A Logic of Trust and Reputation

Emiliano Lorini
(joint work with A. Herzig, J. F. Hübner, L. Vercouter)

IRIT-CNRS, Toulouse, France

Workshop on Trust, Advice, and Reputation
Toulouse, 22-24 October 2009

Motivation: formal analysis of trust and reputation

- many computational models [Sabater & Sierra 02; Huynh et al. 04]
  - quantitative
  - no qualitative analysis
- few logical models
  - mostly focused on trust in information sources [Liau 03; Jones & Firozabadi 01; Dastani et al. 04]
- cognitive model of trust [Castelfranchi & Falcone 98, 01]
  - informal definitions

$\Rightarrow$ formal definitions?
$\Rightarrow$ relations between trust and reputation?

Contribution: definitions in a BDI logic

[Lorini & Demolombe 2008, 2009; Lorini, Falcone, Castelfranchi 2008; Herzig, Lorini, Hübner, Vercouter 2009]

- 'top-down' qualitative analysis: reduction of trust to more primitive concepts
  - trust $\Rightarrow$ belief, goal, intention, action
  - analysis in terms of logics of action and time, BDI logics
- two versions of trust
  - 'occurrent trust' vs. 'dispositional trust'
  - refinement of Castelfranchi & Falcone's definition
- from dispositional trust to reputation

Trust: the ingredients [Castelfranchi & Falcone]

"*i* trusts *j* to do α in order to achieve φ"

- truster *i*,
- trustee *j*,
- action α of *j*,
- goal φ of *i*.
  - important: trust $\neq$ thinking and foreseeing

Informal definition [Castelfranchi & Falcone]

$\Rightarrow$ trust $\approx$ evaluation of the truster about certain properties of trustee (that are relevant for a goal/task φ)

"*i* trusts *j* to do α in order to achieve φ" if and only if:

1. *i* has the *goal* φ;
2. *i* believes that
   1. *j* is *capable* to do α;
   2. *j* has the *power* to achieve φ by doing α;
   3. *j intends* to do α.

$\Rightarrow$ trust defined from four more primitive concepts

$\Rightarrow$ logic of goal, ability, power and intention?

Two kinds of trust: occurrent vs. dispositional

Occurrent trust vs. dispositional trust

two perspectives on trustee's action α:

- truster believes trustee is going to do α *here and now*
  $\Rightarrow$ **occurrent trust**
- truster believes trustee is going to do α *whenever some conditions are satisfied*.
  $\Rightarrow$ **dispositional trust**

(cf. occurrent belief vs. dispositional belief [Searle 92])

Occurrent trust

i wants φ to be true and believes *j* is going to perform α *here and now* so that φ will be ensured

$$\begin{aligned} \mathit{OccTrust}(i,j,\alpha,\varphi) \quad \stackrel{\texttt{def}}{=} \quad & \mathit{OccGoal}(i,\varphi) \wedge \\ & \mathit{Belief}(i, \\ & \mathit{OccCap}(j,\alpha) \wedge \\ & \mathit{OccPower}(j,\alpha,\varphi) \wedge \\ & \mathit{OccIntends}(j,\alpha)) \end{aligned}$$

$\Rightarrow$ predicates to be defined: *Belief*, *OccGoal*, *OccCap*, *OccPower*, *OccIntends*

Example

1's *occurrent trust* in 2 to send a certain product in view of satisfying 1's goal of possessing the product:

- 1 wants to possess the product,
- 1 believes that
  - 2 is capable to send the product,
  - 2's act of sending the product will result in 1 possessing it,
  - 2 has the intention to send the product.

Dispositional trust

i expects that there will be some situations κ in which he will have the goal of achieving φ (*potential goal*), and believes that in all these situations *j* will ensure φ by doing action α

$$\begin{aligned} \mathit{DispTrust}(i,j,\alpha,\varphi,\kappa) \quad \stackrel{\mathtt{def}}{=} \quad & \mathit{PotGoal}(i,\varphi,\kappa) \wedge \\ & \mathit{Belief}(i, \mathit{Henceforth} \\ & ((\kappa \wedge \mathit{OccGoal}(i,\varphi)) \rightarrow \\ & (\mathit{OccCap}(j,\alpha) \wedge \\ & \mathit{OccPower}(j,\alpha,\varphi) \wedge \\ & \mathit{OccIntends}(j,\alpha))) \end{aligned}$$

$\Rightarrow$ predicates to be defined: *PotGoal*, *Belief*, *Henceforth*, *OccGoal*, *OccCap*, *OccPower*, *OccIntends*

Example

1's *dispositional trust* in 2 (a mechanic) to repair his car so that the car will be in order, in the circumstances in which 1 will ask the mechanic to repair his car

Logical formalization

Occurrent trust and dispositional trust: components

- definiendum:
  - belief,
  - occurrent goal,
  - occurrent capability,
  - occurrent power,
  - occurrent intention.
- definiens: formulas of a modal logic of time, belief, preference, and action
  - 'spell out' predicates *Belief*(i, φ), *OccGoal*(i, φ), *PotGoal*(i, φ, κ), *OccCap*(j, α), *OccPower*(j, α, φ), *OccIntends*(j, α) using modal operators of time, belief, preference, and action

convention:

- $\texttt{Bel}_i$, etc. = logical operators

Temporal operators

$\texttt{Henceforth}\,\varphi$ = "φ henceforth holds"

$$\texttt{Eventually}\,\varphi \stackrel{\texttt{def}}{=} \neg\,\texttt{Henceforth}\,\neg\varphi$$

$\texttt{Eventually}\,\varphi$ = "φ eventually holds"

- will be used to define $\mathit{OccGoal}(i, \varphi)$, etc.

Belief operators

$\texttt{Bel}_i\,\varphi$ = "agent *i* believes that φ"

$$\texttt{Poss}_i\,\varphi \stackrel{\texttt{def}}{=} \neg\texttt{Bel}_i\,\neg\varphi$$

$\texttt{Poss}_i\,\varphi$ = "agent *i* thinks that φ is possible"

- epistemic and doxastic logic [Hintikka 62]
- express truster's beliefs about trustee's properties

$$\mathit{Belief}(i,\varphi) \stackrel{\texttt{def}}{=} \texttt{Bel}_i\,\varphi$$

Goals as preferences about the future

$\texttt{Pref}_i\,\varphi$ = "agent i wants φ to be true"

- binary preferences [Cohen & Levesque 90]
- positive introspection: $\vdash \texttt{Pref}_i\,\varphi \rightarrow \texttt{Bel}_i\,\texttt{Pref}_i\,\varphi$
- negative introspection: $\vdash \neg\texttt{Pref}_i\,\varphi \rightarrow \texttt{Bel}_i\,\neg\texttt{Pref}_i\,\varphi$
- realism: $\vdash \texttt{Bel}_i\,\varphi \rightarrow \texttt{Pref}_i\,\varphi$

$$\mathit{OccGoal}(i,\varphi) \stackrel{\texttt{def}}{=} \texttt{Pref}_i\,\texttt{Eventually}\,\varphi$$

Capability in dynamic logic

$\texttt{After}_{i:\alpha}\,\varphi$ = "φ will be true after *every possible* execution of action α by agent *i*"

- propositional dynamic logic (PDL)
- formulas $= \varphi, \psi, \ldots =$ state descriptions
- actions $= \alpha, \beta, \ldots =$ state transition descriptions
  - action $\neq$ formula
  - $i : \alpha$ = action α with author *i*
- $\texttt{After}_{i:\alpha}\,\bot$ = "*i* cannot do α"

$$\mathit{OccCap}(j, \alpha) \stackrel{\texttt{def}}{=} \neg\texttt{After}_{j:\alpha}\,\bot$$

- "*j* is capable to perform α" = "*j* can do α"

Power in dynamic logic

$$\mathit{OccPower}(j, \alpha, \varphi) \stackrel{\text{def}}{=} \texttt{After}_{j:\alpha}\, \varphi$$

- relates *j*'s action α with *i*'s goal φ
- missing: epistemic aspect of power ($\Rightarrow$ 'knowing how to play')

Intention-to-do as preferred action

$\texttt{Does}_{i:\alpha}\,\varphi$ = "agent *i* is going to do α and φ will be true afterwards"

- $\texttt{Does}_{i:\alpha}\,\top$ = *i* does α

$$\mathit{OccIntends}(j,\alpha) \stackrel{\texttt{def}}{=} \texttt{Pref}_j\,\texttt{Does}_{j:\alpha}\,\top$$

- Some axioms:
  1. $\texttt{Does}_{i:\alpha}\,\varphi \rightarrow \neg\texttt{After}_{i:\alpha}\,\neg\varphi$ ($\vdash \texttt{Does}_{i:\alpha}\,\top \rightarrow \neg\texttt{After}_{i:\alpha}\,\bot$)
  2. $(\texttt{Pref}_i\,\texttt{Does}_{i:\alpha}\,\top \wedge \neg\texttt{After}_{i:\alpha}\,\bot) \rightarrow \texttt{Does}_{i:\alpha}\,\top$
  3. $\texttt{Does}_{i:\alpha}\,\top \rightarrow \texttt{Pref}_i\,\texttt{Does}_{i:\alpha}\,\top$

Theorem

$\vdash (\texttt{Pref}_i\,\texttt{Does}_{i:\alpha}\,\top \wedge \neg\texttt{After}_{i:\alpha}\,\bot) \leftrightarrow \texttt{Does}_{i:\alpha}\,\top$

Formal definition of occurrent trust

$$\mathit{OccTrust}(i,j,\alpha,\varphi) \stackrel{\text{def}}{=} \begin{array}{l} \mathtt{Pref}_i\,\mathtt{Eventually}\,\varphi \wedge \\ \mathtt{Bel}_i\,(\neg\mathtt{After}_{j:\alpha}\,\bot \wedge \mathtt{After}_{j:\alpha}\,\varphi \wedge \\ \mathtt{Pref}_j\,\mathtt{Does}_{j:\alpha}\,\top) \end{array}$$

Example

1's *occurrent trust* in 2 to send a certain product in view of satisfying 1's goal of possessing the product

$$
\begin{aligned}
&\mathit{OccTrust}(1, 2, \mathit{sendP1}, \mathit{hasP1}) \stackrel{\text{def}}{=} \\
&\mathtt{Pref}_1\, \mathtt{Eventually}\, \mathit{hasP1} \wedge \\
&\mathtt{Bel}_1\, (\neg \mathtt{After}_{2:\mathit{sendP1}}\, \bot \wedge \\
&\mathtt{After}_{2:\mathit{sendP1}}\, \mathit{hasP1} \wedge \\
&\mathtt{Pref}_2\, \mathtt{Does}_{2:\mathit{sendP1}}\, \top)
\end{aligned}
$$

Some properties of occurrent trust

Theorem

- $\vdash \mathit{OccTrust}(i,j,\alpha,\varphi) \leftrightarrow$
  $(\mathtt{Pref}_i\ \mathtt{Eventually}\,\varphi \wedge \mathtt{Bel}_i\,(\mathtt{Does}_{i:\alpha}\,\top \wedge \mathtt{After}_{j:\alpha}\,\varphi))$
- $\vdash \mathit{OccTrust}(i,j,\alpha,\varphi) \leftrightarrow \mathtt{Bel}_i\,\mathit{OccTrust}(i,j,\alpha,\varphi)$
- $\vdash \mathit{OccTrust}(i,j,\alpha,\varphi) \rightarrow \mathtt{Bel}_i\ \mathtt{Eventually}\,\varphi$

Potential goal

$$\mathit{PotGoal}(i, \varphi, \kappa) \stackrel{\text{def}}{=}$$

$$\mathtt{Poss}_i\, \mathtt{Eventually}\,(\kappa \wedge \mathtt{Pref}_i\, \mathtt{Eventually}\, \varphi)$$

Formal definition of dispositional trust

$$\mathit{DispTrust}(i, j, \alpha, \varphi, \kappa) \stackrel{\text{def}}{=}$$

$$\begin{aligned}
&\mathtt{Poss}_i\,\mathtt{Eventually}\,(\kappa \wedge \mathtt{Pref}_i\,\mathtt{Eventually}\,\varphi) \wedge \\
&\mathtt{Bel}_i\,\mathtt{Henceforth}\,((\kappa \wedge \mathtt{Pref}_i\,\mathtt{Eventually}\,\varphi) \rightarrow \\
&(\neg\mathtt{After}_{j:\alpha}\,\bot \wedge \mathtt{After}_{j:\alpha}\,\varphi \wedge \mathtt{Pref}_j\,\mathtt{Does}_{j:\alpha}\,\top))
\end{aligned}$$

Example

1's *dispositional trust* in 2 (a mechanic) to repair his car so that the car will be in order, in the circumstances in which 1 will ask the mechanic to repair his car

$$\mathit{DispTrust}(1, 2, \mathit{rep1c}, \mathit{OK1c}, \mathit{ask12}) \stackrel{\text{def}}{=}$$

$$\begin{aligned}&\mathtt{Poss}_1\ \mathtt{Eventually}\,(\mathit{ask12} \wedge \mathtt{Pref}_1\ \mathtt{Eventually}\,\mathit{OK1c}) \wedge \\ &\mathtt{Bel}_1\ \mathtt{Henceforth}\,((\mathit{ask12} \wedge \mathtt{Pref}_1\ \mathtt{Eventually}\,\mathit{OK1c}) \rightarrow \\ &(\neg\mathtt{After}_{2:\mathit{rep1c}}\,\bot \wedge \mathtt{After}_{2:\mathit{rep1c}}\,\mathit{OK1c} \wedge \mathtt{Pref}_2\,\mathtt{Does}_{2:\mathit{rep1c}}\,\top))\end{aligned}$$

Some properties of dispositional trust

Theorem

- $\vdash \mathit{DispTrust}(i, j, \alpha, \varphi, \kappa) \leftrightarrow$
  $(\mathtt{Poss}_i\ \mathtt{Eventually}\,(\kappa \wedge \mathtt{Pref}_i\ \mathtt{Eventually}\,\varphi) \wedge$
  $\mathtt{Bel}_i\ \mathtt{Henceforth}\,((\kappa \wedge \mathtt{Pref}_i\ \mathtt{Eventually}\,\varphi) \rightarrow$
  $(\mathtt{Does}_{i:\alpha}\,\top\ \wedge \mathtt{After}_{j:\alpha}\,\varphi)))$
- $\vdash \mathit{DispTrust}(i, j, \alpha, \varphi, \kappa) \leftrightarrow \mathtt{Bel}_i\ \mathit{DispTrust}(i, j, \alpha, \varphi, \kappa)$

From dispositional to occurrent trust

Theorem

$$\vdash (\mathit{DispTrust}(i, j, \alpha, \varphi, \kappa) \wedge \mathtt{Pref}_i\, \mathtt{Eventually}\, \varphi \wedge \mathtt{Bel}_i\, \kappa) \rightarrow \mathit{OccTrust}(i, j, \alpha, \varphi)$$

Discussion: other definition of trust

- Deutsch's definition (1958): trust involves risk perception (trusting is doing a real bet on the trustee)
  - Uncertain beliefs are needed to capture it
- Genuine trust (Trust game, Baier 1986, Holton 1994): *i*'s belief that *j* has goodwill towards him (*j* is willing to sacrifice his wellbeing for *i*)
  - It is just a special case of our definition
- Jones's definition (2002): rule belief+conformity belief
  - Trust $\neq$ thinking or foreseeing

From trust to reputation

Reputation: the building blocks

Rep$(I, j, \alpha, \varphi, \kappa)$ = "*j* has reputation in group *I* to ensure φ by doing α in the circumstances κ"

- Reputation as the collective counterpart of trust
- TRUST = agent *i*'s individual evaluation (belief) about some properties of agent *j* that are relevant for a goal of *i*
- REPUTATION = group *I*'s collective evaluation about some properties of agent *j* that are relevant for a (group) goal of *I*
- to be defined: collective evaluation, group goals

Public facts

For every I such that $|I| > 2$,
$\texttt{Public}_I\,\varphi$ = "φ is public in the group of agents I" ("φ is said in I")

- "φ is public in I" $\neq$ common belief in I that φ
- "φ is public in I" does not imply "every agent in I believes φ"

Logic of public facts: some principles

1. $\neg(\mathtt{Public}_I\,\varphi \wedge \mathtt{Public}_I\,\neg\varphi)$
2. $\mathtt{Public}_I\,\varphi \to \mathtt{Public}_J\,\mathtt{Public}_I\,\varphi$
   if $J \subseteq I$
3. $\neg\mathtt{Public}_I\,\varphi \to \mathtt{Public}_J\,\neg\mathtt{Public}_I\,\varphi$
   if $J \subseteq I$
4. $\mathtt{Public}_I\,\varphi \to \mathtt{Bel}_i\,\mathtt{Public}_I\,\varphi$
   if $i \in I$
5. $\neg\mathtt{Public}_I\,\varphi \to \mathtt{Bel}_i\,\neg\mathtt{Public}_I\,\varphi$
   if $i \in I$
6. $\mathtt{Public}_I\,(\bigwedge_{i\in I}\mathtt{Bel}_i\,\varphi \to \varphi)$

Group goals

$\texttt{GroupPref}_I\,\varphi$ = "the agents in I want that φ"

- group goals: broad sense (individual preference aggregation)
- weaker than joint goals and joint intentions [Grosz & Kraus 96]

Some examples of $\texttt{GroupPref}_I\,\varphi$ definition:

- $\exists$-group goal: $\bigvee_{i \in I} \texttt{Pref}_i\,\varphi$
- $\forall$-group goal: $\bigwedge_{i \in I} \texttt{Pref}_i\,\varphi$
- Group goal based on majority: $\bigvee_{J \subseteq I, |J| > |I \setminus J|} \bigwedge_{i \in J} \texttt{Pref}_i\,\varphi$

Formal definition of reputation

$$\mathit{Rep}(I, j, \alpha, \varphi, \kappa) \stackrel{\text{def}}{=}$$

$$\mathit{PotGoal}(I, \varphi) \wedge \mathtt{Poss}_I\, \mathtt{Henceforth}\,((\kappa \wedge \mathtt{GroupPref}_I\, \mathtt{Eventually}\, \varphi) \rightarrow (\neg \mathtt{After}_{j:\alpha}\, \bot \wedge \mathtt{After}_{j:\alpha}\, \varphi \wedge \mathtt{Pref}_j\, \mathtt{Does}_{j:\alpha}\, \top))$$

- $\mathtt{Poss}_I\, \varphi \stackrel{\text{def}}{=} \neg \mathtt{Public}_I\, \neg \varphi$
  - $\mathtt{Poss}_I\, \varphi$ = "according to the group I, φ is possible"
- $\mathit{PotGoal}(I, \varphi) \stackrel{\text{def}}{=}$
  $\mathtt{Poss}_I\, \mathtt{Eventually}\, (\kappa \wedge \mathtt{GroupPref}_I\, \mathtt{Eventually}\, \varphi)$

Conclusion

Conclusion

- Contribution.
  - formal definitions of occurrent trust and dispositional trust
  - formal definition of reputation and its relation with trust
- Our related works
  - semantics for the different modalities
  - mathematical properties of the logic of belief, preference, action and time (soundness, completeness)
  - trust in information sources and in communication systems
  - from binary trust to graded trust
  - trust in agents vs. trust in roles
  - Implementation of a BDI agent reasoning about trust (see Vercouter's presentation)

THANK YOU!