Chapter 9

Trust and Reputation in Multiagent Systems

Dr. Jordi Sabater-Mir

III A – Artificial Intelligence Research Institute
CSIC – Spanish National Research Council

CSIC

Dr. Laurent Vercouter

LITIS Laboratory
INSA de Rouen

MULTIAGENT SYSTEMS
http://www.the-mas-book.info
From a local perspective, they are integrated into an agent decision process when it involves other agents in order to decide with whom to interact.

From a global perspective, they can be used as social control mechanisms.
2 - Computational representation of trust and reputation values
Exist many different ways to represent trust and reputation values

- **Simplicity**
  - Facilitates the calculation functions and the reasoning mechanisms.
  - Less information, the kind of reasoning that can be done is less sophisticated.

- **Expressiveness**
  - Allows elaborated reasoning mechanisms and sophisticated models.
  - More computational and storage capacity as well as complex reasoning algorithms.
• **Boolean**

True -> the trustee is trustworthy
False -> the trustee is untrustworthy

Not very useful because Trust (like reputation) is a notion eminently graded and therefore it is important to be able to express how much do you trust.

• **Numerical values**

Real or integer values in a range. (ex. [-1.0,1.0], [0,3000])

*Examples:*

- The trust in an agent X is 0.4
- The reputation of agent Y is -1

The most used representation by far.
• Qualitative labels

Finite sets of labels in an ordered set.

*Examples:*

\{very_bad, bad, neutral, good, very_good\}

Is a trust of 0.6 really different from a trust of 0.7 in terms of taking trust decisions?

These sets are mapped to integer numbers so in fact it is a way of reducing the number of output values to simplify the decision making process.

The loss of a fine grain comparison of trust and reputation values is compensated by a universally recognized semantics
- **Probability distribution**

Discrete probability distribution over a sorted discrete set.

*Examples:*

a) With a probability of 0.75 the behaviour of the agent will be very bad, with a probability of 0.25 it will be bad.

b) Bipolar agent, very bad or very good, never in the middle.

c) Unpredictable agent.
• **Fuzzy sets**

The reputation value is a fuzzy set over a range. The linguistic modifiers affect the fuzzy set to express the degree of precisison of the reputation value.

*Example:*

![Graph showing fuzzy sets with different linguistic modifiers](image)

The reliability of reputation is implicitly represented in the shape of the fuzzy set.
• Trust and reputation as beliefs

In a BDI architecture, the trust and reputation values should be represented in terms of beliefs.

Using beliefs to represent trust or reputation raises two main issues:

1. To define the content and the semantics of the specific belief.

   Example: Take the socio-cognitive theory proposed by Castelfranchi and Falcone claiming that "an agent $i$ trusts another agent $j$ in order to do an action $\alpha$ with respect to a goal $\phi$"

   Trust is about an agent and has to be relative to a given action and a given goal.

   ForTrust model. Definition of a specific predicate $\text{OccTrust}(i, j, \alpha, \phi)$ holding for specific instances of a trustor ($i$), a trustee ($j$), an action ($\alpha$) and a goal ($\phi$). The $\text{OccTrust}(i,j,\alpha,\phi)$ predicate is used to represent the concept of occurrent trust that refers to a trust belief holding here and now.
Trust and reputation as beliefs

In a BDI architecture, the trust and reputation values should be represented in terms of beliefs.

Using beliefs to represent trust or reputation raises two main issues:
2. To link the belief to the aggregated data grounding it

Example: In BDI+RepAge the link consists in transforming each one of the probability values of the probability distribution used in RepAge into a belief.
• The reliability of a value

To which extent do we have to take into account a trust or reputation value in order to take a decision?

Are the foundations of that value strong enough to base a decision on it?

Some models add a measure of the reliability that the trust or reputation value has.

*Examples:*

  Associate a number to the trust or reputation value that reflects how reliable it is (ex. ReGreT).

  The widthness of the fuzzy set reflects the reliability of the value (ex. AFRAS).
3 – Trust processes in multiagent systems
Trust evaluation or trust decision

A dual nature of trust:

• Trust as an **evaluation**
  – « *Trust is the subjective probability by which an individual, A, expects that another individual, B, performs a given action on which its welfare depends* » [Gambetta, 88]
  – e.g.: I trust that my medical doctor is a good surgeon

• Trust as an **act**
  – Trust is also the « *decision and the act of relying on, counting on, depending on [the trustee]* » [Castelfranchi & Falcone, 10]
  – E.g.: I decide that my medical doctor will perform a surgery on me
General overview of trust processes

• Trust evaluation
  – A trustor X uses various information sources to decide if a trustee Y is trustworthy
  – It consists in a set of social evaluations (either images or reputations)

• Trust decision
  – A trustor X decides if a trustee Y can be relied on for a given task
  – It is a decision process taking into account trust evaluations
Trust processes

- Direct experiences
- Communicated experiences
- Social information
- Images
- Reputations
- Context
- Self motivations

Trust evaluations

Trust decisions

Actions

3 – Trust processes in multiagent systems
Trust evaluations (1)

Inputs coming from different sources

- **Direct experiences**
  - Direct interactions between the trustor and the trustee

- **Communicated experiences**
  - Interactions between the trustee and another agent communicated to the trustor

- **Social information**
  - Social relations and position of the trustee in the society
Trust evaluations (2)

Inputs need to be filtered or adapted for image building to

- ... consider only relevant inputs for the context of an image
  
  e.g.: if I’m building an image of a medical doctor as a surgeon, I won’t consider her past experiences as a wine recommender

- ... avoid using fake communicated experiences sent by malicious agents

  e.g.: if I detected that an agent sends false communicated experiences about others, I should ignore them

- ... adjust the communicated values if subjective trust computation functions exist

  e.g.: Alice is more severe than Bob and when she communicates a trust value of X, Bob should interpret it as X+2

3 – Trust processes in multiagent systems
Trust evaluation
by a statistical evaluation

**Approach:** Compute a single value from a set of input

- One example with qualitative values [Abdul-Rahman & Hailes, 00]
  - feedback values in the set {very good, good, bad, very bad}
  - aggregation function consists = keeping the most represented feedback about agent a in a context c
  \[ T(a,c,td) \text{ with } td \text{ is defined in} \]
  \{very trustworthy, trustworthy, untrustworthy, very untrustworthy\}

- Another example with numerical values [Schillo et al, 00]
  - Trustor i had n experiences with the trustee j, in which p were positive
  - Aggregation function = a percentage of positive experiences
  \[ T^i_j = p/n \]

- A third example is to keep all the experiences in a probability distribution [Sierra & Debenhram, 00]
Trust evaluation
by logical beliefs generation

**Approach:** Infer a trust evaluation from a set of beliefs

- Example from [Herzig et al, 10], « dispositional trust »:
  
  \[
  \begin{align*}
  \text{DispTrust}(\text{Alice, Bob, write}(p), \text{written}(p), \text{Done}(\text{request}(	ext{Alice, Bob, write}(p)))) & = \\
  \text{PotGoal}_{\text{Alice}}(\text{written}(p), \text{request}(	ext{Alice, Bob, write}(p))) \land \\
  \text{Bel}_{\text{Alice}}(\text{G^*}(\text{request}(	ext{Alice, Bob, write}(p)) \land \text{Choice}_{\text{Alice}}(\text{written}(p)) \rightarrow \\
  \text{Intends}_{\text{Bob}}(\text{write}(p)) \land \text{Capable}_{\text{Bob}}(\text{write}(p)) \land \text{After}_{\text{Bob:write}(p)}(\text{written}(p)))
  \end{align*}
  \]

*Informally:* Alice trusts Bob to write a paper $p$ if
- she may have the goal to have a paper $p$ written and,
- she believes that when she has this goal and when she asked Bob to write the paper
  - Bob intends to write the paper
  - Bob is capable of writing the paper
  - After Bob does the action write($p$) the paper is written
Trust decision

Trust as an act

• The trust decision process takes into account
  – trust evaluations (images and reputations)
  – the context of the decision
  – the motivations of the trustor

• The trust decision process depends on the representation formalism of trust evaluations
Trust value thresholds

Decision relying on thresholds

- If $\Theta_{\text{trust}} \neq \Theta_{\text{distrust}}$, uncertainty in the decision should be handled.
- The trust thresholds can be directly adjusted:
  - with higher values if the trustor’s motivations are important or the context risky
  - with lower values in opposite cases
Trust decision as a belief

• Example from [Herzig et al, 10], « occurrent trust » :

\[
\text{DispTrust(Alice, Bob, write(p), written(p), Done(request(Alice, Bob, write(p))))} \land \\
\text{Choice}_{\text{Alice}} F \text{written(p)} \land \\
\text{Bel}_{\text{Alice}}(\text{request(Alice, Bob, write(p)))} \\
\text{ -> OccTrust(Alice, Bob, write(p), written(p))}
\]

\textit{Alice trusts here and now Bob to write a paper p in order to achieve the goal of having the paper p written}
Diversity of trust models

A current challenge is to propose solutions for T&R interoperability in 3 situations

Value domain

\[ T(John, \text{« good »}) \]

\[ T(John, \{(0.3,0.7), (0.4,0.25), (0.7,0.05)\}) \]

Semantics of values

\[ T(John, 0.6) \]

\[ T(John, 0.6) \]

I’m happy with John

I can do better

Semantics of concepts

\[ \text{SharedEvaluation}(John, \text{seller}) = (0.9,1) \]

\[ \text{DirectTrust}(John, \text{seller}, -0.4) \]
4 - Reputation in multiagent societies
"After death, a tiger leaves behind his skin, a man his reputation"

Vietnamese proverb
Reputation

“What a social entity says about a target regarding his/her behavior”

- The social evaluation linked to the reputation is not necessarily a belief of the issuer.
- Reputation cannot exist without communication.

Set of individuals plus a set of social relations among these individuals or properties that identify them as a group in front of its own members and the society at large.

It is always associated to a specific behaviour/property
What is reputation good for?

• Reputation is one of the elements that allow us to build trust.

• Reputation has also a social dimension. It is not only useful for the individual but also for the society as a mechanism for social order.
4 - Reputation in multiagent societies

The sources for reputation
4 - Reputation in multiagent societies

• **Communicated Image as a source for reputation**

It consists of aggregating the images that other members in the society communicate, taking this aggregation as the reputation value.

![Diagram showing the concept of communicated image as a source for reputation.]

Assumptions:
• the evaluation is being communicated
• the individuals that share the image are a good sample of what the whole social entity thinks.
• Communicated reputation

It is based on the aggregation of information about reputation received from third parties.

"agent A, B and C say that the reputation of D in the social entity $\alpha$ is good"

Shared Voice

"The reputation of D according to the social entity $\alpha$ is good"

Reputation evaluation

Context dependent:
- Number of communications
- Credibility of informers
- ...

• The level of individual compromise the informant is taking here is quite less than that in the communication of images.
• Inherited reputation

We call inherited reputation the reputation that
(i) is directly inherited from third party agents with whom the subject has some kind of social relation

   Example:
   An employee that works for a certain company inherits the reputation of that company.

   The member of a family inherits the reputation of his/her ancestors.

(ii) is associated to the role the subject is playing in the society.

   Example:
   The director of a research institute that has a good reputation is supposed to have a good reputation as a researcher because of the role she/he is playing in that institution.
4 - Reputation in multiagent societies

• Putting all together

Example: ReGreT

\[
R_{a \rightarrow b}(\phi) = \sum_{i \in \{W,N,S,D\}} \xi_i \cdot R_{a \rightarrow b}(\phi)
\]

\[
\begin{align*}
\xi_W &= RL_{a \rightarrow b}(\phi) \\
\xi_N &= RL_{a \rightarrow b}(\phi) \cdot (1 - \xi_W) \\
\xi_S &= RL_{a \rightarrow b}(\phi) \cdot (1 - \xi_W - \xi_N) \\
\xi_D &= 1 - \xi_W - \xi_N - \xi_S
\end{align*}
\]

* R -> Reputation value
* RL -> Reliability value
* W,N,S,D -> witness, neighborhood, system, default reputation
Centralized vs Decentralized models

- Centralized
  - All the information available in the society can be used.
  - Wrong or biased information has a lesser bad impact on the final value.
  - First comers can benefit from the information from the beginning.
  - The individuals have to trust the central service regarding the impartiality of the calculation.
  - Do not takes into account personal preferences and biases.
  - The central repository is a bottleneck for the system.
  - Security problems.

- Central service

Example: eBay

Copyright © 1995-2012 eBay Inc.
Centralized vs Decentralized models

• Decentralized

- No trust on external central entity is necessary.
- They do not introduce any bottleneck.
- Each agent can decide the method that wants to use to calculate reputation.
- It can take some time for the agent to obtain enough information to calculate a reliable reputation value. It is harder for newcomers.
- It demands more complex and “intelligent” agents.

Example: ReGreT, Travos, FIRE
4 - Reputation in multiagent societies

• Using reputation

Reputation as a source for trust

Reputation is one of the elements that can contribute to build trust on a trustee. Usually it is used when there is a lack of direct information.

Reputation for social order

Reputation incentives “socially acceptable conducts (like benevolence or altruism) and/or forbid socially unacceptable ones”.

Ostracism is the main deterrent used by reputation mechanisms.

Social order -> set of linked social structures, social institutions and social practices which conserve, maintain and enforce “normal” ways of relating and behaving.
Pitfalls when using reputation

**Attacks to reputation mechanisms**

*Compromise between waiting for clearer signals and acting against the attack*

**Unfair Ratings**

*Attack*: an agent sends deliberately wrong feedback about interactions with another agent.

*Solution*: to give more weight to the opinions of those agents that in the past have demonstrated to be more certain.

**Ballot-Stuffing**

*Attack*: an agent sends more feedback than interactions it has been partner in.

*Solution*: filtering feedback that comes from peers suspect to be ballot-stuffing and using feedback per interaction rates instead of accumulation of feedback.
• Pitfalls when using reputation

*Attacks to reputation mechanisms*

**Dynamic Personality**

*Attack:* an agent that achieves a high reputation attempts to deceive other agents taking advantage of this high reputation (“value imbalance exploitation”).

*Solution:* to have a memory window so that not all the past history is taken into account.

**Whitewashing**

*Attack:* an agent changes its identifier in order to escape previous bad feedback.

**Sybil Attacks**

*Attack:* an agent creates enough identities so it can subvert the normal functioning of the system.
• Pitfalls when using reputation

*Attacks to reputation mechanisms*

**Collusion**

*Attack:* this is not an attack “per se” but an enhancer of other attacks. A group of agents co-operate with one another in order to take advantage of the system and other agents

*Solution:* difficult to detect. Detect an important and recurrent deviation in the feedbacks of different agents regarding the same targets.

**Reputation Lag Exploitation**

*Attack:* the agent uses the lag that the reputation mechanism needs to reflect the new reality (usually a decrease in reputation) and exploits it to get benefit. Then it recovers the previous reputation value and starts again exploiting it.

*Solution:* (i) to adjust the reaction time of the reputation mechanism so it reacts quickly enough to changes in the behavior. (ii) to give the agent the possibility to detect patterns that show a cyclic behavior in the reputation value.
5 – Trust, reputation & agreement technologies
Connecting T&R with AT

T&R are meant to be used within an agent reasoning process together with other agreement technologies

- Argumentation
- Negotiation
- Norms
- Organization
- Semantics
5 – Trust, reputation & agreement technologies

Argumentation for T/R

Arguments to support/challenge social evaluations
T/R for argumentation

• Trust and reputation used to assess the strength of an argument
  – Reliability of the argument source
  – Confidence in the informative content of the argument

• Impact on the argumentation process
  – Selection of accepted arguments
  – Weighting of arguments
T/R & Negotiation

• The field of trust negotiation is interested in establishing an incremental exchange of trust evidences between two parties (ex: the Keynote system [Blaze et al, 96])

• In a multi-agent negotiation, T&R are useful for local agent decisions
  – for the acceptance of a proposal
  – for selecting partners
T/R & Norms

• A tool for social control of norm obedience
• Scope of norms
  – Individual, group or global
  – May correspond to rules, protocols, contracts, ...
• The satisfaction of a norm can be the context of a T&R evaluation
  – Alice distrusts Bob for respecting his commitments towards her
  – Charles has the reputation of sending answers to any queries as it is defined in the interaction protocols of the society
Multi-agent organizational models provide interesting elements for T&R

• from concepts
  – Reputation in a given *group*
  – Trust in a given *role*

• from infrastructures
  – Repository artefacts to share reputation in a group
T/R & Semantics

Using a common ontology is an approach used to solve the problem of heterogeneity
6 – Conclusions
In summary...

- Trust and reputation have become essential concepts in a multi-agent systems
  - Firstly introduced to implement social control
  - Now vital when dealing with open, heterogeneous multi-agent applications
- Trust models include both a representation formalism and a decisions process
- Reputation is a social evaluation that circulates in a society. It is a source of trust
Current challenges

• There exists now many trust and reputation models
• Current research challenge is now more on their deployment under specific conditions...
  – when an agent have no or multiple identities
  – when human users and software agents interact in mixed virtual communities
  – when privacy issues should be considered when sharing social evaluations
• ... and in their integration in an agent architecture
  – Intertwining T&R with agreement technologies
Selected references
(go to the book chapter for a complete list)