Artif Intell Rev (2013) 40:1–25
DOI 10.1007/s10462-011-9277-z

Computational trust and reputation models for open multi-agent systems: a review

Isaac Pinyol · Jordi Sabater-Mir

Published online: 7 July 2011

Abstract In open environments, agents depend on reputation and trust mechanisms to evaluate the behavior of potential partners. 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 multi-agent systems. In this paper we provide a survey that, far from being exhaustive, intends to show the most representative models that currently exist in the literature. For this enterprise we consider several dimensions of analysis that appeared in three existing surveys, and provide new dimensions that can be complementary to the existing ones and that have not been treated directly. Moreover, besides showing the original classification that each one of the surveys provide, we also classify models that where not taken into account by the original surveys. The paper illustrates the proliferation in the past few years of models that follow a more cognitive approach, in which trust and reputation representation as mental attitudes is as important as the final values of trust and reputation. Furthermore, we provide an objective definition of trust, based on Castelfranchi's idea that *trust* implies a decision to rely on someone.

Keywords Computational trust and reputation models · Multiagent systems · Cognitive trust and reputation

I. Pinyol (✉)
ASCAMM Technology Center, Cerdanyola del Vallès,
Barcelona, Spain
e-mail: ipinyol@ascamm.com

I. Pinyol
PlastiaSite S.A., Cerdanyola del Vallès, Barcelona, Spain
e-mail: ipinyol@plastia.com

J. Sabater-Mir
IIIA - Artificial Intelligence Research Institute,
CSIC - Spanish Scientific Research Council, Bellaterra,
Barcelona, Spain
e-mail: jsabater@iiia.csic.es

1 Introduction

The importance of reputation and trust is out of question in both human and virtual societies. The sociologist Luhmann (1979) wrote: *"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 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 coalition with etc. and it has been studied from different perspectives. 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 the fields that most use these concepts is the field of multi-agent systems (MAS). These systems are composed of autonomous agents 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 and 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 depends 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, by 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 mechanisms 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 knowledge on how agents use reputation information to select partners, how agents communicate and spread reputation, and how agents handle communicated reputation information. It is important to remark that these three approaches are complementary and that each one covers a different typology of problems, all related to the control of interactions on 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). In this paper we provide a survey that, far from being exhaustive, intends to show the most representative models that currently exist in the literature. For this enterprise we consider several dimensions of analysis that appeared in three existing surveys, and provide new dimensions that can be complementary to the existing ones and that have not been treated directly. Moreover, besides showing the original classification that each one of the surveys provide, we also classify models that where not taken into account by the original reviews (often because the models are newer than when the survey appeared).

Many surveys exist in the literature and along with them, different dimensions to classify and characterize computation trust and reputation models. Some of them are based on online trust and reputation related systems (Jsang et al. 2007; 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 2006; Lu et al. 2007; Mui et al. 2002).

The first survey we review in Sect. 2.1 is maybe one of the most cited and used one, and was developed by Sabater & Sierra (2005). Their dimensions of analysis were prepared to enhance the properties of the Regret model (Sabater and Sierra 2001) and show basic characteristics of the models. In Sect. 2.2, we explain the dimensions introduced by Balke et al. (2009). The paper defines a 5-stage process model that accordingly to the authors, all trust and reputation models use between transactions, and classifies the models depending on the characteristics of each stage. In Sect.2.3 we present the classification extracted from the European project eRep (2006). They report four dimensions that give a spectrum of the approaches that such models can follow.

In Sect. 3 we introduce four more dimensions to enhance the characteristics of the most recent models. In particular, we provide an objective definition of what we call a *trust* model, a *name* that traditionally has been left to subjective interpretation by each authors. We provide a brief description of each reviewed model in Sects. 4, and 5 we conclude our analysis.

2 Survey of different classification

2.1 Sabater et al.'s classification dimensions

The first classification dimensions we introduce are those defined by Sabater-Mir & Sierra (Sabater and Sierra 2005; Sabater-Mir 2003). 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, we believe that it represents a very good starting point. The dimensions are:

Paradigm Type: Following Sabater and Sierra (2005), 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. The social trust model introduced by Castelfranchi & Falcone (1998) 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 and that do not have any explicit representation of cognitive attitudes to descrive trust and reputation.

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 instead, it can be considered a private and subjective property that each agent build (S). This dimension is one of the most popular and has been used also in multiple surveys, in particular, in the eRep project (eRep 2006), which refines 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 agents' architectures are considered subjective.

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 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 behavior 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 (see Table 1).

Type of Exchanged Information: Sabater-Mir & Sierra (2005) consider 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 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). We include a set of models that where not taken into account in the original review. In particular we include Repage (Sabater-Mir et al. 2006) and the BDI+Repage (Pinyol and Sabater-Mir 2009; Pinyol et al. 2010) model. We classify Repage as a numerical and cognitive model, since it uses elements of both approaches. It also uses both witness information and direct interaction. It is a subjective contextual model that tolerates cheating. The exchanged information can be communicated

Table 1 Legend for the Table 2

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
Type–Model type	T–Trust
	R–Reputation

Source: Sabater and Sierra (2005)

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 paper.

2.2 Balke et al.'s classification

The classification presented by 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 stage and storage stage. The fourth stage refers to the recall for cooperative behavior, concluding with an adaptation or learning of the strategy in the fifth stage. After these five stages, transaction $i + 1$ can be procesed.

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 models can be considered multi-context or single-context, coinciding with the granularity dimension defined by Sabater & Sierra (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 (1990), and more cognitive approaches where a new rate affects 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 the whole mental state

Table 2 Summary of models' characteristics defined by Sabater-Mir & Sierra dimensions

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*

Models in *italic* were not present in the original classification. Legend on Table 1

of the agent. This classification is related to the paradigm dimension defined by Sabater & Sierra (see previous sections).

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 & Sierra, but from our point of view the former 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 or infer trust and reputation. In this stage, the authors 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.

Table 3 Legend for classification of Table 4

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
	L1–Level 1
	L2–Cheating

While this differentiation is very questionable and somehow contrary to our vision of trust 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 (see Table 3).

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 & 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 implicitly, 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 4 shows the summary of the surveyed models classified according to these dimensions.

2.3 eRep project's classification

The European project *eRep: Social Knowledge for e-Governance* (eRep 2006) 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 et al.'s classification (see previous subsection), and (2) the combination of the information source and visibility dimensions of 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

Table 4 Classification of the models developed by Balke et al. (2009)

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*

The models in *italic* were not present in the original work. Legend on Table 3

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, by summing all the scores and showing it with a system of colored stars.

Table 5 Classification of the models according to the eRep project (eRep 2006)

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	✓	–	–	–

The models in *italic* were not present in the original work

This classification mainly focuses on two big dimensions, agent-oriented and external evaluation agencies. This division corresponds to the visibility dimension of Sabater & Sierra, and the storage stage in Balke et al.'s classification.

Summarizing, Table 5 shows the classification of the models reviewed in eRep Project (eRep 2006). 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.

3 Yet another classification

We could not finish the paper 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. For the nature of the work, this classification only faces distributed models, or in terms of the survey explained in Sect. 2.3, agent-oriented approaches. The dimensions we define here are the following:

Table 6 Computational models against our classification dimensions

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	✓	✓	✓	✓

3.1 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 & Sierra (2005).

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 from Castelfranchi & Falcone (1998), occurrent and dispositional trust by Herzig et al. (2008) and pragmatic-strategic decisions pointed out by Conte & Paolucci (2002), we can deduce that *trust* implies a decision. 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 by Conte & Paolucci (2002).

Table 6 summarizes the models that from our definition should be considered trust models. We mark them with '✓'. For instance, the model defined by 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 (2002). Even when this model is usually considered a reputation model, it defines a decision making process that identifies to whom to interact with, and then, fits into 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 presented by Carbo et al. (2002) 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, in which the model only gives support to the creation of image or reputation predicates, not on how to use such information to choose parners for instance.

Finally, we use ∼ 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 by Sabater-Mir & 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.[1] Hence, the goals of the agent are somehow codified in this ontology, and the final trust value obtained is 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 introduced by Huynh et al. (2006) shows a similar situation.

3.2 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 possible (Castelfranchi and Paglieri 2007). In this sense, for the 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 and can be expressed with analititcal formulas.

In Table 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

[1] 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%.

model with '∼' 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 uses into first-order-like predicates, mixing also epistemic and motivational attitudes. The BDI + Repage model (Pinyol et al. 2010; Pinyol and Sabater-Mir 2009) makes explicit these missing cognitive components.

3.3 Procedural dimension

Often, models offer a nice way to represent and deal with trust and reputation, but there is no explanations 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. However, 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 (1998) 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. 2002) and Regret (Sabater and Sierra 2001; Sabater-Mir 2003) describes until the last detail how evaluations are created and how they are aggregated.

We point out here that the models by Marsh (1994) and Abdul-Rahman & Hailes (Abdul-Rahman and Hailes 2000) are marked with '∼' to indicate that in general they provide all the calculations, but left some initial values. For instance, the former 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.

3.4 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 & Hailes (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 (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 performances.

On the contrary, models built for general purposes can be adapted to multiple scenarios and are perfect then for general agents architectures. Regret (Sabater and Sierra 2001; Sabater-Mir 2003) and BDI + Repage model (Pinyol et al. 2010; Pinyol and Sabater-Mir 2009) are good examples of such models. Table 6 summarizes in the last column this property against the surveyed models.

4 Brief description of the reviewed models

4.1 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 have described in our work.

4.1.1 Online reputation models

These models are used in e-commerce sites such as eBay (2002), Amazon (2002) and 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 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 have a very intuitive understanding, making them ideal for human-based application. However, they lack in robustness; no reliability measures, no consideration of false information or cheating, no temporal issues. For instance, feedbacks remain countable forever. So, a seller with very high reputation value could start acting as a bad seller without having an immediate effect on its reputation value. In general, they lack of the main characteristics that make special each one of the following models.

4.1.2 Sporas and Histos

Sporas and Histos introduced by 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 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 Schillo et al. (2000; 1999).

4.1.3 Carter et al.

The underlying idea of the model introduced by 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 argue 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.

4.1.4 Kuhlen

The model presented by Kuhlen (1999)[2] 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

[2] We extracted the explanation of this model from eRep (2007) and ForTrust (2009) because the original article is in Germany.

measures. It has not been applied to multi-agent systems yet, but the idea should work as well.

4.1.5 Dirichlet reputation systems

This family of reputation systems (see Josang et al. 2007) 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 have been proved to be a good option.

4.2 Agent-oriented approaches

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

4.2.1 Abdul-Rahman and Hailes

The Abdul-Rahman & Hailes 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 the case of a 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).

4.2.2 AFRAS

The model presented by Carbo et al. (2002) 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 a 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 than direct interactions.

4.2.3 Castelfranchi & Falcone

The model by Castelfranchi & Falcone (1998) was the first cognitive trust model. According to the authors, trust is a mental 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:[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. The novelty of this approach was that for the first time a model described the internal components that make agents *trust* other agents.

4.2.4 ForTrust

ForTrust model proposed by Herzig et al. (ForTrust 2009) is a refinement of the previous one. The authors propose a cognitive model that differentiates *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 & Falcone (1998). 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.

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 by Castelfranchi & Falcone, the components embrace an occurrent goal, an occurrent capability belief, an occurrent power belief and an occurrent intention belief. More formally:

[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.

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

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 they move from *occurrent* goals to *potential* goals, and from *occurrent* beliefs to *potential* beliefs. Dispositional trust is defined as follows:

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

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.

Dispositional trust permits the authors to define reputation in terms of its internal components. The authors argue that the notion of reputation described by 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:

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

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 $i, j \in G$,

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

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 (2002), since accepting a group belief does not mean to accept the individual belief.

[4] Formally, $CommonBel_{\{i,j\}}\varphi =_{def} Bel_i\varphi \wedge Bel_j\varphi \wedge Bel_i Bel j\varphi \wedge Bel_j Bel_i\varphi \wedge \cdots$

4.2.5 Esfandiari et al.

Esfandiari & Chandrasekharan (2001) defined a model where trust evaluation is considered in 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.

4.2.6 ReGreT

The ReGreT system presented by Sabater-Mir & 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.

4.2.7 FIRE

The FIRE model introduced by Huynh et al. (2006) 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 except that FIRE considers that agents are honest when communication information) 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*. 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.

4.2.8 Marsh

The Marsh's model (Marsh 1994), one of the first computational models that appeared in literature, talks explicitly about trust, and only considers 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:

$$T_x(y, \alpha)^t = U_x(\alpha)^t \cdot I_x(\alpha)^t \cdot \overline{T_x(y)^t} \tag{1}$$

  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.

4.2.9 Yu and Singh

The model introduced by Yu and Singh (2003), 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. (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.

4.2.10 Mui et al.

Mui et al.'s model (Mui et al. 2002, 2001) suggests a similar approach as the one proposed by Yu & 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.

4.2.11 LIAR

The LIAR model presented by Muller & 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 and check the trust of agents as information sources. Because of that, it has been classified as a context-dependent in our classifications. For the same reason, when we talk about direct experience-based and observation-based reputation, the model refers to the quality of the exchanges information.

4.2.12 Padovan et al.

The model introduced by Padovan et al. (2002) uses a combination of agent oriented approaches and external approaches. The model is designed over 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. Agents use such information (which we consider witness information) together with the history of interactions (if any) of potential partners to compute a *reputation coefficient*, from 0 (bad reputation) to 1 (good reputation). The authors define it as the expected cooperative behavior of potential partners. Reputation is seen by the agents as a subjective property.

4.2.13 Regan & Cohen

Regan & 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.

4.2.14 Sierra & Debenham

Sierra & 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 accordingly to α has β about proposition δ. So, the model accept witness information.
- The power that β has in the social group. The model incorporates sociological information, similar to the Regret model (Sabater & Sierra 2001).

– 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) 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 *direct trust* in the Regret model.

4.2.15 Schillo et al.

The model presented by Schillo et al. (2000; 1999) was designed for societies or environments where the evaluation of interactions between agents has a boolean nature. 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 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.

4.2.16 Ripperger

In the book by Ripperger (1998)[5] the author describes from a pure economical perspective the *creation* of trust as a mechanism to stabilize uncertain expectations when choosing actions, from the point of view of the trustee and the trustor. It models trust relationships based on the trade-off between effective transactions and cost. Thus, the author considers trust as expectations, and use economic theories to calculate it, developing detailed decision making processes for the partner selection. Since the approach is far away from the multi-agent paradigm that we are considering, it does not consider communication between partners nor most of the normal features that would be expected for a trust model. However, it is definitively an alternative model that in some environments could be applied.

4.2.17 Rasmusson & Janson

Rasmusson & Janson (1996; 1997) propose a mechanism similar to Padovan et al. (2002) explained above, with the introduction of special agents as trusted third parties or reviewer agents. The authors' main focus is online marketplaces although their results can be applied to open multi-agent systems in general. 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.

[5] The description of this model is based on the one found in eRep (2007), because it only exists a Germany edition of the book.

4.2.18 Sen and Sajja

In the model presented by Sen and Sajja (2002) 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 that 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, in a linear function.

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.

4.2.19 Repage

Repage (Sabater-Mir et al. 2006) is a computational model that gives support to agent's architectures that want to distinguish between image and reputation. While both objects are social evaluations, *image* is a simple evaluative belief that tells whether a given target agent is *good* or *bad* with respect to a certain context (roles). Instead, reputation is a meta-belief that tells what other people *think* about a given target in a given context. Thus, an agent A may have a very good image of agent B as a *car driver* and at the same time acknowledge that B has a bad reputation as a *car driver*.

Repage organizes social evaluations (represented as firt-order-like predicates) in different levels of abstraction and inter-connected, locating at the top of the hierarchy, image and reputation predicates. Each predicate that belongs to one of the main types (image, reputation, shared voice, shared evaluation) 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 of the 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). The authors call this representation a *weighted labeled tuple* and models a probability distribution.

The model calculates image and reputation through aggregations from elementary predicates, such like contracts and fulfillments (which implement direct interactions) and third-party communications. In fact, direct experiences plus communicated images contribute to the creation of image predicates, while only communication reputations contribute to the inference of reputation. Then, the distinction between image and reputation is always present. Nevertheless, the authors do not define how such information is used in the decision making.

4.2.20 BDI + Repage

The BDI + Repage model (Pinyol et al. 2010) integrates the Repage model explained above within a multi-valued BDI architecture, providing a logic-based framework to perform practical reasoning using image and reputation information. It is specified as a multi-context system, allowing several distinct theoretical components to be specified with the mechanisms that link them together by 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 used to infer information from one context to another. Each bridge rule has a set of antecedents (preconditions) and a consequent (postcondition). The consequent is a formula that becomes true in the target context when the antecedents hold in their respective contexts.

The BDI + Repage model incorporates one context for each attitude (Belief, Desire and Intention), one for the Repage model, and two functional contexts (Communication and Planner). The **Belief context** (BC) includes the believed knowledge of the agent. The deductive system in this context is a probabilistic dynamic belief logic in which formulas like $(B_i[\alpha]\varphi, p)$ indicate that agent i believes that after the execution of action α, the probability that formula φ holds is p. The **Desire context** (DC) is defined as a logic of preferences based on Lukasiewicz logic. Theories in this logic determine the desires of the agent and have the form $(D_i^+\varphi, d)$ and $(D_i^-\varphi, d)$ $(d \in [0, 1] \cap \mathbb{Q})$ meaning that agent i will have a level of satisfaction (the former) or disgust (the latter) d, if φ holds. Finally, the **Intention Context** (IC) holds formulas like $(I_i\varphi, d)$, where $d \in [0, 1] \cap \mathbb{Q}$. Intentions determine the trade-off between positive and negative countereffects of trying to achieve φ. Through the appropriate bridge rules the image and reputation from Repage are introduced as beliefs in the into the BC.

The authors do not talk about *trust* in their approach. However, the architecture performs a practical reasoning to determine to whom to interact with, becoming a trust model according to our definition. Furthermore, when a decision is made, the mental state of the agent can be described in terms of beliefs, desires and intentions. Thus, it can be seen as a cognitive model that completely specifies how the components are constructed.

5 Conclusions

After the analysis performed in this paper, there are several things that we can conclude.

The first one is that the interest in trust and reputation models has not decreased. It can be seen how the amount of different models that currently exist in the literature keeps increasing. It is remarkable though 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, as a set of cognitive attitudes. The reason seems obvious: when designing trust and reputation models, game theoretical approaches work perfectly in simple environments. Nevertheless, if we want to undertake problems found in socially complex virtual societies, like negotiation issues, more sophisticated models based on solid cognitive theories are needed. One of the advantages of cognitive-based modeling is that the structure of the mental state can be as important as the final value. Thus, processes like argumentation, automated negotiation etc. can take place. Another advantage is the proximity with human comprehension. For a human being, it is easier to understand an explanation based on beliefs, desires and intentions than an explanation full of numbers.

Obviously, there are drawbacks in cognitive modeling. Because of their complexity, some of the models remain at a descriptive level. Table 6 shows that the only cognitive model that contemplates procedural aspects is the BDI + Repage model (Pinyol et al. 2010). We believe that this model summarizes one of most prominent future research line in trust and reputation models: *implementable cognitive models*.

The paper also introduces an objective definition of trust model that allows us to classify models independently of the subjective interpretation of each author. The definition is not new, but based on Castelfranchi and Falcone's idea that there is no trust without delegation. Thus, trust implies the decision to rely in someone. Table 6 shows the classification of models based on this definition, and we classify the models with ✓ only if there exist well-defined decision making procedure that permits to select the *best* partner according to the current golas of the agent.

Acknowledgments This work was supported by the EC by the project LiquidPub (STREP FP7-213360), by the Spanish Education and Science Ministry with the projects AEI (TIN2006-15662-C02-01), AT (CONSOLIDER CSD2007-0022, INGENIO 2010) and RepBDI (Intramural 200850I136), and by the Generalitat de Catalunya under the grants 2009-SGR-1433 and 2009-SGR-1434.

References

Abdul-Rahman A, Hailes S (2000) Supporting trust in virtual communities. In: Proceedings of the Hawaii's international conference on systems sciences. Maui, Hawaii

Amazon (2002) Amazon auctions. http://auctions.amazon.com

Artz D, Gil Y (2007) A survey of trust in computer science and the semantic web. Web Semant Sci Serv Agents World Wide Web, Softw Eng Semant Web 5(2):58–71

Balke T, Knig S, Torsten E (2009) A survey on reputation systems for artificial societies. Technical Report 46, Bayreuth University

Carbo J, Molina J, Davila J (2002) Trust management through fuzzy reputation. Int J Co-op Inf Syst pp (in press)

Carter J, Bitting E, Ghorbani A (2002) Reputation formalization for an information-sharing multi-agent sytem. Comput Intell 18(2):515–534

Castelfranchi C, Falcone R (1998) Social trust. In: Proceedings of the first workshop on deception, fraud and trust in agent societies. Minneapolis, USA pp 35–49

Castelfranchi C, Paglieri F (2007) The role of beliefs in goal dynamics: prolegomena to a constructive theory of intentions. Synthese 155:237–263

Conte R, Paolucci M (2002) Reputation in artificial societies: social beliefs for social order. Kluwer, Dordrecht

eBay (2002) eBay. http://www.eBay.com

eRep (2006) eRep: social knowledge for e-governance. http://megatron.iiia.csic.es/eRep

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 B, 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, pp 27–34

ForTrust (2009) ForTrust: social trust analysis and formalization. http://www.irit.fr/ForTrust/

Gambetta D (1990) Trust: making and breaking cooperative relations, chapter can we trust trust?. Basil Blackwell, Oxford 213–237

Grabner-Kruter S, Kaluscha EA (2003) Empirical research in on-line trust: a review and critical assessment. Int J Hum-Comput Stud 58(6):783–812

Grandison T, Sloman M (2000) A survey of trust in internet applications. IEEE Commun Surv Tutor 3(4)

Herzig A, Lorini E, Hubner JF, Ben-Naim J, Castelfranchi C, Demolombe R, Longin D, Vercouter L (2008) Prolegomena for a logic of trust and reputation. In: NORMAS'08. pp 143–157

Hoffman K, Zage D, Nita-Rotaru C (2007) A survey of attack and defense techniques for reputation systems. Technical Report CSD TR 07-013, Purdue University

Huynh T, Jennings N, Shadbolt N (2006) An integrated trust and reputation model for open multi-agent systems. J AAMAS 2(13):119–154

Jsang A, Ismail R, Boyd C (2007) A survey of trust and reputation systems for online service provision. Decis Support Syst 43(2):618–644 Emerging Issues in Collaborative Commerce

Koutrouli E, Tsalgatidou A (2006) Reputation-based trust systems for p2p applications: design issues and comparison framework. In: Trust and privacy in digital business, vol 4083 of LNCS. Springer, Berlin pp 152–161

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, Frankfurt

Lu G, Lu J, Yao S, Yip J (2007) A review on computational trust models for multi-agent systems. In: International conference on internet computing. pp 325–331

Luck M, McBurney P, Shehory O, Willmott S (2005) Agent technology: computing as interaction (a roadmap for agent based computing), AgentLink

Luhmann N (1979) Trust and Power. Wiley, Chichester

Marsh S (1994) Formalising trust as a computational concept. PhD thesis, Department of Mathematics and Computer Science, University of Stirling

Mui L, Halberstadt A, Mohtashemi M (2002) 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, pp 280–287

Mui L, Mohtashemi M, Ang C, Szolovits P, 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 L, Mohtashemi M, Halberstadt A (2002) A computational model for trust and reputation. In: Proceedings of the 35th Hawaii international conference on system sciences

Muller G, Vercouter L (2005) Decentralized monitoring of agent communications with a reputation model. In: Falcone R, Barber KS, Sabater-Mir J, Singh MP (eds) Trusting agents for trusting electronic societies, theory and applications in HCI and e-commerce, volume 3577 of lecture notes in computer science. Springer, Berlin

OnSale (2002) OnSale. http://www.onsale.com

Padovan B, Sackmann S, Eymann T, Pippow I (2002) A prototype for an agent-based secure electronic marketplace including reputation-tracking mechanisms. Int J Electron Commer 6(4):93–113

Pinyol I, Sabater-Mir J (2009) Pragmatic-strategic reputation-based decisions in bdi agents. In: Proceedings of the AAMAS'09, Budapest, Hungary, pp 1001–1008

Pinyol I, Sabater-Mir J, Dellunde P, Paolucci M (2010) Reputation-based decisions for logic-based cognitive agents. Auton Agents Multi-Agent Syst pp 1–42

Rasmusson L, Janson ARS (1997) Using agents to secure the internet marketplace. In: Proceedings of the practical applications of agents as multi-agent systems

Rasmusson L, Janson S (1996) Simulated social control for secure internet commerce. In: Proceedings of the 1996 workshop on new security paradigms. ACM Press, London, UK, pp 18–25

Regan K, Cohen R (2005) Indirect reputation assessment for adaptive buying agents in electronic markets. Business Agents and the Semantic Web workshop, 1

Ripperger T (1998) konomik des Vertrauens–Analyse eines Organisationsprinzips Tbinger

Ruohomaa S, Kutvonen L, Koutrouli E (2007) Reputation management survey. In: ARES '07: proceedings of the the second international conference on availability, reliability and security, IEEE Computer Society. Washington, pp 103–111

Sabater J, 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, pp 61–69

Sabater J, Sierra C (2002) Reputation and social network analysis in multi-agent systems. In: Proceedings of AAMAS-02. Bologna, pp 475–482

Sabater J, Sierra C (2005) Review on computational trust and reputation models. Artif Intel Rev 24(1):33–60

Sabater-Mir J (2003) Trust and reputation for agent societies. PhD thesis, IIIA-CSIC, Barcelona, Spain

Sabater-Mir J, Paolucci M, Conte R (2006) Repage: reputation and image among limited autonomous partners. JASSS 9(2)

Schillo M, Funk P, 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, pp 95–106

Schillo M, Funk P, 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 S, Sajja N (2002) Robustness of reputation-based trust: Boolean case. In: Proceedings of the first international joint conference on autonomous agents and multiagent systems (AAMAS-02). Bologna, pp 288–293

Sierra C, Debenham J (2005) An information-based model for trust. In: Proceedings of the fourth international joint conference on autonomous agents and multiagent systems, AAMAS '05. ACM, New York, pp 497–504

Suryanarayana G, Taylo RN (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

Yu B, Singh MP (2003) Detecting deception in reputation management. In: AAMAS '03: proceedings of the second international joint conference on autonomous agents and multiagent systems. ACM Press, New York, pp 73–80
Zacharia G (1999) Collaborative reputation mechanisms for online communities. Master's thesis, Massachusetts Institute of Technology, September 1999