Prolegomena for a logic of trust and reputation

Emiliano Lorini

Institut de Recherche en Informatique de Toulouse (IRIT), France

NorMAS 2008, 15th July 2008

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; Demolombe 01; 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

- 'top-down' qualitative analysis: reduction of trust to more primitive concepts
  - trust $\Rightarrow$ belief, goal, intention, action
  - analysis in terms of existing 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 trust and dispositional trust

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 believes *j* is going to perform α *here and now*

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

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

Dispositional trust

i believes *henceforth*, *j* will perform α *under some conditions*

$$\begin{aligned} \mathit{DispTrust}(i,j,\alpha,\varphi) \quad \stackrel{\texttt{def}}{=} \quad & \mathit{PotGoal}(i,\varphi) \wedge \\ & \mathit{Believes}(i, \mathit{CondCap}(j,\alpha)) \wedge \\ & \mathit{Believes}(i, \mathit{CondPower}(j,\alpha,\varphi)) \wedge \\ & \mathit{Believes}(i, \mathit{CondIntends}(j,\alpha)) \end{aligned}$$

$\Rightarrow$ predicates to be defined: *PotGoal*, *CondCap*, *CondPower*, *CondIntends*

$\Rightarrow$ relationship between occurrent and dispositional trust:
$\vdash$ (*DispTrust*(i,j,α,φ) & conditions) $\rightarrow$ *OccTrust*(i,j,α,φ)

Occurrent trust

Occurrent 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 $\mathit{Believes}(i, \psi)$, $\mathit{OccGoal}(i, \varphi)$ using modal operators of time, belief, preference, and action

convention:

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

Temporal operators

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

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

Belief operators

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

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

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

Goals as preferences about the future

$\texttt{Pref}_i\,\varphi$ = "agent *i* prefers that φ"

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

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

- add a negative condition: $\texttt{Pref}_i\,\texttt{Eventually}\,\varphi \wedge \neg\texttt{Bel}_i\,\varphi$

Capability in dynamic logic

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

- 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" = "α can be executed by j"

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"

- $\neg\texttt{After}_{i:\alpha}\,\bot$ = *i* can do α
- $\texttt{Does}_{i:\alpha}\,\top$ = *i* does α
- logical relationships [Lorini & Demolombe DEON'08]:
  $\texttt{Does}_{i:\alpha}\,\varphi \rightarrow \neg\texttt{After}_{i:\alpha}\,\neg\varphi$ ($\texttt{Does}_{i:\alpha}\,\top \rightarrow \neg\texttt{After}_{i:\alpha}\,\bot$)

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

- present-directed intention
- vs. future-directed intention:
  $\texttt{Pref}_j\,\texttt{Eventually}\,\texttt{Does}_{j:\alpha}\,\top$ [Bratman 87, Lorini & Herzig 08]

Formal definition of occurrent trust

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

Dispositional trust

Dispositional trust: components

- definiendum:
  - belief,
  - potential goal,
  - conditional capability,
  - conditional power,
  - conditional intention.
- definiens: formulas of a modal logic of time, belief, preference, and action
- conditional capability, etc. = conditioned occurrent capability etc., prefixed by ‘henceforth’

Potential goal

$$\mathit{PotGoal}(i, \varphi) \stackrel{\text{def}}{=} \neg \mathtt{Bel}_i\, \mathtt{Henceforth}\, \neg \mathit{OccGoal}(i, \varphi)$$

Conditional capability

$$\mathit{CondCap}(j,\alpha) \stackrel{\text{def}}{=} \texttt{Henceforth}\,(\kappa_{\mathit{OccCap}(j,\alpha)} \rightarrow \mathit{OccCap}(j,\alpha))$$

- $\kappa_{\mathit{OccCap}(j,\alpha)}$ = circumstances under which *i* expects *j* has (occurrent) capability to perform α
- **executability precondition of** α

Conditional power

$$\mathit{CondPower}(j,\alpha,\varphi) \stackrel{\texttt{def}}{=} \texttt{Henceforth}\,(\kappa_{\mathit{OccPower}(j,\alpha,\varphi)} \rightarrow \mathit{OccPower}(j,\alpha,\varphi))$$

- $\kappa_{\mathit{OccPower}(j,\alpha,\varphi)}$ = circumstances under which *i* expects that *j* has the power to obtain φ via α
- **effect precondition for α causing φ**

Conditional intention

$$\mathit{CondIntends}(j,\alpha) \stackrel{\texttt{def}}{=} \texttt{Henceforth}\,(\kappa_{\mathit{OccIntends}(j,\alpha)} \rightarrow \mathit{OccIntends}(j,\alpha))$$

- $\kappa_{\mathit{OccIntends}(j,\alpha)}$ = circumstances under which *i* expects *j* will intend to perform α
- $\mathit{CondIntends}(j,\alpha)$ = "willingness"
- different forms of willingness [Lorini & Demolombe DEON'08]:
  - when *j* believes that *i* wants him to do α, *j* forms the intention to do α ($\Rightarrow$ *i* is adoptive toward *j*)
  - when *j* believes to be obliged to do α, *j* forms the intention to do α ($\Rightarrow$ *i* is obedient).

Formal definition of dispositional trust

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

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

- $\kappa_{\mathit{OccCap}(j,\alpha)}$, $\kappa_{\mathit{OccPower}(j,\alpha,\varphi)}$, $\kappa_{\mathit{OccIntends}(j,\alpha)}$ = nonlogical conditions
- should be jointly true when i has occurrent goal that φ
- replace $\mathit{PotGoal}(i,\varphi)$ by:
  $\neg \mathtt{Bel}_i\, \mathtt{Henceforth}\, \neg (\mathtt{Pref}_i\, \mathtt{Eventually}\, \varphi \wedge \kappa_{\mathit{OccCap}(j,\alpha)} \wedge \kappa_{\mathit{OccPower}(j,\alpha,\varphi)} \wedge \kappa_{\mathit{OccIntends}(j,\alpha)})$.

From dispositional to occurrent trust

Theorem

Suppose $\texttt{Bel}_i$ *and* $\texttt{Henceforth}$ *are normal modal operators, and suppose* $\texttt{Henceforth}\,\varphi \rightarrow \varphi$. *Then*

$$\begin{aligned}\vdash (\quad & \mathit{DispTrust}(i,j,\alpha,\varphi)\\ & \wedge \mathit{OccGoal}(i,\varphi)\\ & \wedge \mathit{Believes}(i,\kappa_{\mathit{OccCap}(j,\alpha)})\\ & \wedge \mathit{Believes}(i,\kappa_{\mathit{OccPower}(j,\alpha,\varphi)})\\ & \wedge \mathit{Believes}(i,\kappa_{\mathit{OccIntends}(j,\alpha)}) \quad) \rightarrow \mathit{OccTrust}(i,j,\alpha,\varphi)\end{aligned}$$

Formal definition of reputation

Reputation: the building blocks

$Rep(I, j, \alpha, \varphi)$ = "j has reputation to do α in order to achieve φ in group I"

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

Group beliefs and group goals

$\texttt{GroupBelief}_I\varphi$ = "the group of agents I believes that φ"
[Tuomela 02; Gaudou et al. 06; Lorini et al. 08]

- "I has group belief that φ" = "φ is publicly established in the group of agents I"
- group belief $\neq$ common belief
  - group belief does not imply individual belief
- "I has group belief that φ" $\approx$ "shared voice in I that φ" [Conte & Paolucci 02]

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

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

Formal definition of reputation

$$\mathit{Rep}(i,j,\alpha,\varphi) \stackrel{\texttt{def}}{=} \mathit{PotGoal}(i,\varphi) \wedge \texttt{GroupBelief}_I\, \mathit{CondCap}(j,\alpha) \wedge \texttt{GroupBelief}_I\, \mathit{CondPower}(j,\alpha,\varphi) \wedge \texttt{GroupBelief}_I\, \mathit{CondIntends}(j,\alpha)$$

- $\mathit{PotGoal}(i,\varphi) \stackrel{\texttt{def}}{=} \neg\texttt{GroupBelief}_I\, \texttt{Henceforth}\, \neg\texttt{GroupPref}_I\, \texttt{Eventually}\, \varphi$
- $\mathit{CondCap}(j,\alpha)$, $\mathit{CondPower}(j,\alpha,\varphi)$ and $\mathit{CondIntends}(j,\alpha)$ defined as before

Related works and conclusion

Existing implemented models of reputation: a comparison (1)

Cognitive (Cog): mental states like beliefs are used to define the reputation.

Evaluation group (Grp): a target agent can have different reputations for different groups.

Group goal (GG): the reputation is with respect to a goal of the group.

Action capability (Cap): the definition of reputation considers the agent's capabilities to do a certain action.

Action power (Pow): the definition of reputation considers the power of the agent to achieve the group goal by means of an action.

Intention (Int): the definition of reputation considers the agent's intentions to perform the action or goal for the group.

Existing computational models of reputation: a comparison (2)

Definition	Cog	Grp	GG	Cap	Pow	Int
Our model	yes	yes	yes	yes	yes	action
Repage [Sabater et al. 06]	yes	yes	yes	no	no	goal
FIRE [Huynh et al. 04]	no	no	no	no	no	action
LIAR [Muller & Vercouter 05]	no	no	yes	no	no	no
Regret [Sabater & Sierra 02]	no	yes	no	yes	no	no
e-Bay	no	no	no	no	no	no

Table: Properties of definitions of reputation

Conclusion

- Contribution.
  - formal definitions of occurrent trust, dispositional trust, and reputation
  - evaluation of existing reputation systems w.r.t. the formal definition
- Our related works [DEON'08, KRAMAS'08].
  - mathematical properties of the logic of belief, preference, action and time (soundness, completeness)
  - interactions between the different modalities
  - trust in information sources and communication systems

THANK YOU!