respectively. As expected, ARG does not perform better than NO-ARG until certain amount of data is managed by the agents. Figure 6.11 is maybe the most illustrative situation. There it can be observed how from 10 turns on, ARG is always better than NO-ARG. This is an indicator of the amount of information needed to take advantage of argumentation. The following table summarizes the intervals where ARG improves significantly NO-ARG.

pctQBuyers	**Intervals (turns)**
80% (fig. 6.9)	18-20
50% (fig. 6.10)	17-20
20% (fig. 6.11)	10-20

The intervals show the points in which the difference is statistically significant with a *p_value* < 0.01. However, some other points achieve a *p_value* < 0.05, which in many cases it would be enough to consider it a significant improvement.

It is also interesting the behavior of *pctQBuyers*. When it is 20% the improvement can be already appreciated in the turn 10, while when it is 50% and 80% the improvement is appreciated much later. We recall here that the studied agent is always a *QBuyer*, and then, when *pctQBuyers* is low, there are few *QBuyers*, so, few agents with the exact same goals. Thus, the results confirm that when the percentage of *QBuyers* is low, few bootstrap turns are enough to encourage the use of argumentation. Notice that when *pctQBuyers* is high, the achieved improvement by NO-ARG is already high, while it is low when *pctQBuyers* is low. We can extrapolate here that when everybody has similar goals it is not worth it to argue, while when it is not the case, argumentation can improve significantly the performance. The problem is that in real scenarios, it is hard to know the goals of the agents beforehand. Nevertheless, they can be learned by the agents.

6.5.3 Discussion

We have performed a set of simulations to empirically validate the utility of the argumentation protocol. We compare the accuracy obtained by agents using our argumentation protocol (ARG), and those not using it (NO-ARG). We show how in most of the checked conditions, ARG improves significantly (p_value ≤ 0.01) the accuracy obtained by NO-ARG. We explore several parameters and provide

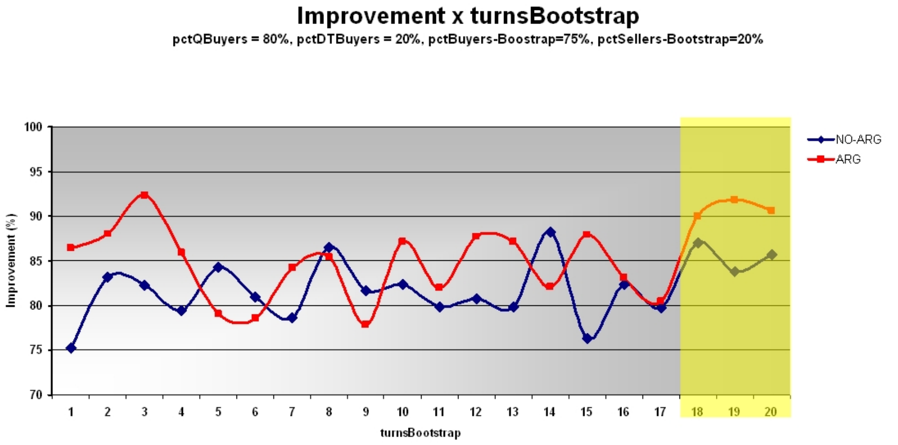

Figure 6.9: The performance of both experimental conditions varying the *turns-Bootstrap*, with *pctQBuyers*=80% and *pctDTBuyers* = 20%

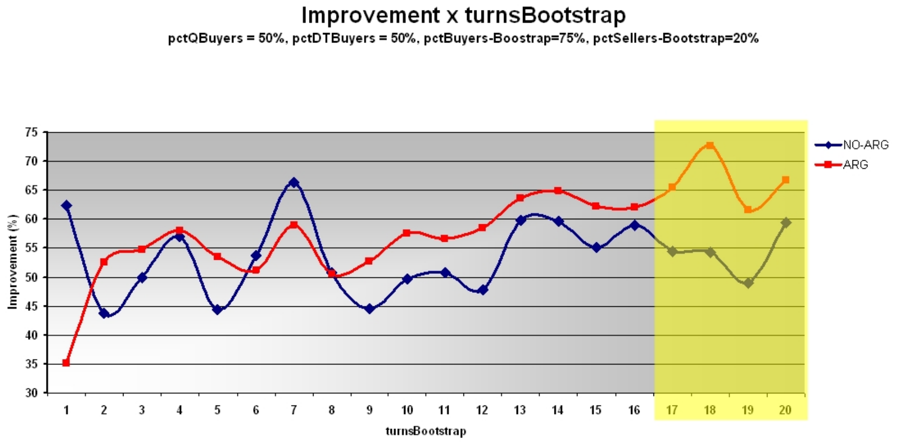

Figure 6.10: The performance of both experimental conditions varying the *turns-Bootstrap*, with *pctQBuyers*=50% and *pctDTBuyers* = 50%

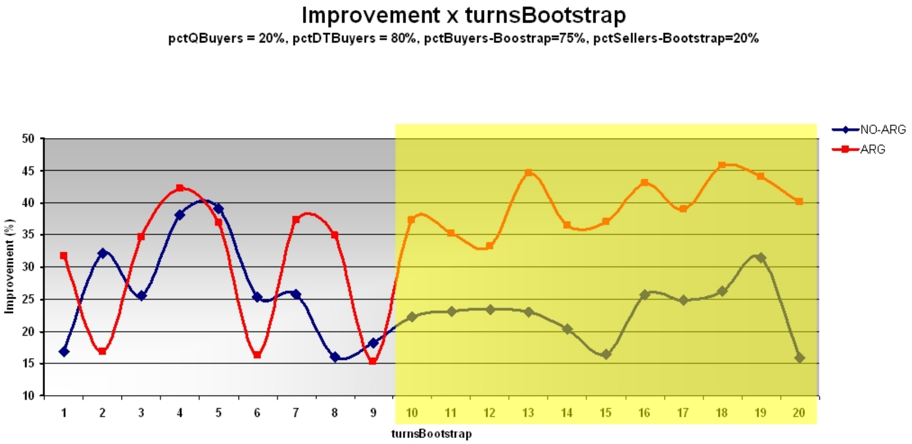

Figure 6.11: The performance of both experimental conditions varying the *turnsBootstrap*, with *pctQBuyers*=20% and *pctDTBuyers* = 80%

intervals in which ARG works *better* than NO-ARG. In concrete, we show how when (1) there is an heterogeneity of agents (not everybody has the same goals) and (2) agents do not base all their inferences in direct experiences, agents using argumentation achieve significantly a better accuracy that agents not using it.

When everybody has similar goals the gain in accuracy obtained using ARG may not be significant (1). The reason is that though argumentation, agents can reject information that they consider not reliable. In our settings where no cheating is present, when the goals are similar the inclusion of the communications into the reputation theories does not produce much bias in the new deductions. In other words, the reputation mechanism by itself obtains very good accuracy levels that are difficult to be improved by argumentation techniques.

The empirical analysis show the importance of the bootstrapping phase, which models the fact that argumentation is useful when agents are endowed with some knowledge (2). Regarding this, the experiments reveal that given a set of parameters, there is always a need for a certain number of bootstrapping turns to make ARG better than NO-ARG. This situation is specially illustrated by the figure 6.11.

We want to remark that the presented simulations were performed independently for ARG and NO-ARG. It means that for each simulation, a bootstrap phase was executed and either ARG or NO-ARG where executed. We partially explored what happens when after the same bootstrap phase, we run ARG and NO-ARG. We observe that in most of the cases (almost all of them) ARG performs better than NO-ARG. Preliminary results are illustrated in figure 6.12. We let for future work a more exhaustive exploration in this direction, plus the study of other parameters.

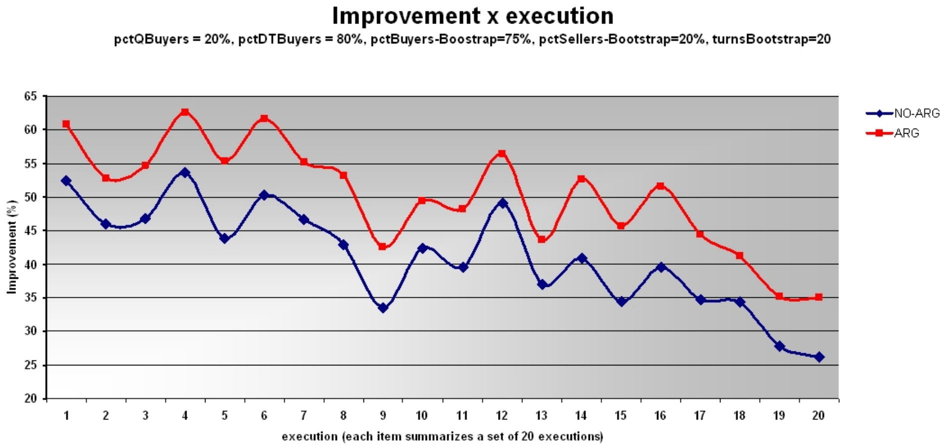

Figure 6.12: The performance of both experimental conditions keeping the same bootstrap data, with fix parameters. We illustrate how when keeping the same bootstrap data, argumentation always performs better than not using it.

6.6 Conclusions

In this chapter, we have defined an argumentation-based protocol for the exchange of reputation-related information that allows agents to judge whether a given piece of information is reliable or not. We use argumentation techniques to give semantics to the protocol.

We have made an important assumption: the agents use the same language to *talk* about reputation concepts. This requires that the concepts described by the language have the same semantics for both agents. We allow though the use of different deduction rules to infer the predicates. In the case agents use different semantics they should engage first in a process of ontology alignment.

At the theoretical level, the next step regarding this work will be the inclusion of defeats among arguments. We plan to use the typology of ground elements to give strength to the arguments, independent of their attack relations. For instance, one may consider that arguments based on direct experiences are stronger than those based on communications.

It is also important to remark that as shown in [Prakken, 2000], dialog games in dynamic contexts may be neither sound nor complete. This implies that we can only ensure the correctness of the presented dialog protocol in static environments, so, when the information that agents have remains static during the game. This constrain can be too strong in certain scenarios, specially when each step of the dialog can be considered a communication that may change the internal mental state of the recipient agents. This is a research line that should be explored and investigated in the future.

Chapter 7

Conclusions and Future Work

In this final chapter we summarize the main contributions of this book that have already appeared partially in the respective sections. We also state the future lines of research that this work has opened.

7.1 Reasoning Using Social Evaluations

We have presented the BDI+Repage model, a BDI agent architecture that integrates in the reasoning process reputation information from the existing Repage model. The main characteristics of the model have been already mentioned in chapter 1, but in any case, we want to remark the following features:

- The model is defined as a multi-context system (MCS), a framework that allows several distinct theoretical components to be specified together, with a mechanism to relate these components. From a software engineering perspective, MCS supports modular architectures and encapsulation. This modular architecture permits smooth integrations of possible modules that could extend the functionalities of the original one

- We use DC-logic and IC-logic extracted from [Casali, 2008] to model the desires and intentions of the agent. This allows the agents to be endowed with graded attitudes. Also, we introduce a new logic to deal with the beliefs of the agent. The belief logic is a classical first-oder many-sorted logic, deals with probabilities and is capable of representing and combine the information that the reputation model Repage computes.

- It handles *image* and *reputation*. The Repage model is based on a cognitive theory of reputation that states a main difference between image and reputation. The belief logic that we develop captures both concepts and combines them, defining a family of agents depending on how such combination is performed. It is important to remark that we use Repage as a paradigmatic example, but any model whose information can be captured by the reputation language L_{rep} could be placed into the system.

Considering trust as a decision to rely on somebody, the BDI+Repage model can be seen as a trust model, becoming as far as we know the only cognitive trust model which defines analytically the components of the trust mental state. In this sense, when an agent that uses the BDI+Repage architecture makes a decision, the mental state is composed of a set of beliefs, desires and intentions. Some of the beliefs are generated from Repage and model several aspects that other cognitive models of trust suggest, like competence and disposition beliefs [Castelfranchi and Falcone, 1998b, Herzig et al., 2008]. The BDI+Repage model assumes that the general desires of the agents are already in the system (with their respective grades). These generic desires correspond in fact to the *wishes* from Castelfranchi & Falcone's model [Castelfranchi and Falcone, 1998b], and the final intention generated though the reasoning process, to the goal attitude defined in [Castelfranchi and Falcone, 1998b].

We also provide the BDI+Repage+Norm model, an extension of the original BDI+Repage architecture that shows the flexibility of the model. The BDI+Repage+Norm architecture uses the reputation model to evaluate the accomplishment or not of the norms, which are specified in a new *Norm Context*. Even when a big amount of effort has been put into the study of normative systems from an organizational perspective or through deontic aspects that restrict the possible actions agents can perform, few attention has been paid to how agents evaluate such norms and use them to reason. The BDI+Repage+Norm proposes a solution for this by specifying the normative language L_{norm} to express norms, and integrating it to the BDI+Repage system.

From the examples we place in chapter 5 it should be clear that on one hand, epistemic decisions play a crucial role in the pragmatic-strategic decisions of the agent, and that a formal model for its integration improves the conceptualization of the reasoning process. On the other hand, the consequences of pragmatic-strategic decisions may also effect the epistemic decisions, implementing somehow the loop that Conte and Paolucci in [Conte and Paolucci, 2002] state.

7.1.1 Future Work

The future work regarding this part concerns several aspects that exploit the usage of the BDI+Repage model and the empirical evaluation of certain properties regarding the relationship between image and reputation. In appendix B we already explore an heuristic implemented through a simple Q-learning algorithm that helps agents to decide whether to rely on image more that reputation information or vice versa. We statistically validate that the method is effective in the simple and artificial scenario we provide. However such exploration requires more development:

- It would be nice to generalize the heuristic presented in the appendix to include not only the classes $\mathcal{H}_3$, $\mathcal{H}_4$ and $\mathcal{H}_5$, but the classes $\mathcal{H}_1$ and $\mathcal{H}_2$, which consider only image and only reputation respectively. This is to establish under which (environmental or social) conditions the agent can withdraw image or reputation from the system.

- Also, the experiments performed in the appendix B considers a very simple environment. Future work requires an exhaustive study of a different typology of scenarios. For example, reproducing the simulations described in chapter 2, with societies where the number of cheaters differ, or where the fear of retaliation is present, with few/many resources, or few/many bad/good informers, etc.

- An empirical comparison of the model with other existing models is a challenging objective. The reason is that there are no models that make use of the distinction between image and reputation, which is one of the main characteristics of the model presented here. However, some models could be slidely modified to capture such difference.

Regarding the future research on the development of the model, we want to integrate it in argumentation-based negotiation processes. Some published work [Parsons et al., 1998] already makes use of a multicontext BDI agent to define processes of negotiation through argumentation, that we could exploit using the BDI+Repage model. The challenges regarding this issue are:

- The adaptation of/to the negotiation model with graded information. The BDI+Repage model incorporates graded attitudes that should be taken into account when arguing in a process of negotiation. In our case, agents not only *desire* certain goals but they do it with a degree.

- Also, since our model incorporates a reputation model, such valuable source of information should be integrated in the negotiation process. The integration will enrich the accuracy of the process and help agents determine *better* choices accordingly to their individual objectives.

- In chapter 6 we propose an argumentation system that already consider graded information. Its focus though relies on the internal elements of the reputation model. To develop argumentation-based negotiation protocols we need to develop argumentation systems at the BDI level, where beliefs, desires and intentions justify themselves through the bridge rules. Some work regarding practical reasoning have been done where these three attitudes are the main part of the arguments.

7.2 Dialogs and Argumentation

The other main contribution faces definition of an argumentation-based dialog protocol for the exchange of reputation-related information. The protocol intends to give an alternative solution to one of the main problems in the field of trust and reputation models, regarding the subjectivity of reputation information and its harmful consequences when it is communicated. Due to the subjectivity of reputation information, a social evaluation totally reliable for an agent A may not be reliable for B, because the bases under which A has inferred the social evaluation cannot be accepted by B. This can happen because agents have different inference rules, have had different experiences, have different goals, etc. When such information is communicated this can become very problematic, specially if the reputation model assigns a reliability measure to the communicated information, because it depends on the source agent.

The system we propose offers a possible solution for this, and can complement already existing methods. Taking advantage of the internal structure of reputation-related information, rather than allow only single communications, we allow agents to *justify* their communications following the guidelines of the argumentation-based protocol. Then, the agent can incrementally construct a tree of arguments with their attack relations that can be used to decide on the reliability (and thus acceptance) of a communicated social evaluation. The main features of the system are:

- The recipient agent is who decides about the reliability of a communicated evaluation. This makes more difficult for dishonest agent to intentionally send fraudulent information, because they must be aware of the knowledge of the recipient.

- It handles quantitative and qualitative graded information. One of the main characteristics of reputation information is that it is graded. Nowadays it is strange to find a model that provides crisp evaluations of the agents. For instance, an agent A may be *bad*, *very bad* or *very good* etc. as a car driver, and this has to be taken into account when arguing about evaluations. For this, we make use of the weighted argument system defined in [Dunne et al., 2009].

The system is generic enough to permit interactions among agents that use different reputation models. We only require that the language of such models must be captured by L_{rep}.

We also provide some empirical validations of the system. The simulation experiments confirm that when (i) there is an heterogeneity of agents, (ii) they do not base all their inferences in direct experiences (they have not explored all the environment by direct interactions), and (iii) agents are partially endowed with a moderate amount of information, agents that use our argumentation protocol improve significantly the accuracy when modeling sellers. The results assume that the cost of direct trades is high, while the cost of communication is very low. If this is not the case, the agents do not have the motivation to communicate.

7.2.1 Future Work

We have several research lines in mind regarding this topic:

- We want to introduce explicitly the notion of defeat among arguments. In this sense, we plan to use the typology of ground elements to give strength to the arguments, independently of their attack relations. For instance, one may consider that arguments based on direct experiences are stronger than those based on communications. This requires further development, since as far as we know, no argument system has included yet strength in the attacks and weights in the arguments.

- Another research line involves the extension of L_{rep} to include possible arguments about ground elements, specially direct experiences. The framework presented in this work omit such kind of arguments, arguing that this could effect the privacy of the agents. Nevertheless, some promising work [Koster et al., 2009] makes use of such basic interactions to establish alignments between different trust models. We think both approaches are complementary and could be used together.

- Also, more empirical valuations are required. We want to perform an exhaustive exploration of scenarios where the application of argumentation about reputation-related concepts makes a difference. In particular, we would like to play with the percentage of cheaters and fraudulent information, the heterogeneity level of the agents in the society, the level of inconsistencies etc...

- It would be also nice to study the impact of the most classical attacks in virtual societies, specially whitewashing. We theorize that an argumentation system like ours could discourage such behaviors by adjusting inconsistency budgets.

- Also, we would like to study the effect of allowing argumentation in P2P systems, with different topologies of networks. The literature on P2P is extensive and the attempts to minimize the impact of whitewashing or free-raiders attacks are the main issues. Some of them are based on reputation-based trust (other are policy-based trust, where the trust on peers in built through the exchange of credentials), and a similar argumentation framework presented in this book could be very useful.

Appendix A

Entropy of the Representations

A.1 Introduction

As stated before, in a system where participants may be using different kinds of reputation and trust models, the necessity of exchanging social evaluations to achieve their goals may drive in a situation where an agent that uses a boolean representation needs to communicate with one that uses probabilistic distribution, and then, a conversion of representations must take place. However, type conversions carry lose of precision and addition of uncertainty. As an example, some evaluation represented as a boolean that is Bad, when is converted to a real representation may have an evaluation from 0 to 0.5 (not included), when is converted to discrete set, it may be one of these elements $\{VB, B\}$ etc... This factor of uncertainty that is added when we convert a value to a more expressive representation is what we call Conversion Uncertainty (CU), and is an information that the recipient should know.

A.2 Entropy of the Representations

In order to calculate the CU we use the information theory approach introduced by Shannon [Shannon, 1948]. In this context, the entropy of a random variable X ($H(X)$) can be understood as the *uncertainty* of X, and is defined as

$$H(X) = -\sum_{x \in X} p(X = x) \log(p(X = x)) \tag{A.1}$$

From Shannons's theory we can define the conditional entropy as follows:

$$H(X|Y = x) = -\sum_{x \in X} p(X = x|y = Y) \log(p(X = x|Y = y)) \tag{A.2}$$

Type	Entropy
BO	1.00
DS	2.31
RE	6.64
PD	22.19

Table A.1: Entropies of the type representation

and finally,

$$H(X|Y) = -\sum_{y \in Y} p(Y=y)H(X|Y=y) \tag{A.3}$$

Now, we consider each one of the representations as discrete random variables. Without lose of generality we can discretize the Real representation using two digits (in base 10), having a hundred possible values. The fact of using 100 divisions for the interval and not a bigger amount is because we think that a greater precision is completely unnecessary (and even counterproductive) given the nature of the measure that is represented with this value, that is, a measure of a social evaluation. At the same time, taking into account the hundred possible values of a Real number, we can count the number of elements of the Probabilistic Distribution representation considering all possible combinations of distribution values that need to achieve the unit[1]. Let A be this number, the following equation holds:

$$A = \sum_{i=0}^{4} \binom{5}{i}\binom{100}{4-i} = 4780230 \tag{A.4}$$

Each random variable has as elements each possible element of the representations and its probability distribution is totally equiprobable. Then, we define the conversion uncertainty of the source random variable X to the target random variable Y as

$$CU(X,Y) = H(Y|X) \tag{A.5}$$

In other words, CU is the increment of uncertainty produced when a value is represented in X and it is converted to a value of type Y, which is more expressive. There is a set of candidate values that makes conditional entropy increase. The values of the entropy of each type is showed in table A.1. See appendix A for the details of the calculus.

The CU values for each conversion is showed in table A.2. Each row is the source and each column is the target.

An example will illustrate the usage of the CU value. Let's suppose agent A is using a Boolean representation, and generates and sends an evaluation to agent B that uses a discrete set representation. Agent B would reach the evaluation with a CU value of 1.29. If agent B send the same evaluation to agent C that

[1] This is a combinatorial problem related to the famous Balls and Bins problem

	BO	DS	RE	PD
BO	0	1.29	5.64	21.19
DS	0	0	4.32	19.89
RE	0	0	0	15.55
PD	0	0	0	0

Table A.2: CU values

uses a probabilistic distribution, agent C would receive the evaluation with a CU value of 1.29 (the base value coming from the communication) plus 19.89 (from the type conversion between DS to PD), it means, a CU value of 21.18. The idea is that the uncertainty of the evaluations is accumulative, without allowing loops (if the evaluation goes back to an agent using a representation type that have already been used in some transformation there is no addition of uncertainty)

A.3 Calculus of CU

In this section we provide the calculus to compute each one of the CU that are summarized in table A.2.

$CU(BO, DS) = 1.29$

Considering True as t and False as f:

$$CU(BO, DS) = H(DS|BO) \tag{A.6}$$

Knowing that $p(BO = t) = p(BO = f) = \frac{1}{2}$ we can write that

$$CU(BO, DS) = \frac{1}{2}H(DS|BO = t) + \frac{1}{2}H(DS|BO = f) \tag{A.7}$$

At this point, when $BO = t$ and following our semantic interpretation we know that it may refer to one value of the set $\{N, G, VG\}$, and if $BO = f$ of the set {VB,B}. Then $P(DS = \{VB\}|BO = f) = P(DS = \{B\}|BO = f) = 1/2$ (zero in other values of DS) and $P(DS = \{N\}|BO = t) = P(DS = \{G\}|BO = t) = P(DS = \{VG\}|BO = t) = 1/3$ (zero in other values of LL). Then, following the previous equations and developing the entropy formula we have that

$$H(DS|BO = t) = -3\frac{1}{3}\log(\frac{1}{3}) \approx 1.58 \tag{A.8}$$

$$H(DS|BO = f) = -2\frac{1}{2}\log(\frac{1}{2}) = 1 \tag{A.9}$$

finally, computing the equation A.7 we have

$$CU(BO, DS) = 0.79 + 0.5 = 1.29 \tag{A.10}$$

$CU(BO, RE) = 5.64$

Here, knowing that $BO = t$ our semantic indicates that as a real, it could be a value from 0.50 and 1, then $\forall_{i \in [0,1]} p(RE = i|BO = t) = p(RE = i|BO = f) = 1/50$ and therefore,

$$H(RE|BO = t) = H(RE|BO = f) = -50\frac{1}{50}\log(\frac{1}{50}) \approx 5.64 \tag{A.11}$$

$$CU(BO, RE) = 5.64 \tag{A.12}$$

$CU(BO, PD) = 21.19$

Having in mind the total number possible elements in PD (see equation A.4), we know that $BO = t$ implies that whatever representation of PD will tend towards a good evaluation, it means that the probability of being good is higher that the opposite. That eliminates exactly 50% of all the representations, and therefore

$$\forall_{i \in PD} p(PD = i|BO = t) = p(PD = i|BO = f) = \frac{2}{A} \tag{A.13}$$

$$H(PD|BO = t) = H(PD|BO = f) = -\frac{A}{2}\frac{2}{A}\log(\frac{2}{A}) \approx 21.19 \tag{A.14}$$

$$CU(BO, PD) = 21.19 \tag{A.15}$$

$CU(DS, RE) = 4.32$

Following the same reasoning:

$$CU(DS, RE) = H(RE|DS) \tag{A.16}$$

$$CU(DS, RE) = \sum_{i \in \{vb,b,n,g,vg\}} \frac{1}{5} H(RE|DS = i) \tag{A.17}$$

Notice that in this case, the difference between a Real and DS is that the first is continuous and the second discrete. Then, dividing the $[0, 1]$ interval into five identical parts, and assigning each of them into a value of DS we have the problem almost done. In this situation, each value of DS correspond to a 20 values of Real, and therefore,

$$\forall_{i \in \{vb,b,n,g,vb\}} \forall_{j \in [0,1]} p(RE = j|DS = i) = \frac{1}{20} \tag{A.18}$$

Then,

$$\forall_{i \in \{vb,b,n,g,vb\}} H(RE|DS = i) = -20\frac{1}{20}\log(\frac{1}{20}) \approx 4.32 \tag{A.19}$$

and then,

$$CU(DS, RE) = 4.32 \tag{A.20}$$

$CU(DS, PD) = 19.89$

The key in all the calculus is in the fact that each element of DS may correspond to a set of elements of PD whose center of mass is included in the interval corresponding to the function defined in R'. In the same way we have discretized the interval $[0, 1]$ in five parts, for each of these intervals we have a total of $\frac{A}{5}$ elements of PD with a center of mass that points inside the interval. Therefore, we can establish the following statement:

$$\forall_{i\in\{vb,b,n,g,vb\}}\forall_{j\in PD}p(PD = j|DS = i) = \frac{5}{A} \tag{A.21}$$

and,

$$\forall_{i\in\{vb,b,n,g,vb\}}H(PD|DS = i) = -\frac{A}{5}\frac{5}{A}\log(\frac{5}{A}) \approx 19.89 \tag{A.22}$$

then,

$$CU(DS, PD) \approx 19.89 \tag{A.23}$$

$CU(RE, PD) = 15.55$

Following the same reasoning than in the previous point, the number of elements of PD whose center of mass is the one being converted is approximately $\frac{A}{100}$, and therefore,

$$\forall_{i\in[0,1]}\forall_{j\in PD}p(PD = j|RE = i) = \frac{100}{A} \tag{A.24}$$

and,

$$\forall_{i\in\{vb,b,n,g,vb\}}H(PD|RE = i) = -\frac{A}{100}\frac{100}{A}\log(\frac{100}{A}) \approx 15.55 \tag{A.25}$$

then,

$$CU(RE, PD) = 15.55 \tag{A.26}$$

Appendix B

An Heuristic for the Axiom IRB

B.1 Introduction

In chapter 5 we have presented the BDI+Repage model which makes use of the belief language L_{BC} defined in chapter 4. There, we define a typology of agents depending on the axiom IRB:

$$\forall axp_1p_2r(E(a,x,p_1,r) \wedge S(a,x,p_2,r)) \rightarrow B(a,x,h(p_1,p_2),r)$$

We propose a family of agents whose h function is defined generically as

$$h(p_E,p_S) = \frac{\delta_E \cdot p_E + \delta_S \cdot p_S}{\delta_E + \delta_S}$$

where $\delta_E, \delta_S \in \mathbb{Q}_{\geq}$. Table 4.1 summarizes the behavior of a family of agents depending on the values of δ_E and δ_S.

When considering this generic definition, one question arises: which is the best function? Notice that by changing it the reasoning process of the agents is touched. We theorize that it is context dependent. To investigate a little bit more in this direction, we propose an heuristic process that decides at each turn which is the best function to use. We validate it through simulations.

The experiments should be also understood as a proof-of-concept platform that shows the viability of possible implementations of the BDI+Repage model. It should not be considered a complete empirical validation of the model. This is only a first step towards a potential set of simulation experiments that can be done using our model to answer questions about image and reputation. Further possible simulations are described in the future work section (chapter 7).

B.2 A Metaprocess for Updating Function h

We propose a mechanism that learns which is the best function h at each stage. For this, we consider the h function as

$$h(p_E, p_S) = (1 - Z) \cdot p_E + Z \cdot p_S$$

where $Z \in [0, 1] \cap \mathbb{Q}$. Then, our process only needs to decide which value to assign to Z. Notice that when Z is 0, the agent only takes into account image information. When Z is 1, reputation information is more important. Previous work on cognitive theories and simulation of image and reputation dynamics (see chapter 2) reveals that the amount of reputation information that circulates in a society is a lot higher than image-based information, due to the implicit commitment that sending image information carries out [1].

However, even when reputation information is mostly inaccurate, open societies perform *better* when reputation information is allowed in the system, and also are more robust with respect to certain level of cheating information [2]. This indicates that agents face mostly inaccurate information but that they need to use it to face real uncertain and unpredictable scenarios.

These studies are very helpful when defining a process to decide Z. Our IRB axiom is in fact a predicate that indicates how much information that circulates in the society can be considered true by the agent. In the way we have defined rules A_I and A_R, settings of Z tending to 0 could be useful when the number of cheaters is considerably big, meanwhile settings of Z close to 1 would be helpful in the opposite way. Then, our process tries to calculate how *different* image and reputation information results to be.

As defined in chapter 4, Repage provides image and reputation information as

$$img(j, r, [v_1, \ldots, v_t])$$

$$rep(j, r, [v_1, \ldots, v_t])$$

where j is the target agent, r is the role and $[v_1, \ldots, v_n]$ is the value of the social evaluation. The value represents a probability distribution over a sorted set of labels, like {Very bad, bad, neutral, good, very good}. The informal idea is to average the distances between all pairs of image and reputation values corresponding to the same target agent and role to estimate the value of Z. Then, if for most of the agents and roles, the values of image and reputation

[1] As explained in [Pinyol et al., 2007a], when an agent communicates image information, she is in fact informing about her thoughts, about what she *thinks*. The source agent is both revealing her identity and certifying that the information is true. Because of the fear of retaliation from the other members of the society, an agent only sends image information if she is mostly certain about it. If this is not the case, it is more likely that this information is not communicated or communicated as reputation, which involves a detachment from the source of information and thus, no commitment.

[2] More than 50% of cheaters in a society still produces a benefit in the overall performance when reputation communication is allowed. See chapter 2.

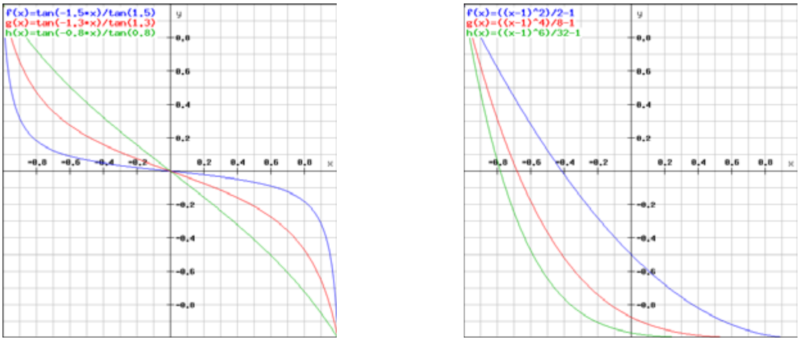

Figure B.1: Possible *scaleZ* functions. x is the average distance, and y the estimated Z

considerably differ, Z should tend to 0. Otherwise, it should tent to 1. In the following lines we formalize this process, whose schema is shown in figure B.2.

Firstly, we state the definition of the function *dist* that provides a measure of the distance between two Repage evaluations.

Definition Let w_1 and w_2 be Repage evaluations with t partitions, and w_1j the j-th value of the evaluation w_1. The distance between them is calculated as

$$dist(w_1, w_2) = \sum_{j=1}^{t} |w_{1j} - w_{2j}| \cdot |(j - 1 - CM(w_1, w_2))| \tag{B.1}$$

where CM (center of mass) is

$$CM(w_1, w_2) = \frac{\sum_{j=1}^{t} |w_{1j} - w_{2j}| \cdot (j - 1)}{\sum_{j=1}^{n} |w_{1j} - w_{2j}|} \tag{B.2}$$

Notice that the maximum possible distance is exactly $t - 1$. This is the case in which one evaluation is $(1, 0, \ldots, 0)$ and the other $(0, \ldots, 0, 1)$. The minimum distance is 0 and this is the case when both evaluations have exactly the same weights. We usually present this value normalized between -1 and 1, where -1 is the minimum difference, and 1 the maximum:

$$dist_N(w_1, w_2) = 2 \cdot dist(w_1, w_2)/(t - 1) - 1 \tag{B.3}$$

We use this function to calculate the general distance between image and reputation predicates about the same agent playing the same role. Let $S = \{s_1, \ldots, s_n\}$ be the set composed of pairs of image and reputation predicates from Repage, where each $s_i = \langle img(a, r, w_1)_i, rep(a, r, w_2)_i \rangle_i$, a is an agent name, r is a role name and w_1 and w_2 are evaluations. The average distance between each pair is calculated as

$$distAvg(S) = \frac{\sum_{i=1}^{n} (dist_N(s_i.w_1, s_i.w_2))}{n} \tag{B.4}$$

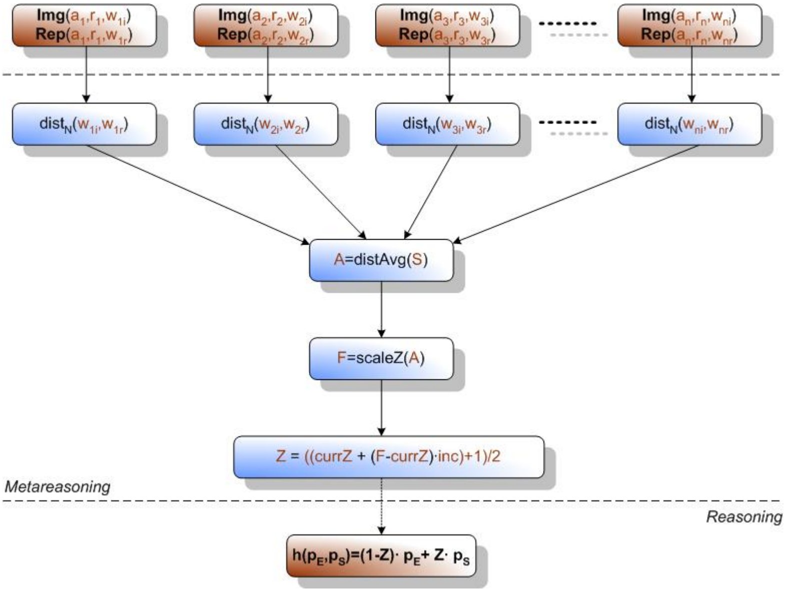
Img(a_1,r_1,w_{1i})
Rep(a_1,r_1,w_{1r})
Img(a_2,r_2,w_{2i})
Rep(a_2,r_2,w_{2r})
Img(a_3,r_3,w_{3i})
Rep(a_3,r_3,w_{3r})
Img(a_n,r_n,w_{ni})
Rep(a_n,r_n,w_{nr})
dist_N(w_{1i},w_{1r})
dist_N(w_{2i},w_{2r})
dist_N(w_{3i},w_{3r})
dist_N(w_{ni},w_{nr})
A=distAvg(S)
F=scaleZ(A)
Z = ((currZ + (F-currZ)·inc)+1)/2
Metareasoning
Reasoning
h(p_E,p_S)=(1-Z)· p_E+ Z· p_S

Figure B.2: Schema for the calculus of the new Z value.

This gives us the average distance between what the agent believes and what the agent believes to be said (in terms of Repage predicates). Aforesaid, lower values should carry higher values of Z, while higher differences, low Z. To scale this measure, we can consider several functions (named *scaleZ*). We show some of them in figure B.1.

Once we have this value we are ready to update the current Z value of the agent. Let *currZ* be the current Z level of the agent, the new Z value is calculated as

$$newZ = currZ + (scaleZ(distAvg(S)) - currZ) * inc \qquad \text{(B.5)}$$

where *inc* is the increment index, or learning rate (from 0 to 1). This schema follows the classical Q-learning equation. Then, with $inc = 0$ the agent does not learn anything and with $inc = 1$ only the last value is taken into account. We need to normalize this value in the interval [0,1]. The final value is then

$$(newZ + 1)/2$$

In the next section we validate the previous process by simulating a simple market.

B.3 Validation of the Proposed Method

B.3.1 Scenario and Simulation Settings

To validate the proposed method, we replicate a simple wine market with buyers, sellers and informants. In this scenario, all sellers offer wine that has certain quality. Also, a delivery time expressed in weeks is associated with the seller. Buyers are BDI agents following the model described in this paper. Therefore, the goals of the agents are described in terms of graded desires. The set of informant agents send out reputation information about the sellers. We control the experiment by setting a percentage of informants that spread *wrong* reputation (liars), the number of sellers and the distribution of qualities and delivery times.

We focus our attention in the performance of the buyer agents. Figure B.3 shows the sequence diagram of a single turn. First, all informant agents send reputation information about the sellers. Each informant communicates at each turn one reputation communication referring to a single seller agent and focusing on either the quality of the products offered by such seller or the delivery time of the products. The buyer agent incorporates all these information into the Repage model. After the communications, the buyer starts the BDI reasoning. The result is a decision. In this case, the purchase of the product from the *best* reasonable seller according to the desires of the buyer agent. Once the purchase is done, the buyer agent receives a fulfillment indicating the quality and the delivery time of the product. This new direct experience is introduced into the Repage model, before starting the metareasoning process to updates bridge rules and axioms. After that, the turn finishes and the buyer agent is evaluated.

Seller configurations

Each seller has two parameters: Quality of the product offered and the delivery time of the product that they achieve. To simplify the simulations we consider that the quality of the wine has four possible values: *excellentWine, goodWine, regularWine, poorWine.* The delivery time of the product is given in terms of days. However, buyer agents evaluate them in terms of five possible outcomes: *days(0,1), days(1,3), days(3,5), days(5,10), days* $(10,\infty)$, where days(x,y) indicates a delivery time between x (inclusive) and y(exclusive) days.

We state a distribution of qualities such a way that the best qualities are scarce. Instead, good delivery times are not rare. The impact of such distributions in the performance of the simulations depends on the desires of the buyer agents, and thus, their importance is subjective. We select the following distributions in this paper:

Quality		**Delivery Time**	
Concept	%	Concept	%
excellentWine	15	*days(0,1)*	30
goodWine	20	*days(1,3)*	25
regularWine	30	*days(3,5)*	15
poorWine	35	*days(5,10)*	15
		days (10,∞)	15

Sellers are completely reactive and always sell the wine under request. We run experiments with different numbers of sellers and informants. But since the distribution of quality products and delivery times is done in percentage, at certain point there is no difference in the performance. Furthermore, unlike previous work [Pinyol et al., 2007a, di Salvatore et al., 2007], sellers are always available.

Informant configurations

Informants are aware of the parameters of the sellers, and can be honest or liars. Honest agents spread accurate reputation information while liars spread wrong reputation. At each turn each informant randomly chooses a seller, and sends a reputation communication to the buyers regarding that seller. If the informant is honest, it will send the exact parameters of the seller. Else, the informant will send reputation values with different parameters of the chosen seller. In the simulations we present in this paper, the parameters are modified to the value of maximum difference. The following table shows the transformation that liar informants do when sending wrong reputation.

Quality		**Delivery Time**	
Real	Send	Real	Send
excellentWine	*poorWine*	*days(0,1)*	*days(10,∞)*
goodWine	*poorWine*	*days(1,3)*	*days(10,∞)*
regularWine	*excellentWine*	*days(3,5)*	*days(0,1)*
poorWine	*excellentWine*	*days(5,10)*	*days(0,1)*
		days (10,∞)	*days(0,1)*

The number of informants also is a parameter of the simulation although we fix it to 5. Then, at each turn the buyer agent receives 5 communicated reputations and only performs one direct experience, simulating the fact that reputation information is more present than image information. In these simulations we do not consider image communications. Therefore, image information is only calculated through direct experience. At each turn one direct experience is contrasted with N reputation communications from the informants (where $N > 1$). The increment in the number of informants increases the effects shown in the following experiments, but shows the same pattern of behavior. Instead, we play with the percentage of honest and liar agents, which directly impact the metareasoning process.

Evaluation of Buyers

We evaluate the performance of buyers at each turn by considering the maximum grade of the positive desires achieved (if any) and subtracting the grades of the negative desires. See the following example.

Example : Let us assume that our agent buyer i wants to get a very good quality product. However, she also would accept a product with less quality but delivered in less than one day. What she does not want at all is a very bad quality product and a delivery time higher than 5 days. These desires can be modeled as follows:

$(D_i^+ excellentWine, 1)$
$(D_i^+ goodWine \wedge dTime < 1, 0.85)$
$(D_i^+ goodWine \wedge dTime < 3, 0.65)$
$(D_i^+ goodWine \wedge dTime < 5, 0.55)$
$(D_i^+ goodWine, 0.45)$
$(D_i^- poorWine, 1)$
$(D_i^- 3 < dTime \leq 5, 0.6)$
$(D_i^- 5 < dTime, 0.8)$

In the following table we exemplify the performance evaluation given some fulfillments:

Fulfillment	**Positive**	**Negative**	**Eval.**
$excellentWine$ $dTime = 6$	$(D_i^+ excellentWine, 1)$	$(D_i^- 5 < dTime, 0.8)$	0.2
$goodWine$ $dTime = 2$	$(D_i^+ goodWine \wedge dTime \leq 3, 0.65)$ $(D_i^+ goodWine \wedge dTime \leq 5, 0.55)$ $(D_i^+ goodWine, 0.45)$	-	0.65
$goodWine$ $dTime = 5$	$(D_i^+ goodWine, 0.45)$	$(D_i^- 3 < dTime \leq 5, 0.6)$	-0.15
$poorWine$ $dTime = 4$	-	$(D_i^- poorWine, 1)$ $(D_i^- 3 < dTime \leq 5, 0.6)$	-1.6

The maximum possible performance is the maximum grade of the positive desires, while the minimum could be as low as the sum of all negative grades. We could consider other forms of evaluation. However, we consider that this is the most reasonable because goes in tune with the semantics given to the desire context [Casali et al., 2004] and the reasoning process led by the set of desires [Pinyol and Sabater-Mir, 2009a] (see also chapter 5).

In the specification, we are considering the evolution of a single buyer with 10 sellers and 5 informants. We executed 10 times each experiment and consider the average level of satisfaction for each turn [3].

[3] For the implementation we use the JASON platform [Bordini et al., 2007], which offers to logic-based agents (prolog-like) a multiagent communication layer. The source code, together with the exact parameters and the set of desires used to run the experiments can be found at `http://www.iiia.csic.es/~ipinyol/sourceJASSS.zip`

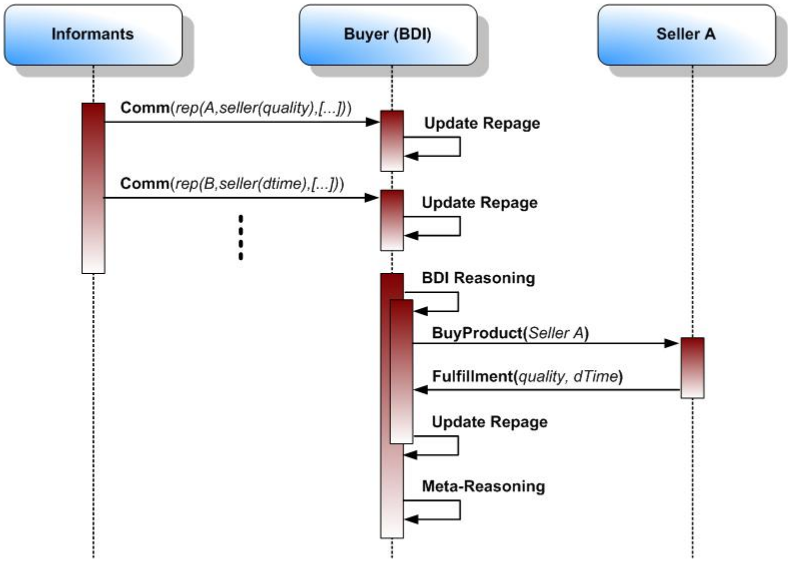

Figure B.3: Sequence diagram of one turn of the simulations.

B.3.2 Experimental Results and Discussion

Static Experiments

It is easy to show the effects of a static Z value in different situations. Figure B.4 shows the accumulated average level of satisfaction obtained by a buyer at each turn in an environment where all informants are honest, and when all informants are liars, considering $Z = 0$ and $Z = 1$. Since when $Z = 0$ reputation information is not taken into account, the performance in this case does not depend on the quality of the reputation information.

The graphic shows that when $Z = 1$, in the case of a scenario with honest informants (0% liars), the level of satisfaction obtained by the agent increases considerable with respect to the case in which $Z = 0$. Assuming normality in the data, from the turn 10, the difference is already statistically significant with a 95% of confidence (p_value$\leq$ 0.05), and from the turn 20 on, the difference becomes significant with a 99% of confidence (p_value$\leq$ 0.01).

Also, when $Z = 1$ and in the scenario all informants spread false reputation, the performance of the buyer decreases considerably with respect to the case in which $Z = 0$. In fact, from the very first turns, the difference becomes already significant with a confidence of 99%.

These results are quite obvious. Since image information is only created from direct experiences (1 at each turn) and reputation information through communicated reputation (5 at each turn) if the communicated information corresponds to the reality and the agent believes what circulates in the society ($Z = 1$) the buyer should discover faster which are the sellers that accomplish her objectives.

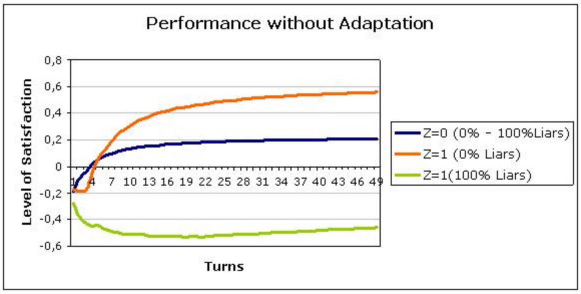

Figure B.4: Level of satisfaction obtained with no adaptation. When Z=0m the percentage of liars does not affect the performance, because no reputation information is considered.

As well, if reputation information if mostly false, and the agent believes it, for a long time the buyer would not be able to fulfill her objectives.

Dynamic Experiments

The main idea behind the updating of Z is that in scenarios where mostly false reputation information circulates Z should tend to 0. On the contrary, scenarios where reputation information is mostly accurate, Z should tend to 1. In this very preliminary paper, we study the effects of an adaptation strategy in the same situations tested in the previous extreme experiments.

The strategy is very simple, but effective. If most of the image information coincide with reputation information (about the same agent/role), the Z value should increase from the current value (in certain proportion). On the contrary, it should decrease. This algorithm contains the parameter *Increment* (inc), which could be also considered as another degree of freedom. For the sake of simplicity we consider it as a constant value.

Figure B.5 and B.6 show the performance obtained in both scenarios. It can be observed how the final performance tends to the theoretical optimum in each situation. In both scenarios there is no statistical significant difference between the performance and the theoretical optimum, with p-values higher than 0.2 with most of the points of the graph.

B.4 Conclusions

In this appendix we propose an heuristics to choose, at run-time, the best h function, which determines the axiom IRB of the L_{BC} theory. Our function h is defined as $h(pE, pS) = (1 - Z) \cdot pE + Z \cdot pS$,where $Z \in [0,1] \cap \mathbb{Q}$. When $Z < 0.5$ function h defines an agent in class $\mathcal{H}_4$ (image is prevalent over reputation). Instead, when $Z > 0.5$ the implemented agent belongs to the class $\mathcal{H}_5$ (reputation is more important than image). The heuristics defined in this appendix computes

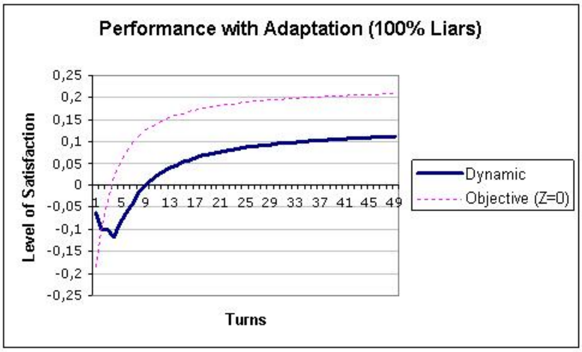

Figure B.5: Level of satisfaction obtained with agents using adaptation in a scenario with 100% of liars. Dot line represents the theoretical best possible performance

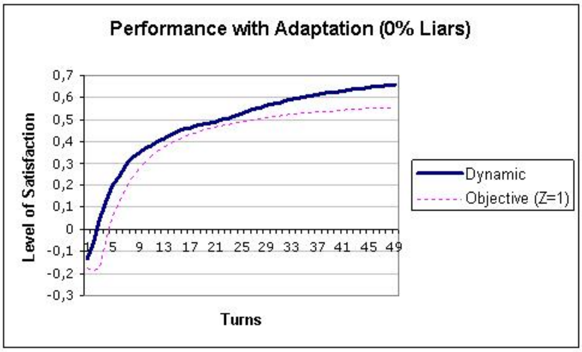

Figure B.6: Level of satisfaction obtained with agents using adaptation in a scenario with no liars . Dot line represents the theoretical best possible performance

the parameter Z at each run-time, in order to maximize the satisfaction level of the agent in its interactions.

Since image information is computed from direct experiences and reputation from communications, the main underlying idea is that when reputation information is similar to image information, Z should increase. This is because we hold the assumption that the amount of reputation information is much higher than the amount of image information, and thus, if it can be estimated that reputation and image information coincide, the use of reputation allows a faster way of discovering *good* partners. Such assumption is justified in [Conte and Paolucci, 2002].

The proposed heuristics is validated though computational simulations. The results show that agents do not archive statistically significant differences in the level of satisfaction obtained compared to the theoretical optimum. We also validate that when the heuristics is not used, agents archive statistically a lower level of satisfaction. This demonstrates the validity of the heuristics.

Bibliography

[Abdul-Rahman and Hailes, 2000] Abdul-Rahman, A. and Hailes, S. (2000). Supporting trust in virtual communities. In *Proceedings of the Hawaii's International Conference on Systems Sciences, Maui, Hawaii.*

[Amazon, 2002] Amazon (2002). *Amazon Auctions.* http://auctions.amazon.com.

[Amgoud et al., 2000] Amgoud, L., Maudet, N., and Parsons, S. (2000). Modelling dialogues using argumentation. In *Proceedings of the Fourth International Conference on MultiAgent Systems*, pages 31–38.

[Artz and Gil, 2007] Artz, D. and Gil, Y. (2007). A survey of trust in computer science and the semantic web. *Web Semantics: Science, Services and Agents on the World Wide Web*, 5(2):58 – 71. Software Engineering and the Semantic Web.

[Balke et al., 2009] Balke, T., Knig, S., and Torsten, E. (2009). A survey on reputation systems for artificial societies. Technical Report 46, Bayreuth University.

[Bentahar et al., 2007] Bentahar, J., Meyer, J. C., and Moulin, B. (2007). Securing agent-oriented systems: An argumentation and reputation-based approach. In *ITNG*, pages 507–515. IEEE Computer Society.

[Bordini et al., 2007] Bordini, R. H., Hbner, J. F., and Wooldridge, M. (2007). *Programming Multi-Agent Systems in AgentSpeak Using Jason.* John Wiley and Sons, Ltd.

[Bromley, 1993] Bromley, D. B. (1993). *Reputation, Image and Impression Management.* John Wiley & Sons.

[Buskens, 1998] Buskens, V. (1998). The social structure of trust. *Social Networks*, (20):265—298.

[Carbo et al., 2002a] Carbo, J., Molina, J., and Davila, J. (2002a). Comparing predictions of sporas vs. a fuzzy reputation agent system. In *3rd International Conference on Fuzzy Sets and Fuzzy Systems, Interlaken*, pages 147—153.

[Carbo et al., 2002b] Carbo, J., Molina, J., and Davila, J. (2002b). Trust management through fuzzy reputation. *Int. Journal in Cooperative Information Systems*, pages in–press.

[Carter et al., 2002] Carter, J., Bitting, E., and Ghorbani, A. (2002). Reputation formalization for an information-sharing multi-agent sytem. *Computational Intelligence*, 18(2):515—534.

[Casali, 2008] Casali, A. (2008). *On Intentional and Social Agents with Graded Attitudes*. PhD thesis, Universitat de Girona.

[Casali et al., 2004] Casali, A., Godo, L., and Sierra, C. (2004). Graded models for bdi agents. In Leite, J. and Torroni, P., editors, *CLIMA V, Lisboa, Portugal*, pages 18—33.

[Casali et al., 2008] Casali, A., Godo, L., and Sierra, C. (2008). A logical framework to represent and reason about graded preferences and intentions. In *Proc. of KR'08, Sydney, Australia.*

[Casare and Sichman, 2005] Casare, S. J. and Sichman, J. S. (2005). Towards a functional ontology of reputation. In *Proceedings of the Fourth International Joint Conference on Autonomous Agents and Multi Agent Systems, Utrecht, The Netherlands*, pages 505–511. ACM Press.

[Castelfranchi and Falcone, 1998a] Castelfranchi, C. and Falcone, R. (1998a). Principles of trust for mas: Cognitive anatomy, social importance, and quantification. In *Proceedings of the International Conference on Multi-Agent Systems (ICMAS'98),Paris,France*, pages 72—79.

[Castelfranchi and Falcone, 1998b] Castelfranchi, C. and Falcone, R. (1998b). Social trust. In *Proceedings of the First Workshop on Deception, Fraud and Trust in Agent Societies, Minneapolis, USA*, pages 35—49.

[Castelfranchi and Paglieri, 2007] Castelfranchi, C. and Paglieri, F. (2007). The role of beliefs in goal dynamics: Prolegomena to a constructive theory of intentions. *Synthese*, 155:237–263.

[Celentani et al., 1966] Celentani, M., Fudenberg, D., Levine, D. K., and Psendorfer, W. (1966). Maintaining a reputation against a long-lived opponent. *Econometrica*, 64(3):691—704.

[Centeno et al., 2009a] Centeno, R., Billhardt, H., Hermoso, R., and Ossowski, S. (2009a). Organising mas: A formal model based on organizational mechanisms. In *Proc. of SAC*.

[Centeno et al., 2009b] Centeno, R., da Silva, V. T., and Hermoso, R. (2009b). A reputation model for organisational supply chain formation. In *COIN@AAMAS*.

[Chesevar and Simari, 2007] Chesevar, C. and Simari, G. (2007). Modelling inference in argumentation through labeled deduction: Formalization and logical properties. *Logica Universalis*, 1(1):93—124.

[Conte and Paolucci, 2002] Conte, R. and Paolucci, M. (2002). *Reputation in artificial societies: Social beliefs for social order.* Kluwer Academic Publishers.

[Demolombe and Liau, 2001] Demolombe, R. and Liau, C. (2001). A logic of graded trust and belief fusion. In *Proc. of the 4th Workshop on Deception, Fraud and Trust in Agent Societies*, pages 13–25.

[Demolombe and Lorini, 2008] Demolombe, R. and Lorini, E. (2008). A logical account of trust in information sources. In *Eleventh International Workshop on Trust In Agent Societies.*

[di Salvatore et al., 2007] di Salvatore, A., Pinyol, I., Paolucci, M., and Sabater, J. (2007). Grounding reputation experiments. a replication of a simple market with image exchange. In *Proceedings of the M2M'07, Marseille, France*, pages 32—45.

[Dung, 1995] Dung, P. M. (1995). On the acceptability of arguments and its fundamental role in nonmonotonic reasoning, logic programming and n-person games. *AI*, 77(2):321–358.

[Dunne and Bench-Capon, 2003] Dunne, P. and Bench-Capon, T. (2003). Two party immediate response disputes: Properties and efficiency. *Artificial Intelligence*, 149(2):221 – 250.

[Dunne et al., 2009] Dunne, P., Hunter, A., McBurney, P., Parsons, S., and Wooldridge, M. (2009). Inconsistency tolerance in weighted argument systems. In *Proc. of the AAMAS'09, Budapest, Hungary.*, pages 851–858.

[eBay, 2002] eBay (2002). *eBay.* http://www.eBay.com.

[Enderton, 1972] Enderton, H. B. (1972). *A mathematical introduction to logic.* Academic Press, New York.

[eRep, 2006a] eRep (2006a). *eRep Wiki.* http://megatron.iiia.csic.es/mediawiki.

[eRep, 2006b] eRep (2006b). *eRep:Social Knowledge for e-Governance.* http://megatron.iiia.csic.es/eRep.

[eRep, 2007] eRep (2007). *Deliverable 1.1: Review of Internet User-oriented Reputation Applications and Application Layer Networks.* http://megatron.iiia.csic.es/eRep/?q=node/37.

[Esfandiari and Chandrasekharan, 2001] Esfandiari, B. and Chandrasekharan, S. (2001). On how agents make friends: Mechanisms for trust acquisition. In *Proceedings of the Fourth Workshop on Deception, Fraud and Trust in Agent Societies, Montreal, Canada*, pages 27—34.

[Eymann, 2000] Eymann, T. (2000). *AVALANCHE - an agent-based decentralized coordination mechanism for electronic marketplaces*. PhD thesis, Albert-Ludwigs-Universitt, Freiburg, Germany.

[Fagin and Halpern, 1994] Fagin, R. and Halpern, J. (1994). Reasoning about knowledge and probability. *J. ACM*, 41(2):340–367.

[Flaminio et al., 2008] Flaminio, T., Pinna, G. M., and Tiezzi, E. B. P. (2008). A complete fuzzy logical system to deal with trust management systems. *Fuzzy Sets Syst.*, 159(10):1191–1207.

[ForTrust, 2009] ForTrust (2009). *ForTrust:Social Trust Analysis and Formalization*. http://www.irit.fr/ForTrust/.

[Gaertner et al., 2006] Gaertner, D., Noriega, P., and Sierra, C. (2006). Extending the bdi architecture with commitments. In *Proceedings of the CCIA'06*, pages 247—257.

[Gambetta, 1990] Gambetta, D. (1990). *Trust: Making and Breaking Cooperative Relations*, chapter Can We Trust Trust?, pages 213—237. Basil Blackwell, Oxford.

[Gaudou et al., 2006] Gaudou, B., Herzig, A., and Longin, D. (2006). Grounding and the expression of belief. In *Proceedings of the Tenth International Conference on Principles of Knowledge Representation and Reasoning*, pages 221–229.

[Giunchiglia and Serafini, 1994] Giunchiglia, F. and Serafini, L. (1994). Multilanguage hierarchical logic (or: How we can do without modal logics). *Journal of AI*, 65:29—70.

[Grabner-Kruter and Kaluscha, 2003] Grabner-Kruter, S. and Kaluscha, E. A. (2003). Empirical research in on-line trust: a review and critical assessment. *International Journal of Human-Computer Studies*, 58(6):783 – 812.

[Grandison and Sloman, 2000] Grandison, T. and Sloman, M. (2000). A survey of trust in internet applications. *IEEE Communications Surveys and Tutorials*, 3(4).

[Grant et al., 2000] Grant, J., Kraus, S., and Perlis, D. (2000). A logic for characterizing multiple bounded agents. *Autonomous Agents and Multi-Agent Systems*, 3(4):351–387.

[Hájek et al., 1995] Hájek, P., Godo, L., and Esteva, F. (1995). Fuzzy logic and probability. *Uncertainty in Artificial Intelligence Conference*, pages 237–244.

[Heras et al., 2009] Heras, S., M. Navarro, V. B., and Julian, V. (2009). Applying dialogue games to manage recommendation in social networks. In *ArgMAS 2009*.

[Herzig et al., 2008] Herzig, A., Lorini, E., Hubner, J. F., Ben-Naim, J., Castelfranchi, C., Demolombe, R., Longin, D., and Vercouter, L. (2008). Prolegomena for a logic of trust and reputation. In *NORMAS'08*, pages 143–157.

[Hoffman et al., 2007] Hoffman, K., Zage, D., and Nita-Rotaru, C. (2007). A survey of attack and defense techniques for reputation systems. Technical Report CSD TR 07-013, Purdue University.

[Hume, 1975] Hume, D. (1975). *A Treatise of Human Nature (1737).* Oxford: Clarendon Press.

[Huynh et al., 2006a] Huynh, T., Jennings, N., and Shadbolt, N. (2006a). An integrated trust and reputation model for open multi-agent systems. *Autonomous Agents and Multi-Agent Systems*, 13(2):119–154.

[Huynh et al., 2006b] Huynh, T., Jennings, N., and Shadbolt, N. (2006b). An integrated trust and reputation model for open multi-agent systems. *Journal of AAMAS*, 2(13):119–154.

[Jsang et al., 2007a] Jsang, A., Ismail, R., and Boyd, C. (2007a). A survey of trust and reputation systems for online service provision. *Decision Support Systems*, 43(2):618 – 644. Emerging Issues in Collaborative Commerce.

[Jsang et al., 2007b] Jsang, A., Ismail, R., and Boyd, C. (2007b). A survey of trust and reputation systems for online service provision. *Decision Support Systems*, 43(2):618 – 644. Emerging Issues in Collaborative Commerce.

[Karlins and Abelson, 1970] Karlins, M. and Abelson, H. I. (1970). *Persuasion, how opinion and attitudes are changed.* Crosby Lockwood & Son.

[Kooi, 2003] Kooi, B. P. (2003). Probabilistic dynamic epistemic logic. *J. of Logic, Lang. and Inf.*, 12(4):381–408.

[Koster et al., 2009] Koster, A., Sabater-Mir, J., and Schorlemmer, M. (2009). A formalization of trust alignment. In *12th International Conference of the Catalan Association for Artificial Intelligence*, Cardona, Catalonia, Spain.

[Koutrouli and Tsalgatidou, 2006] Koutrouli, E. and Tsalgatidou, A. (2006). Reputation-based trust systems for p2p applications: Design issues and comparison framework. In *Trust and Privacy in Digital Business*, volume 4083 of *LNCS*, pages 152—161. Springer.

[Kozen, 1983] Kozen, D. (1983). A probabilistic pdl. In *STOC '83: Proceedings of the fifteenth annual ACM symposium on Theory of computing*, pages 291–297, New York, NY, USA. ACM.

[Kuhlen, 1999] Kuhlen, R. (1999). *Die Konsequenzen von Informationsassistenten. Was bedeutet informationelle Autonomie oder wie kann Vertrauen in elektronische Dienste in offenen Informationsmrkten gesichert wer-den?* Suhrkamp Verlag.

[Liau, 2003] Liau, C. J. (2003). Belief, information acquisition, and trust in multi-agent systems: a modal logic formulation. *Artif. Intell.*, 149(1):31–60.

[Lu et al., 2007] Lu, G., Lu, J., Yao, S., and Yip, J. (2007). A review on computational trust models for multi-agent systems. In *International Conference on Internet Computing*, pages 325–331.

[Luck et al., 2005] Luck, M., McBurney, P., Shehory, O., and Willmott, S. (2005). *Agent Technology: Computing as Interaction (A Roadmap for Agent Based Computing)*. AgentLink.

[Luhmann, 1979] Luhmann, N. (1979). *Trust and Power*. Chichester: Wiley.

[Marimon et al., 2000] Marimon, R., Nicolini, J. P., and Teles, P. (2000). Competition and reputation. In *Proceedings of the World Conference Econometric Society, Seattle*.

[Marsh, 1994] Marsh, S. (1994). *Formalising Trust as a Computational Concept*. PhD thesis, Department of Mathematics and Computer Science, University of Stirling.

[Maximilien and Singh, 2002] Maximilien, E. and Singh, M. (2002). Reputation and endorsement for web services. *ACM SIGEcom Exchange*, 3(1):24–31.

[Miceli and Castelfranchi, 2000] Miceli, M. and Castelfranchi, C. (2000). *Human Cognition and Social Agent Technology*, chapter The Role of Evaluation in Cognition and Social Interaction, pages 225–259. John Benjamins.

[Morge, 2008] Morge, M. (2008). An argumentation-based computational model of trust for negotiation. In *AISB'08*, volume 4, pages 31–36. The Society for the Study of Artificial Intelligence and Simulation of Behavior.

[Mui et al., 2002a] Mui, L., Halberstadt, A., and Mohtashemi, M. (2002a). Notions of reputation in multi-agent systems: A review. In *Proceedings of the first international joint conference on autonomous agents and multiagent systems (AAMAS-02), Bologna, Italy*, pages 280—287.

[Mui et al., 2001] Mui, L., Mohtashemi, M., Ang, C., Szolovits, P., and Halberstadt, A. (2001). Ratings in distributed systems: A bayesian approach. In *Proceedings of the 11th Workshop on Information Technologies and Systems (WITS), New Orleans, USA*.

[Mui et al., 2002b] Mui, L., Mohtashemi, M., and Halberstadt, A. (2002b). A computational model for trust and reputation. In *Proceedings of the 35th Hawaii International Conference on System Sciences*.

[Muller and Vercouter, 2005] Muller, G. and Vercouter, L. (2005). Decentralized monitoring of agent communications with a reputation model. In Falcone, R., Barber, K. S., Sabater-Mir, J., and Singh, M. P., editors, *Trusting Agents for Trusting Electronic Societies, Theory and Applications in HCI and E-Commerce*, volume 3577 of *Lecture Notes in Computer Science*. Springer.

[OnSale, 2002] OnSale (2002). *OnSale*. http://www.onsale.com.

[Padovan et al., 2002] Padovan, B., Sackmann, S., Eymann, T., and Pippow, I. (2002). A prototype for an agent-based secure electronic marketplace including reputation-tracking mechanisms. *International Journal of Electronic Commerce*, 6(4):93–113.

[Paolucci et al., 2009] Paolucci, M., Eymann, T., Jager, W., Sabater-Mir, J., Conte, R., Marmo, S., Picascia, S., Quattrociocchi, W., KOnig, S., Balke, T., Broekhuizen, T., Trampe, D., Tuk, M., Brito, I., Pinyol, I., and Villatoro, D. (2009). *Social Knowledge for e-Governance: Theory and Technology of Reputation*. ISTC-CNR.

[Paolucci et al., 2005] Paolucci, M., Sabater-Mir, J., and Conte, R. (2005). "what if?" dealing with uncertainity in repage's mental landscape. pages 372–378.

[Parsons et al., 1998] Parsons, S., Sierra, C., and Jennings, N. (1998). Agents that reason and negotiate by arguing. *Journal of Logic and Computation*, 8(3):261–292.

[Pinyol, 2008] Pinyol, I. (2008). Social evaluations for cognitive agents: Image and reputation in multi-context bdi agents. Master's thesis, Polytechnic University of Catalonia(UPC) - Artificial Intelligence Research Institute (IIIA-CSIC), Barcelona, Spain.

[Pinyol et al., 2007a] Pinyol, I., Paolucci, M., Sabater-Mir, J., and Conte, R. (2007a). Beyond accuracy. reputation for partner selection with lies and retaliation. In *Proceedings of the MABS'07. Hawaii, USA.*, volume 5003 of *LNCS*, pages 128–140. Springer.

[Pinyol and Sabater-Mir, 2007] Pinyol, I. and Sabater-Mir, J. (2007). Arguing about reputation. the lrep language. In *8th Annual International Workshop "Engineering Societies in the Agents World"*, volume 4995 of *LNCS*, pages 284–299. Springer.

[Pinyol and Sabater-Mir, 2008] Pinyol, I. and Sabater-Mir, J. (2008). Cognitive social evaluations for multi-context bdi agents. In *9th Annual International Workshop Engineering Societies in the Agents World.*

[Pinyol and Sabater-Mir, 2009a] Pinyol, I. and Sabater-Mir, J. (2009a). Pragmatic-strategic reputation-based decisions in bdi agents. In *Proc. of the AAMAS'09, Budapest, Hungary.*, pages 1001–1008.

[Pinyol and Sabater-Mir, 2009b] Pinyol, I. and Sabater-Mir, J. (2009b). Towards the definition of an argumentation framework using reputation information. In *Proc. of the TRUST@AAMAS'09 workshop, Budapest, Hungary.*, pages 92–103.

[Pinyol et al., 2007b] Pinyol, I., Sabater-Mir, J., and Cuni, G. (2007b). How to talk about reputation using a common ontology: From definition to implementation. In *Proceedings of the Ninth Workshop on Trust in Agent Societies. Hawaii, USA.*, pages 90—101.

[Pinyol et al., 2008] Pinyol, I., Sabater-Mir, J., and Dellunde, P. (2008). Probabilistic dynamic belief logic for image and reputation. In *Proc. of the CCIA'08, Empuries, Spain.*

[Plato, 1955] Plato (1955). *The Republic (370BC).* Viking Press.

[Prakken, 2000] Prakken, H. (2000). Relating protocols for dynamic dispute with logics for defeasible argumentation. *Synthese*, 127:2001.

[Rao and Georgeff, 1991] Rao, A. S. and Georgeff, M. P. (1991). Modeling rational agents within a BDI-architecture. In Allen, J., Fikes, R., and Sandewall, E., editors, *Proc. of KR'91*, pages 473–484. Morgan Kaufmann publishers Inc.: San Mateo, CA, USA.

[Rasmusson and Janson, 1997] Rasmusson, L. and Janson, A. R. S. (1997). Using agents to secure the internet marketplace. In *Proc. of the Practical Applications of Agents as Multi-Agent Systems.*

[Rasmusson and Janson, 1996] Rasmusson, L. and Janson, S. (1996). Simulated social control for secure internet commerce. In *Proc. of the 1996 Workshop on New Security Paradigms, London, UK*, pages 18–25. ACM Press.

[Regan and Cohen, 2005] Regan, K. and Cohen, R. (2005). Indirect reputation assessment for adaptive buying agents in electronic markets. *Business Agents and the Semantic Web workshop*, 1.

[Ripperger, 1998] Ripperger, T. (1998). *konomik des Vertrauens - Analyse eines Organisationsprinzips.* Tbinger.

[Ruohomaa et al., 2007] Ruohomaa, S., Kutvonen, L., and Koutrouli, E. (2007). Reputation management survey. In *ARES '07: Proceedings of the The Second International Conference on Availability, Reliability and Security*, pages 103–111, Washington, DC, USA. IEEE Computer Society.

[Sabater and Sierra, 2001] Sabater, J. and Sierra, C. (2001). Regret: A reputation model for gregarious societies. In *Proceedings of the Fourth Workshop on Deception, Fraud and Trust in Agent Societies, Montreal, Canada*, pages 61—69.

[Sabater and Sierra, 2002] Sabater, J. and Sierra, C. (2002). Reputation and social network analysis in multi-agent systems. In *Proceedings of AAMAS-02, Bologna, Italy*, pages 475—482.

[Sabater and Sierra, 2005] Sabater, J. and Sierra, C. (2005). Review on computational trust and reputation models. *Artif. Intel. Rev.*, 24(1):33–60.

[Sabater-Mir, 2003] Sabater-Mir, J. (2003). *Trust and Reputation for agent societies*. PhD thesis, IIIA-CSIC, Barcelona, Spain.

[Sabater-Mir and Paolucci, 2007] Sabater-Mir, J. and Paolucci, M. (2007). On representation and aggregation of social evaluations in computational trust and reputation models. *International Journal of Approximate Reasoning*, 46(3):458–483.

[Sabater-Mir et al., 2006] Sabater-Mir, J., Paolucci, M., and Conte, R. (2006). Repage: Reputation and image among limited autonomous partners. *JASSS*, 9(2).

[Sabater-Mir et al., 2002] Sabater-Mir, J., Sierra, C., Parsons, S., and Jennings, N. R. (2002). Engineering executable agents using multi-context systems. *J. Logic and Comp.*, 12(3):413–442.

[Schillo et al., 1999] Schillo, M., Funk, P., and Rovatsos, M. (1999). Who can you trust: Dealing with deception. In *Proceedings of the Second Workshop on Deception, Fraud and Trust in Agent Societies, Seattle, USA*, pages 95—106.

[Schillo et al., 2000] Schillo, M., Funk, P., and Rovatsos, M. (2000). Using trust for detecting deceitful agents in artificial societites. *Applied Artificial Intelligence*, (Special Issue on Trust, Deception and Fraud in Agent Societies).

[Sen and Sajja, 2002a] Sen, S. and Sajja, N. (2002a). Robustness of reputation-based trust: boolean case. In *AAMAS '02: Proceedings of the first international joint conference on Autonomous agents and multiagent systems*, pages 288–293, New York, NY, USA. ACM Press.

[Sen and Sajja, 2002b] Sen, S. and Sajja, N. (2002b). Robustness of reputation-based trust: Boolean case. In *Proceedings of the first international joint conference on autonomous agents and multiagent systems (AAMAS-02), Bologna, Italy*, pages 288—293.

[Shannon, 1948] Shannon, C. (1948). A mathematical theory of communication. *Bell System Thecnical Journal*, 27:379—423.

[Sierra and Debenham, 2005] Sierra, C. and Debenham, J. (2005). An informationbased model for trust. In *AAMAS '05: Proceedings of The 4th International Conference on Autonomous Agents and Multiagent Systems*, pages 497–504. International Foundation for Autonomous Agents and Multiagent Systems.

[Stranders et al., 2008] Stranders, R., de Weerdt, M., and Witteveen, C. (2008). Fuzzy argumentation for trust. In *Proceedings of the CLIMA VIII*, volume 5056 of *LNCS*, pages 214–230. Springer.

[Suryanarayana and Taylo, 2004] Suryanarayana, G. and Taylo, R. N. (2004). A survey of trust management and resource discovery technologies in peer-to-peer applications. ISR Technical Report UCI-ISR-04-6, University of California.

[trustProject, 2000] trustProject (2000). *Trust-building in electronic markets from an intercultural point of view.* http://www.inf-wiss.uni-konstanz.de/FG/IV/TRUST/.

[Tuomela, 1992] Tuomela, R. (1992). Group beliefs. *Synthese*, 91(3):285–318.

[Valente, 1995] Valente, A. (1995). *Legal Knowledge Engineering - A modelling approach.* IOS Press.

[W. Quattrociocchi, 2008a] W. Quattrociocchi, M. P. (2008a). Reputation and uncertainty. a fairly optimistic society when cheating is total. In *First International Conference on Reputation: Theory and Technology (ICORE 09), Gargonza Italy.*, pages 215–226.

[W. Quattrociocchi, 2008b] W. Quattrociocchi, M. Paolucci, R. C. (2008b). Dealing with uncertainty: Simulating reputation in an ideal marketplace. In *Proceedings of the Eleventh Workshop on Trust in Agent Societies. Estoril, Portugal.*

[Walton and Krabbe, 1995] Walton, D. and Krabbe, E. (1995). *Commitement in Dialogue: Basic Concepts of Interpersonal Reasoning.* State University of New York Press.

[Yu and Singh, 2001] Yu, B. and Singh, M. P. (2001). Towards a probabilistic model of distributed reputation management. In *Proceedings of the Fourth Workshop on Deception, Fraud and Trust in Agent Societies, Montreal, Canada*, pages 125—137.

[Yu and Singh, 2003] Yu, B. and Singh, M. P. (2003). Detecting deception in reputation management. In *AAMAS '03: Proceedings of the second international joint conference on Autonomous agents and multiagent systems*, pages 73–80, New York, NY, USA. ACM Press.

[Zacharia, 1999] Zacharia, G. (1999). Collaborative reputation mechanisms for online communities. Master's thesis, Massachusetts Institute of Technology.

Monografies de l'Institut d'Investigació en Intel·ligència Artificial

Num. 1 J. Puyol, *MILORD II: A Language for Knowledge–Based Systems*

Num. 2 J. Levy, *The Calculus of Refinements, a Formal Specification Model Based on Inclusions*

Num. 3 Ll. Vila, *On Temporal Representation and Reasoning in Knowledge–Based Systems*

Num. 4 M. Domingo, *An Expert System Architecture for Identification in Biology*

Num. 5 E. Armengol, *A Framework for Integrating Learning and Problem Solving*

Num. 6 J. Ll. Arcos, *The Noos Representation Language*

Num. 7 J. Larrosa, *Algorithms and Heuristics for Total and Partial Constraint Satisfaction*

Num. 8 P. Noriega, *Agent Mediated Auctions: The Fishmarket Metaphor*

Num. 9 F. Manyà, *Proof Procedures for Multiple-Valued Propositional Logics*

Num. 10 W. M. Schorlemmer, *On Specifying and Reasoning with Special Relations*

Num. 11 M. López-Sánchez, *Approaches to Map Generation by means of Collaborative Autonomous Robots*

Num. 12 D. Robertson, *Pragmatics in the Synthesis of Logic Programs*

Num. 13 P. Faratin, *Automated Service Negotiation between Autonomous Computational Agents*

Num. 14 J. A. Rodríguez, *On the Design and Construction of Agent-mediated Electronis Institutions*

Num. 15 T. Alsinet, *Logic Programming with Fuzzy Unification and Imprecise Constants: Possibilistic Semantics and Automated Deduction*

Num. 16 A. Zapico, *On Axiomatic Foundations for Qualitative Decision Theory - A Possibilistic Approach*

Num. 17 A. Valls, *ClusDM: A multiple criteria decision method for heterogeneous data sets*

Num. 18 D. Busquets, *A Multiagent Approach to Qualitative Navigation in Robotics*

Num. 19 M. Esteva, *Electronic Institutions: from specification to development*

Num. 20 J. Sabater, *Trust and Reputation for Agent Societies*

Num. 21 J. Cerquides, *Improving Algorithms for Learning Bayesian Network Classifiers*

Num. 22 M. Villaret, *On Some Variants of Second-Order Unification*

Num. 23 M. Gómez, *Open, Reusable and Configurable Multi-Agent Systems: A Knowledge Modelling Approach*

Num. 24 S. Ramchurn, *Multi-Agent Negotiation Using Trust and Persuasion*

Num. 25 S. Ontañon, *Ensemble Case-Based Learning for Multi-Agent Systems*

Num. 26 M. Sánchez, *Contributions to Search and Inference Algorithms for CSP and Weighted CSP*

Num. 27 C. Noguera, *Algebraic Study of Axiomatic Extensions of Triangular Norm Based Fuzzy Logics*

Num. 28 E. Marchioni, *Functional Definability Issues in Logics Based on Triangular Norms*

Num. 29 M. Grachten, *Expressivity-Aware Tempo Transformations of Music Performances Using Case Based Reasoning*

Num. 30 I. Brito, *Distributed Constraint Satisfaction*

Num. 31 E. Altamirano, *On Non-clausal Horn-like Satisfiability Problems*

Num. 32 A. Giovannucci, *Computationally Manageable Combinatorial Auctions for Supply Chain Automation*

Num. 33 R. Ros, *Action Selection in Cooperative Robot Soccer using Case-Based Reasoning*

Num. 34 A. García-Cerdaña, *On some Implication-free Fragments of Substructural and Fuzzy Logics*

Num. 35 A. García-Camino, *Normative Regulation of Open Multi-agent Systems*

Num. 36 A. Ramisa Ayats, *Localization and Object Recognition for Mobile Robots*

Num. 37 C.G. Baccigalupo, *Poolcasting: an intelligent technique to customise music programmes for their audience*

Num. 38 J. Planes, *Design and Implementation of Exact MAX-SAT Solvers*

Num. 39 A. Bogdanovych, *Virtual Institutions*

Num. 40 J. Nin, *Contributions to Record Linkage for Disclosure Risk Assessment*

Num. 41 J. Argelich Romà, *Max-SAT Formalisms with Hard and Soft Constraints*

Num. 42 A. Casali, *On Intentional and Social Agents with Graded Attitudes*

Num. 43 A. Perreau de Pinnick Bas, *Decentralised Enforcement in Multiagent Networks*

Num. 44 I. Pinyol, *Milking the Reputation Cow: Argumentation, Reasoning and Cognitive Agents*