Formalisation de la confiance sociale et mise en oeuvre dans un système multi-agent

Projet ANR ForTrust (07-11) www.irit.fr/ForTrust

IRIT-LILaC, Toulouse
EMSE-G2I, Saint-Etienne
ISTC-CNR, Rome

October 17, 2008

Context and Motivations

Trust: “everything is based on it”

- context: open and decentralized systems
  - incomplete information
  - erroneous information
  - uncertainty about agents’ identity

Trust: “everything is based on it”

- examples:
  - “Is information from this website correct?”
  - “I can book this hotel via two different sites, and prices are different; which one should I choose?”
  - “Should I take this email serious?”
  - “Will this webservice sell my data without consent?”
  - “Will this intelligent device do what I asked for? (and not more than that?)”
  - . . .

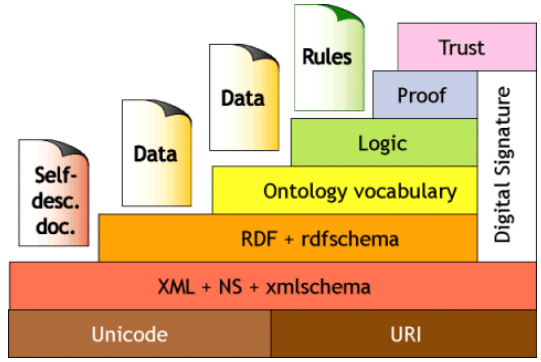
Background: existing trust management systems

"...address only narrow subsets of the overall trust management problem and often do so in a manner that is appropriate to only one application" [Blaze, Feigenbaum & Lacy, 96]

- essentially still true today
  - focus: static form of trust
    $\Rightarrow$ certificates
    $\Rightarrow$ authentication, access control
  - missing: cognitive aspect, dynamics, versatility

Human2Machine vs. Machine2Machine

- Trust for Human to Machine interactions (cf. T. Berners-Lee's cake model of the Internet)



- Trust for Machine to Machine interactions
  - Peer-to-peer systems
  - Business to Business, automatic negotiation
  - Field of **Multi-Agent Systems** (MAS)

Background: existing models of reputation

- Wide variety of multi-agent reputation models: SPORAS [Zacharia, 99], Regret [Sabater,02], socio-cognitive model [Conte & Paolucci, 02], FIRE [Huyns, 04], LIAR [Muller & Vercouter, 06], RepAge [Sabater & Paolucci, 06], ...
- varying focus $\Rightarrow$ difficult to compare
  - reputation
  - reliability
  - sincerity
  - quality of service
  - ...
- semi-formal
- no consensus

The ForTrust project (ANR SETIN 07-11)

- statement
  - many implemented systems
  - many theoretical models
    - quantitative
    - no logical account
- ForTrust project goals
  - cognitive model of trust by Castelfranchi&Falcone [C&F]
  - formalization in modal logic
  - establish relation with reputation
  - implementation in MAS
  - application to security systems

The ForTrust project ANR SETIN (07-11)

- statement
  - many implemented systems
  - many theoretical models
    - quantitative
    - no logical account
- ForTrust project goals
  - cognitive model of trust by Castelfranchi&Falcone [C&F]
  - formalization in modal logic
  - establish relation with reputation
  - implementation in MAS
  - application to security systems

Castelfranchi & Falcone's theory of trust

General properties of trust

- related to truster's goals
  - I trust you only if you can help me to achieve my goals or to solve tasks which are important for me.
- involves complex evaluation of the trustee
  - The decision to trust you results from the evaluation of your capabilities to solve the task, your powers, your willingness,. . .

Trust as a subjective attitude in a social context

- trust based on objective information
  - series of observations (e.g. it has been observed for a long time that this payment protocol is secure)
  - proof (e.g. it has been proved that this cryptographic technology requires at least such amount of resources to be broken)
- trust based on social relationships
  - reputation: the certificate delivered by such institution has the reputation to guarantee safe payments
  - commitment: such internet bookshop commits itself to deliver books in less than 2 days
  - inter-individual relationships: I trust this service because my friend trusts it

C&F's informal definition

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

$\Rightarrow$ defined from four more primitive concepts

- goal, ability, power and intention
- no formal definition of these concepts in [C&F]

$\Rightarrow$ here: represented as formulas of a modal logic of belief, preference, time, and action

A logical framework of 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 95)

Occurrent trust: formal definition

Agent *i* trusts agent *j* in an occurrent manner to do α in order to achieve φ iff

- *i* has the goal that φ;
- *i* believes that *j* has the occurrent capability to do α
- *i* believes that *j* has the occurrent power to ensure φ by doing α
- *i* believes that has the occurrent intention to do α

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

Dispositional trust: formal definition

Agent *i* trusts agent *j* in a dispositional manner to do α in order to achieve φ iff

- *i* has potentially the goal that φ;
- *i* believes that *j* has the conditional capability to do α
- *i* believes that *j* has the conditional power to ensure φ by doing α
- *i* believes that has the conditional intention to do α

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

From dispositional to occurrent trust

Theorem

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

Then

$$\vdash \Big(\begin{array}{l} \mathit{DispTrust}(i, j, \alpha, \varphi) \\ \wedge \texttt{Pref}_i\ \texttt{Eventually}\,\varphi \\ \wedge \texttt{Bel}_i\ \kappa_{\mathit{OccCap}(j,\alpha)} \\ \wedge \texttt{Bel}_i\ \kappa_{\mathit{OccPower}(j,\alpha,\varphi)} \\ \wedge \texttt{Bel}_i\ \kappa_{\mathit{OccIntends}(j,\alpha)} \end{array} \Big) \rightarrow \mathit{OccTrust}(i, j, \alpha, \varphi)$$

Two kinds of occurrent trust: trust in the action vs. trust in the inaction

- truster believes that trustee is in condition to *further* the achievement of his goals, and he will *do* that
  $\Rightarrow$ *trust in the action*
- truster believes that trustee is in condition to *endanger* the achievement of his goals, but he will *refrain* from doing that.
  $\Rightarrow$ *trust in the inaction*

(cf. *doing* vs. *refraining* (or *forbearing*), Von Wright 63, Belnap et al. 01)

Trust in the action vs. trust in the inaction: formal definitions

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

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

A logical framework of reputation

Formal definition of reputation: the building blocks

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

- also to be defined from :
  - j's capability,
  - j's willingness,
  - j's power.
- formally builded from:
  - *group belief* of the evaluating agents (*vs* individual belief in the case of trust)
  - *group goals* of the evaluating agents (*vs* individual goals in the case of trust)

$\Rightarrow$ To be defined: group beliefs, group goals

Group beliefs and group goals

$\texttt{GroupBelief}_I\varphi$ = "it is public in the group I that φ"

- group belief $\neq$ mutual belief
  (group belief does not imply individual belief
  [Tuomela 95, 02; Gaudou et al. 06, 08])

$\texttt{GroupPref}_I\varphi$ = "the group I *prefers* that φ"

- group goals: broad sense (including norms, standards, values . . .)
- does not imply individual preferences

Formal definition of reputation

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

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

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

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

Agent Prototype

Implementation of a prototype

- Implementation of an agent reasoning about trust using our model
- Logical formalization used as a specification
- Operational prototype of the trust framework
  - in a BDI language: Jason
  - deployed on the “standard” testbed: ART
- Reputation framework not yet included

Application to the ART Testbed

- Definition of ART's Occurent Capability

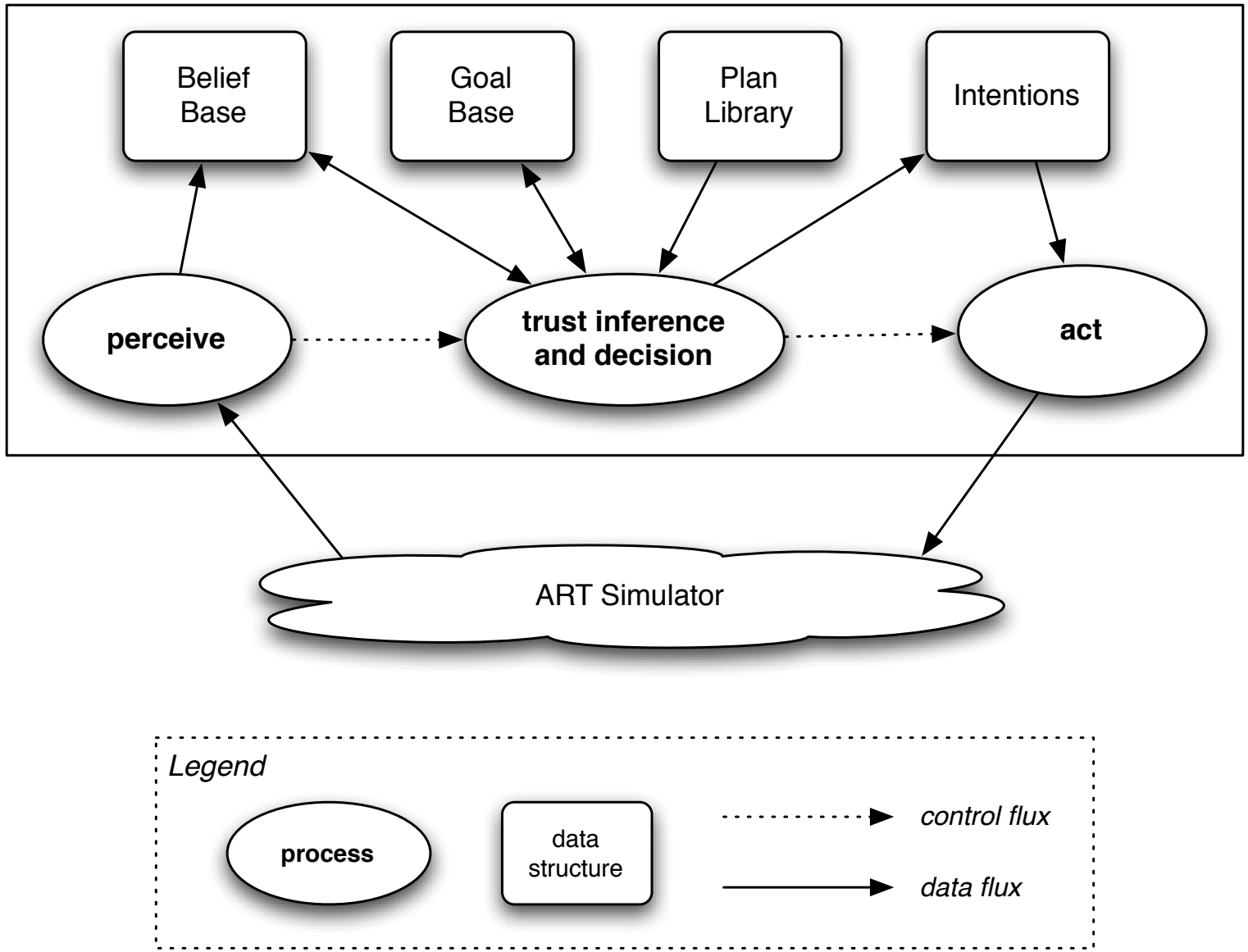
$$\begin{aligned} &\mathit{Believes}(i, \mathit{OccAct}(j, \alpha), x) \\ &\quad \leftarrow \quad \mathit{Believes}(i, \mathit{opinions\_count}(j, \mathit{asked}, \mathit{got}), 1) \wedge \\ &\qquad \mathit{asked} > 0 \ \wedge \ x = \frac{\mathit{got}}{\mathit{asked}} \ \wedge \ x > \epsilon \end{aligned} \tag{1}$$

- Definition of ART's Occurent Power

$$\begin{aligned} &\mathit{Believes}(i, \mathit{OccPower}(j, \alpha, \varphi), y) \\ &\quad \leftarrow \quad \mathit{Believes}(i, \mathit{sincere}(j), 1) \wedge \\ &\qquad \mathit{Believes}(i, \mathit{painting}(e, p), 1) \wedge \\ &\qquad y = \mathit{image}_t(j, e) \ \wedge \ y > \delta \end{aligned} \tag{2}$$

- ...

Agent architecture in Jason



Trust evaluation in Jason

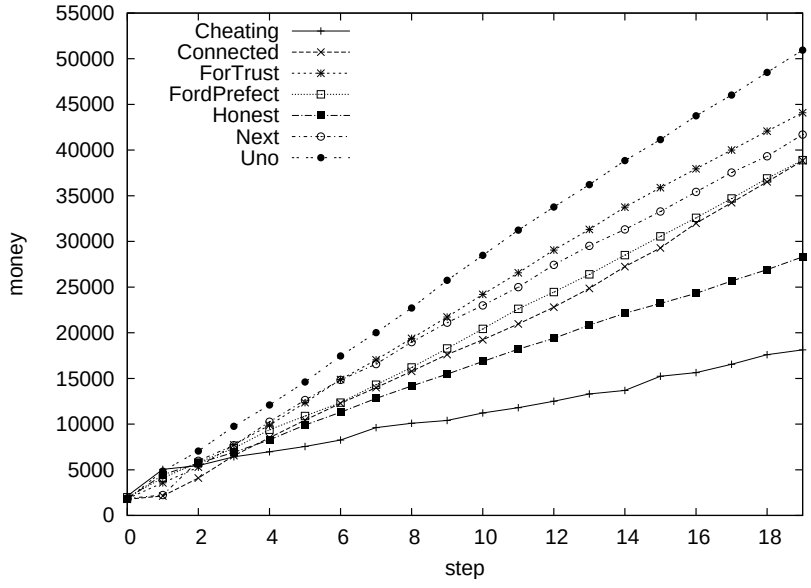
$$
\begin{aligned}
&trust(J, Action, Goal)[strength(C)] \ :- && // \ trust\ inference \\
&\quad .intend(Goal)\ \& && // \ I\ have\ the\ goal \\
&\quad does(J, Action)[strength(X)]\ \& && // \ J\ is\ capable\ and\ intend \\
&\quad after(J, Action, Goal)[strength(Y)]\ \& && // \ he\ has\ the\ power \\
&\quad C = math.min(X, Y). && // \ strength\ of\ the\ trust
\end{aligned}
$$

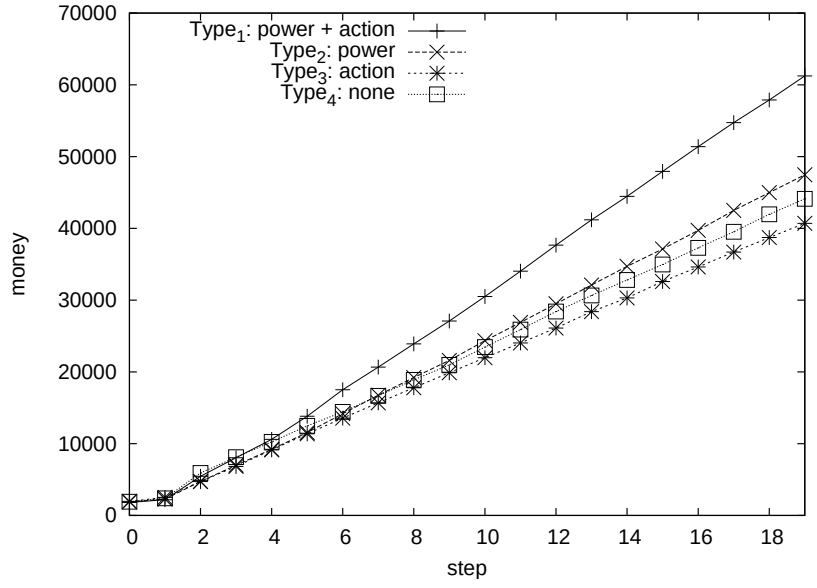
$$
+painting(Era, P) \leftarrow !good\_eval(P).
$$

$$
\begin{aligned}
&does(J, appraise(P))[strength(X)] :- \\
&\quad opinions\_count(J, Asked, Provided)\ \& \\
&\quad Asked > 0\ \& \\
&\quad X = Provided/Asked\ \&X > 0.9.
\end{aligned}
$$

$$
\begin{aligned}
&after(J, appraise(P), \_)[strength(Y)] \ :- \\
&\quad sincere(J)\ \&painting(Era, P)\ \& \\
&\quad image(J, Era, Y)\ \&Y > 0.5.
\end{aligned}
\tag{3}
$$

Experiments





Conclusion & Future Works

- Study of the components of C & F theory
- Proposition of a logical formalization
  - compliant to the properties of C & F
  - extended to reputation
- A first implemented prototype
  - logical formalization used as specifications
  - use of MAS language (Jason) and platform (ART)
- Future works
  - improvement and generalization of the implementation
  - application to security systems