An analytical decomposition of trust in terms of mental attitudes
(work in progress)

Robert Demolombe\textsuperscript{1}

\textsuperscript{1}Institut de Recherche en Informatique de Toulouse

June 2012
Assumption (à la Castelfranchi et al.): the truster believes that \textbf{if} he has some particular goal, \textbf{then} this goal will be reached

Analysis of possible supports for trust:

- Empirical
  - truster’s observations
  - information from trusted sources
- Analytical
  - trust in something can be supported by trust in other things
Objective of this work: systematic analysis of the possible social relationships between the truster and the trustees

Method:

- formalization in modal logic
- no analysis of the logical properties of the involved modalities: belief, action, intention, obligation, ...
- trust has the form of conditional properties
Modalities and operators

\[ \phi \Rightarrow \psi : \phi \text{ entails } \psi \]
\[ Bel_i \phi : i \text{ believes } \phi \]
\[ Goal_i \phi : i \text{'s goal is } \phi \]
\[ Attempt_i \phi : i \text{ attempts to bring it about that } \phi \]
\[ Int_i \phi : i \text{'s intention is to bring it about that } \phi \]
\[ Obg \phi : \text{it is obligatory that } \phi \]
\[ Forb \phi : \text{it is forbidden that } \phi \]
\[ Forb \phi \overset{\text{def}}{=} Obg \neg \phi \]
\[ Ask_{i,j} \phi : i \text{ asks } j \text{ to bring it about that } \phi \]
\[ Commit_i (\phi \mid \psi) : i \text{ has committed himself to bring it about that } \phi \]
\[ \text{if } \psi \text{ holds} \]
\[ \square \phi : \phi \text{ holds now and always in the future} \]
\[ \Diamond \phi : \phi \text{ will hold at some moment in the future} \]
\[ \Diamond \phi \overset{\text{def}}{=} \neg \square \neg \phi \]
Logical properties

Bel_i obeys a K system

Properties of the conditionals (sufficient)

(EQUIV) Si ⊢ φ ↔ φ' et ⊢ ψ ↔ ψ', alors ⊢ (φ → ψ) → (φ' → ψ')

(PROPG) (φ_1 ∧ φ_2 ⇒ φ_3) → (φ_1 ∧ φ_2 ⇒ φ_1 ∧ φ_3)

(TRANS) (φ_1 ⇒ φ_2) ∧ (φ_2 ⇒ φ_3) → (φ_1 ⇒ φ_3)

(DIST) (φ_1 ∧ (φ_1 ⇒ φ_2) ∧ ψ) ⇒ (φ_2 ∧ ψ)

(MATIMP) (φ ⇒ ψ) → (φ → ψ)
Initial form of trust

*i’s goal: to reach a state of affairs*

Example: *i* believes that if he has no cash ($\neg \phi$) and his goal is to get cash ($\Diamond \phi$), then he will get cash ($\Diamond \phi$)

(F1) $Bel_i(\neg \phi \land Goal; \Diamond \phi \Rightarrow \Diamond \phi)$
Initial form of trust

i’s goal: to reach a state of affairs
Example: i believes that if he has no cash (¬φ) and his goal is to get cash (◊φ), then he will get cash (◊φ)
(F1) Bel; (¬φ ∧ Goal; ◊φ ⇒ ◊φ)

i’s goal: to maintain a state of affairs
Example: i believes that if his car works well (φ) and his goal is that it still works well (□φ), then it still works well (□φ)
(M1) Bel; (φ ∧ Goal; □φ ⇒ □φ)
Analysis of trust support
to reach
If there is some agent $j$ who holds some property $Prop(j, \phi)$ such that:
(F2) $Bel_i((\neg \phi \land Goal_i \diamond \phi \Rightarrow \exists j Prop(j, \phi)) \land (\exists j Prop(j, \phi) \Rightarrow \diamond \phi))$
(F2) is a support for (F1)
(F1) $Bel_i((\neg \phi \land Goal_i \diamond \phi \Rightarrow \diamond \phi)$
because (F2) entails (F1)
Analysis of trust support
to reach
If there is some agent $j$ who holds some property $Prop(j, \phi)$ such that:
(F2) $Bel_i((\neg \phi \land Goal_i \land \phi \Rightarrow \exists j Prop(j, \phi)) \land (\exists j Prop(j, \phi) \Rightarrow \Diamond \phi))$
(F2) is a support for (F1)
(F1) $Bel_i(\neg \phi \land Goal_i \land \phi \Rightarrow \Diamond \phi)$
because (F2) entails (F1)
to maintain
If there is no agent who holds some property $Prop'(j, \phi)$ such that:
(M2) $Bel_i((\phi \land Goal_i \land \Box \phi \Rightarrow \neg \exists j Prop'(j, \phi)) \land (\neg \exists j Prop'(j, \phi) \Rightarrow \Box \phi))$
(M2) is a support for (M1)
i’s assumption: the only way to change $\phi$ is that there is an agent $j$
such that $Prop'(j, \phi)$
Trust in ability

(F2) \( Bel_i((\neg \phi \land Goal_i \Diamond \phi \Rightarrow \exists j Prop(j, \phi)) \land (\exists j Prop(j, \phi) \Rightarrow \Diamond \phi)) \)

Ability

\( Able_j \phi \overset{\text{def}}{=} \text{Attempt}_j \phi \Rightarrow \Diamond \phi \)

Trust in ability: \( Bel_i \text{Able}_j \phi \)

\( Prop_1(j, \phi) \overset{\text{def}}{=} \text{Attempt}_j \phi \land (\text{Attempt}_j \phi \Rightarrow \Diamond \phi) \)
Trust in ability

(F2) $Bel_i((\neg \phi \land Goal_i \Diamond \phi \Rightarrow \exists j Prop(j, \phi)) \land (\exists j Prop(j, \phi) \Rightarrow \Diamond \phi))$

Ability

$Able_j \phi \overset{\text{def}}{=} Attempt_j \phi \Rightarrow \Diamond \phi$

Trust in ability: $Bel_i Able_j \phi$

$Prop_1(j, \phi) \overset{\text{def}}{=} Attempt_j \phi \land (Attempt_j \phi \Rightarrow \Diamond \phi)$

$i$’s belief:

$\exists j (Attempt_j \phi \land Able_j \phi) \Rightarrow \Diamond \phi$

If $Prop(j, \phi)$ is $Prop_1(j, \phi)$, (F2) holds
Trust in ability

\[(M2)\]
\[\text{Bel}_i((\phi \land \text{Goal}_i \Box \phi \Rightarrow \neg \exists j \text{Prop}'(j, \phi)) \land (\neg \exists j \text{Prop}'(j, \phi) \Rightarrow \Box \phi))]\]

Ability

\[\text{Prop}'_1(j, \phi) \overset{\text{def}}{=} \text{Attempt}_j \neg \phi \land (\text{Attempt}_j \neg \phi \Rightarrow \Diamond \neg \phi)\]
Trust in ability

(M2)
\[ Bel_i((\phi \land \text{Goal}_i \square \phi \Rightarrow \neg \exists j \text{Prop'}(j, \phi)) \land (\neg \exists j \text{Prop'}(j, \phi) \Rightarrow \square \phi)) \]

Ability

\[ \text{Prop'}_1(j, \phi) \overset{\text{def}}{=} \text{Attempt}_j \neg \phi \land (\text{Attempt}_j \neg \phi \Rightarrow \Diamond \neg \phi) \]

Logical properties:
\[ \vdash \exists j \text{Prop'}_1(j, \phi) \rightarrow \Diamond \neg \phi \]
\[ \vdash \square \phi \rightarrow \neg \exists j \text{Prop'}_1(j, \phi) \]

i’s belief:
\[ \neg \exists j(\text{Attempt}_j \neg \phi \land \text{Able}_j \neg \phi) \Rightarrow \square \phi \]

If \text{Prop'}(j, \phi) is \text{Prop'}_1(j, \phi), (M2) holds
Active

$j$’s intention triggers $j$’s action

\[ \text{Active}_j \phi \overset{\text{def}}{=} \text{Int}_j \phi \Rightarrow \text{Attempt}_j \phi \]

\[ \text{Prop}_2(j, \phi) \overset{\text{def}}{=} \text{Int}_j \phi \land (\text{Int}_j \phi \Rightarrow \text{Attempt}_j \phi) \land \text{Able}_j \phi \]
Active

j’s intention triggers j’s action

\[ Active_j \phi \overset{\text{def}}{=} Int_j \phi \rightarrow \text{Attempt}_j \phi \]

\[ Prop_2(j, \phi) \overset{\text{def}}{=} Int_j \phi \land (Int_j \phi \Rightarrow \text{Attempt}_j \phi) \land \text{Able}_j \phi \]

Logical property:

\[ \vdash \exists j Prop_2(j, \phi) \rightarrow \Diamond \phi \]

i’\'s belief:

\[ \exists j (Int_j \phi \land Active_j \phi \land \text{Able}_j \phi) \Rightarrow \Diamond \phi \]
Active

\[ Prop'_2(j, \phi) \overset{\text{def}}{=} \text{Int}_j \neg \phi \land (\text{Int}_j \neg \phi \Rightarrow \text{Attempt}_j \neg \phi) \land \text{Able}_j \neg \phi \]

Logical property:
\[ \vdash \exists j Prop'_2(j, \phi) \rightarrow \Diamond \neg \phi \]

i’s belief:
\[ \neg \exists j (\text{Int}_j \neg \phi \land \text{Active}_j \neg \phi \land \text{Able}_j \neg \phi) \Rightarrow \Box \phi \]
Intention adoption
Norms fulfillment
To achieve
If $j$ believes that he is obliged to do something, then he intends to do that thing
\[
\text{Obey}_j \phi \overset{\text{def}}{=} \text{Bel}_j \text{OblInt}_j \phi \Rightarrow \text{Int}_j \phi
\]
\[
\text{Prop}_{3.1}(j, \phi) \overset{\text{def}}{=} \text{Bel}_j \text{OblInt}_j \phi \land (\text{Bel}_j \text{OblInt}_j \phi \Rightarrow \text{Int}_j \phi) \land \\
\text{Active}_j \phi \land \text{Able}_j \phi
\]
Intention adoption
Norms fulfillment
To achieve
If \( j \) believes that he is obliged to do something, then he intends to do that thing

\[
\text{Obey}_j \phi \overset{\text{def}}{=} \text{Bel}_j \text{OblInt}_j \phi \Rightarrow \text{Int}_j \phi
\]

\[
\text{Prop}_{3.1}(j, \phi) \overset{\text{def}}{=} \text{Bel}_j \text{OblInt}_j \phi \land (\text{Bel}_j \text{OblInt}_j \phi \Rightarrow \text{Int}_j \phi) \land \text{Active}_j \phi \land \text{Able}_j \phi
\]

Logical property:
\( \vdash \exists j \text{Prop}_{3.1}(j, \phi) \rightarrow \lozenge \phi \)

\( i \)'s belief:
\( \exists j (\text{OblInt}_j \phi \land \text{Obey}_j \phi \land \text{Active}_j \phi \land \text{Able}_j \phi) \Rightarrow \lozenge \phi \)
Norms fulfillment and institutional power

If $i$ asks $j$ to bring it about that $\phi$, then $j$ believes that it is obligatory that he adopts the intention to bring it about that $\phi$

Example: policeman $i$ asks $j$ to stop his car

$$\text{InstPower}_{i,j}\phi \overset{\text{def}}{=} \text{Ask}_{i,j}\phi \Rightarrow \text{Bel}_j\text{OblInt}_j\phi$$

$$\text{Prop}_{4.1}(j, \phi) \overset{\text{def}}{=} \text{Ask}_{i,j}\phi \land (\text{Ask}_{i,j}\phi \Rightarrow \text{Bel}_j\text{OblInt}_j\phi) \land \text{Obey}_j\phi \land \text{Active}_j\phi \land \text{Able}_j\phi$$
Norms fulfillment and institutional power
If \( i \) asks \( j \) to bring it about that \( \phi \), then \( j \) believes that it is obligatory that he adopts the intention to bring it about that \( \phi \)
Example: policeman \( i \) asks \( j \) to stop his car
\[
\text{InstPower}_{i,j}\phi \overset{\text{def}}{=} \text{Ask}_{i,j}\phi \Rightarrow \text{Bel}_j\text{OblInt}_j\phi
\]
\[
\text{Prop}_{4.1}(j,\phi) \overset{\text{def}}{=} \text{Ask}_{i,j}\phi \land (\text{Ask}_{i,j}\phi \Rightarrow \text{Bel}_j\text{OblInt}_j\phi) \land \text{Obe}_{j}\phi \land \text{Act}_{i}\phi \land \text{Able}_{j}\phi
\]
Logical property:
\[
\vdash \exists j\text{Prop}_{4.1}(j,\phi) \rightarrow \diamond \phi
\]
i’s belief:
\[
\exists j(\text{Ask}_{i,j}\phi \land \text{Obe}_{j}\phi \land \text{InstPower}_{i,j}\phi \land \text{Act}_{i}\phi \land \text{Able}_{j}\phi) \Rightarrow \diamond \phi
\]
Norms fulfillment
To maintain

\[ Prop'_{3.1}(j, \phi) \overset{\text{def}}{=} Bel_j \neg \text{ObgInt}_j \neg \phi \land (Bel_j \neg \text{ObgInt}_j \neg \phi \Rightarrow \text{Int}_j \neg \phi) \land \]
\[ \text{Active}_j \neg \phi \land \text{Able}_j \neg \phi \]
Norms fulfillment
To maintain
\[ \text{Prop}'_{3.1}(j, \phi) \overset{\text{def}}{=} \text{Bel}_j \text{OblInt}_j \neg \phi \land (\text{Bel}_j \text{OblInt}_j \neg \phi \Rightarrow \text{Int}_j \neg \phi) \land \text{Active}_j \neg \phi \land \text{Able}_j \neg \phi \]
Logical property:
\[ \vdash \exists j \text{Prop}'_{3.1}(j, \phi) \rightarrow \lozenge \neg \phi \]
\(i\)'s belief:
\[ \neg \exists j (\text{Bel}_j \text{OblInt}_j \neg \phi \land \text{Obey}_j \neg \phi \land \text{Active}_j \neg \phi \land \text{Able}_j \neg \phi) \Rightarrow \Box \phi \]
\[ \neg \exists j \text{Prop}'_{3.1}(j, \phi) : \text{no agent who fulfills the norms believes that} \]
\(\text{ObglInt}_j \neg \phi \)
\[ \text{Prop}''_{3.1}(j, \phi) \overset{\text{def}}{=} \text{Bel}_j \text{ForbInt}_j \neg \phi \land (\text{Bel}_j \text{ForbInt}_j \neg \phi \Rightarrow \text{Int}_j \neg \phi) \land \text{Active}_j \neg \phi \land \text{Able}_j \neg \phi \]
\[ \neg \exists j \text{Prop}'_{3.1}(j, \phi) : \text{no agent who violates the norms believes that} \]
\(\text{ForbInt}_j \neg \phi \)
Contract
To achieve
If $i$ asks to a taxi driver $j$ to commit himself to bring $i$ at the airport in a context where $i$ commits himself to pay the taxi driver, then the taxi driver $j$ adopts the intention to bring $i$ at the airport

$\text{Contract}_{i,j}(\phi, \psi) \overset{\text{def}}{=} \text{Ask}_{i,j}\text{Commit}_j(\phi \mid \text{Commit}_i\psi) \Rightarrow \text{Int}_j\phi$

$\text{Prop}_{3.2}(j, \phi) \overset{\text{def}}{=} \text{Ask}_{i,j}\text{Commit}_j(\phi \mid \text{Commit}_i\psi) \land$

$(\text{Ask}_{i,j}\text{Commit}_j(\phi \mid \text{Commit}_i\psi) \Rightarrow \text{Int}_j\phi) \land$

$\text{Active}_j\phi \land \text{Able}_j\phi$
**Contract**

**To achieve**

If $i$ asks to a taxi driver $j$ to commit himself to bring $i$ at the airport in a context where $i$ commits himself to pay the taxi driver, then the taxi driver $j$ adopts the intention to bring $i$ at the airport

$Contract_{i,j}(\phi, \psi) \overset{\text{def}}{=} \text{Ask}_{i,j}\text{Commit}_j(\phi \mid \text{Commit}_i\psi) \Rightarrow \text{Int}_j\phi$

$Prop_{3.2}(j, \phi) \overset{\text{def}}{=} \text{Ask}_{i,j}\text{Commit}_j(\phi \mid \text{Commit}_i\psi) \land$

$(\text{Ask}_{i,j}\text{Commit}_j(\phi \mid \text{Commit}_i\psi) \Rightarrow \text{Int}_j\phi) \land$

$Active_j\phi \land Able_j\phi$

Logical property:

$\vdash \exists j Prop_{3.2}(j, \phi) \rightarrow \Diamond \phi$

$i$’s belief:

$\exists j(\text{Ask}_{i,j}\text{Commit}_j(\phi \mid \text{Commit}_i\psi) \land Contract_{i,j}(\phi, \psi) \land Active_j\phi \land Able_j\phi) \Rightarrow \Diamond \phi$
To maintain

\[
\begin{align*}
\text{Prop}_3.2(j, \phi) & \overset{\text{def}}{=} \exists k (\text{Ask}_{j,k} \text{Commit}_k (\neg \phi \mid \text{Commit}_j \psi) \land \\
& (\text{Ask}_{j,k} \text{Commit}_k (\neg \phi \mid \text{Commit}_j \psi) \Rightarrow \text{Int}_k \neg \phi) \land \\
& \text{Active}_k \neg \phi \land \text{Able}_k \neg \phi
\end{align*}
\]

Example: i’s goal is not be killed and j asks to some mafia member to kill i
To maintain

\[ Prop'_{3.2}(j, \phi) \overset{\text{def}}{=} \exists k (\text{Ask}_{j,k} \text{Commit}_k(\neg \phi \mid \text{Commit}_j \psi) \land \text{Active}_k \neg \phi \land \text{Able}_k \neg \phi) \]

Example: \( i \)'s goal is not be killed and \( j \) asks to some mafia member to kill \( i \)

Logical property:

\[ \vdash \exists j Prop'_{3.2}(j, \phi) \rightarrow \Diamond \neg \phi \]

\( i \)'s belief:

\[ \neg \exists j Prop'_{3.2}(j, \phi) \rightarrow \Box \phi \]
Altruism
If $j$ believes that $i$’s goal is $\phi$, then $j$ adopts the intention to bring it about that $\phi$
Example: $i$ is an old man who wants to find some help to cross the road and $j$ is aware of $i$’s goal

$Altruis_{j,i}\phi \overset{\text{def}}{=} Bel_j(\neg \phi \land \text{Goal}_i \diamond \phi) \Rightarrow Int_j \phi$

$Prop_{3.3}(j, \phi) \overset{\text{def}}{=} Bel_j(\neg \phi \land \text{Goal}_i \diamond \phi) \land$

$(Bel_j(\neg \phi \land \text{Goal}_i \diamond \phi) \Rightarrow Int_j \phi) \land$

$Active_{j} \phi \land Able_{j} \phi$
Altruism

If $j$ believes that $i$’s goal is $\phi$, then $j$ adopts the intention to bring it about that $\phi$

Example: $i$ is an old man who wants to find some help to cross the road and $j$ is aware of $i$’s goal

$Altruis_{ij}\phi \overset{\text{def}}{=} Bel_j(\neg \phi \land \text{Goal}_i \Diamond \phi) \Rightarrow Int_j \phi$

$Prop_{3.3}(j, \phi) \overset{\text{def}}{=} Bel_j(\neg \phi \land \text{Goal}_i \Diamond \phi) \land$

$(Bel_j(\neg \phi \land \text{Goal}_i \Diamond \phi) \Rightarrow Int_j \phi) \land$

$Active_j \phi \land Able_j \phi$

Logical property:

$\vdash \exists j Prop_{3.3}(j, \phi) \rightarrow \Diamond \phi$

$i$’s belief:

$\exists j (Bel_j(\neg \phi \land \text{Goal}_i \Diamond \phi) \land Altruis_{ij}\phi \land Active_j \phi \land Able_j \phi) \Rightarrow \Diamond \phi$

$i$ believes that if there is some altruist agent, his goal $\phi$ will be reached
Perversion

\[ Pervert_{j,i}\phi \stackrel{\text{def}}{=} Bel_j(\neg\phi \land \text{Goal}_i \diamond \phi) \Rightarrow Int_j \neg\phi \]

\[ j \text{ intends to prevent } i \text{ to reach his goal} \]

\[ Prop'_{3.3}(j, \phi) \stackrel{\text{def}}{=} Bel_j(\neg\phi \land \text{Goal}_i \diamond \phi) \land \]

\[ (Bel_j(\neg\phi \land \text{Goal}_i \diamond \phi) \Rightarrow Int_j \neg\phi) \land \]

\[ \text{Active}_j \neg\phi \land \text{Able}_j \neg\phi \]
Perversion

\[ \text{Pervert}_{j,i} \phi \overset{\text{def}}{=} \text{Bel}_j(\neg \phi \land \text{Goal}_i \diamond \phi) \Rightarrow \text{Int}_j \neg \phi \]

\( j \) intends to prevent \( i \) to reach his goal

\[ \text{Prop}^{'}_{3.3}(j, \phi) \overset{\text{def}}{=} \text{Bel}_j(\neg \phi \land \text{Goal}_i \diamond \phi) \land \]
\[ (\text{Bel}_j(\neg \phi \land \text{Goal}_i \diamond \phi) \Rightarrow \text{Int}_j \neg \phi) \land \]

\( \text{Active}_j \neg \phi \land \text{Able}_j \neg \phi \)

Logical property:

\[ \vdash \exists j \text{Prop}^{'}_{3.3}(j, \phi) \rightarrow \diamond \neg \phi \]

\( i \)'s belief:

\[ \neg \exists j (\text{Bel}_j(\neg \phi \land \text{Goal}_i \diamond \phi) \land \text{Pervert}_{j,i} \phi \land \text{Active}_j \neg \phi \land \text{Able}_j \neg \phi) \Rightarrow \diamond \phi \]

\( i \) believes that if there is no pervert agent the state of affairs \( \phi \) will be maintained
Summary

Initial trust definition:

- to reach a state of affairs
- to maintain a state of affairs

Analysis of possible trust supports:

- Action and ability
- Intention and "activeness"
  - Obligation and obedience (norms fulfillment)
  - Request and contract fulfillment (mutual interest)
  - Altruism

Duality: to reach vs to maintain

- there is a "good" agent
- there is no "bad" agent
Formalization
The proposed axiomatic of conditionals is sufficient to prove that each property $Prop_1$, $Prop_2$, $Prop_3$, ... logically entails $Prop$