Delegation and mental states

Nicolas Troquard (IRIT/LOA)
joint work with Cristiano Castelfranchi, Emiliano Lorini (ISTC-CNR)
and Andreas Herzig (IRIT-CNRS)

ILIKS – Trento – December, 1st
Present account of Intention

Cohen & Levesque’s formalization of Bratman’s theory

\[ AGoal_i^{CL} \phi \overset{\text{def}}{=} \text{Pref}_i F \phi \land \text{Bel}_i \neg \phi \]

\[ PGoal_i^{CL} \phi \overset{\text{def}}{=} AGoal_i^{CL} \phi \land (\text{Bel}_i \phi \lor \text{Bel}_i \text{G}\phi) \text{Before} \neg \text{Pref}_i F \phi \]

\[ Int_i^{CL} \phi \overset{\text{def}}{=} PGoal_i \phi \land \text{Pref}_i F \exists i: \alpha (i: \alpha) \phi \]

Problems:

- too strong definition: e.g. in cooperative contexts, intentions cannot entail to build plans triggering other agents’ actions
- too weak definition: e.g. intention of trivialities
What can logic of agency do for us?

- theories of agency: causal connection between action and goal
  - Kanger, Pörn and col.
  - Belnap, Horyt, Chellas et col.: seeing to it that (STIT)

- objective: combine C&L approach with STIT operator, for a logical theory of *intention* and its application to *delegation*
A logic of agency, belief and preference (semantics)

\[ M = \langle \text{Mom}, <, \text{ATM}, \text{AGT}, \text{Choice}, \text{Belief}, \text{Preference}, \nu \rangle \]

- \( \langle \text{Mom}, < \rangle = \text{branching-time, discrete structure} \)
  - history = maximal \(<\)-ordered subset of \text{Mom}
  - \(\text{Hist} = \text{set of all histories}\)
  - \(H_w = \text{set of histories passing through } w\)
  - \(\text{Ctxt} \overset{\text{def}}{=} \{m/h \mid w \in \text{Mom}, h \in H_w\} = \text{set of contexts}\)
A logic of agency, belief and preference (semantics)

\( M = \langle \text{Mom}, <, \text{ATM}, \text{AGT}, \text{Choice}, \text{Belief}, \text{Preference}, v \rangle \)

- \( \langle \text{Mom}, < \rangle = \text{branching-time, discrete structure} \)
  - history = maximal \(<\)-ordered subset of \( \text{Mom} \)
  - \( \text{Hist} \) = set of all histories
  - \( H_w \) = set of histories passing through \( w \)
  - \( \text{Ctx} \overset{\text{def}}{=} \{ m/h \mid w \in \text{Mom}, h \in H_w \} = \text{set of contexts} \)

- \( \text{Choice} : 2^{\text{AGT}} \times \text{Mom} \rightarrow 2^{2^{\text{Hist}}} \)
  - \( \text{Choice}_w^a(h) = a's \text{ particular at moment } w \text{ choice containing history } h \)
A logic of agency, belief and preference (semantics)

\[ M = \langle \text{Mom}, <, \text{ATM}, \text{AGT}, \text{Choice}, \text{Belief}, \text{Preference}, v \rangle \]

- \( \langle \text{Mom}, < \rangle = \text{branching-time, discrete structure} \)
  - history = maximal \(<\)-ordered subset of \text{Mom}
  - Hist = set of all histories
  - \( H_w \) = set of histories passing through \( w \)
  - \( Ctxt \overset{\text{def}}{=} \{ m/h \mid w \in \text{Mom}, h \in H_w \} \) = set of contexts

- \text{Choice} : \( 2^{\text{AGT}} \times \text{Mom} \rightarrow 2^{2^{\text{Hist}}} \)
  - \( \text{Choice}^w_a(h) = a's \text{ particular at moment } w \text{ choice containing history } h \)

- \text{Belief}_i \subseteq Ctxt \times Ctxt
A logic of agency, belief and preference (semantics)

\[ M = \langle \text{Mom}, <, \text{ATM}, \text{AGT}, \text{Choice}, \text{Belief}, \text{Preference}, v \rangle \]

- \( \langle \text{Mom}, < \rangle = \text{branching-time, discrete structure} \)
  - history = maximal \(<\)-ordered subset of \text{Mom}
  - \( \text{Hist} = \text{set of all histories} \)
  - \( H_w = \text{set of histories passing through } w \)
  - \( \text{Ctx} \overset{\text{def}}{=} \{ m/h \mid w \in \text{Mom}, h \in H_w \} = \text{set of contexts} \)

- \text{Choice} : 2^{\text{AGT}} \times \text{Mom} \rightarrow 2^{2^{\text{Hist}}}
  - \( \text{Choice}_a^w(h) = a's \text{ particular at moment } w \text{ choice containing history } h \)

- \( \text{Belief}_i \subseteq \text{Ctx} \times \text{Ctx} \)
- \( \text{Preference}_i \subseteq \text{Ctx} \times \text{Ctx} \)
agents’ choices are always compatible
  ▶ at least one common history to each possible combination of agent’s choices
  ▶ for groups: $\text{Choice}_j^w(h) = \bigcap_{i \in J} \text{Choice}_i^w(h) \neq \emptyset$

Belief$_i$ and Preference$_i$
  ▶ serial, transitive and euclidean
  ▶ $\text{Preference}_i \subseteq \text{Belief}_i$ (realism)
  ▶ if $w \text{Belief}_i w'$ then $\text{Preference}_i(w) = \text{Preference}_i(w')$ (introspection)
Semantics of operators

- $M, w/h \models □\phi$ iff $M, w/h' \models \phi$ for all $h' \in H_w$
- $M, w/h \models Stit_j\phi$ iff $M, w/h' \models \phi$ for every $h' \in Choice^w_j(h)$
Semantics of operators

- $M, w/h \models \square \phi$ iff $M, w/h' \models \phi$ for all $h' \in H_w$
- $M, w/h \models \text{Stit}_J \phi$ iff $M, w/h' \models \phi$ for every $h' \in \text{Choice}_J^w(h)$
- $M, w/h \models \text{Bel}_i \phi$ iff $M, w'/h' \models \phi$ for every $w'/h' \in \text{Belief}_i(w/h)$
- $M, w/h \models \text{Pref}_i \phi$ iff $M, w'/h' \models \phi$ for every $w'/h' \in \text{Preference}_i(w/h)$
Semantics of operators

- \( M, w/h \models \Box \phi \) iff \( M, w/h' \models \phi \) for all \( h' \in H_w \)
- \( M, w/h \models \text{Stit}_j \phi \) iff \( M, w/h' \models \phi \) for every \( h' \in \text{Choice}^w_j(h) \)
- \( M, w/h \models \text{Bel}_i \phi \) iff \( M, w'/h' \models \phi \) for every \( w'/h' \in \text{Belief}_i(w/h) \)
- \( M, w/h \models \text{Pref}_i \phi \) iff \( M, w'/h' \models \phi \) for every \( w'/h' \in \text{Preference}_i(w/h) \)
- \( M, m/h \models X \phi \) iff \( M, w'/h \models \phi \), \( w' \) immediate successor of \( w \) in history \( h \)
  
  - \( G \phi \) = "from now on, \( \phi \) always true on this history"
  - \( F \phi \overset{\text{def}}{=} \neg G \neg \phi \) = "\( \phi \) is true at some future point on this history"
Some validities

(Stit) \quad \text{S5 axioms for } Stit_J

(Box) \quad \text{S5 axioms for } □

(BoxStit) \quad □φ \to Stit_iφ

(Monotony) \quad Stit_Iφ \to Stit_Jφ, \text{ for } I \subseteq J

(LTL) \quad \text{axioms of LTL}

(Bel/Pref) \quad \text{KD45 axioms for } Bel_i \text{ and } Pref_i

(Inclusion) \quad Bel_iφ \to Pref_iφ

(Pos. introspection) \quad Pref_iφ \to Bel_iPref_iφ

(Neg. introspection) \quad \neg Pref_iφ \to Bel_i\neg Pref_iφ
Future directed intention to be

- \( AGoal_i \phi \overset{\text{def}}{=} Pref_i F \phi \land \neg Bel_i \phi \)
  - C&L’s negative condition was \( Bel_i \neg \phi \)

**Definition**

\[
Int_i \phi \overset{\text{def}}{=} AGoal_i \phi \land Bel_i \neg Stit_{AGT \setminus \{i\}} F \phi
\]

- \( i \) has the achievement goal that \( \phi \)
- \( i \) believes that \( \phi \) will not be achieved without \( i \)'s intervention
  - *dependence clause*
Properties of intention

- $\text{Int}_i \phi \land \text{Int}_i \neg \phi$ is satisfiable
  - future-directed intentions: *indeterminate* moment in the future

- $\text{Indep}(\phi, i) \overset{\text{def}}{=} \phi \rightarrow \text{Stit}_{\text{AGT}} \{i\} \phi$
  - $\models \text{Bel}_i \text{Indep}(F \phi, i) \land \text{Int}_i \phi \rightarrow \bot$

- $\text{Veto}(i, j, \phi) \overset{\text{def}}{=} \neg \diamond \text{Stit}_{\text{AGT}} \{i\} F \phi \land \text{AGoal}_j \phi$
  - $\models \text{Bel}_i \text{Veto}(i, i, \phi) \rightarrow \text{Int}_i \phi$

- intentions to believe persist (under *no forgetting* for Pref)
  - $\models \text{Int}_i \text{Bel}_i \phi \rightarrow X (\text{Bel}_i \phi \lor \text{Int}_i \text{Bel}_i \phi \lor \neg \text{Bel}_i \neg \text{Stit}_{\text{AGT}} \{i\} F \text{Bel}_i \phi)$
we take inspiration from goal-based theory of Falcone & Castelfranchi (1998)
- logical modeling purpose: some slight differences
- weak delegation
- mild delegation
- strict delegation (contracts, explicit agreement)

we focus on two notions of delegation
- **passive**: Gabriela expects her flatmate the task of cleaning the bathroom
- **active**: Gabriela forces her flatmate to clean the bathroom
Passive delegation

Definition

\[
\text{PassiveDel}(i, j, \phi) \overset{\text{def}}{=} \neg \text{Bel}_i \phi \land \text{Pref}_i F \text{Stit}_j \phi \land \neg \text{Bel}_i \neg \text{Stit}_{\text{AGT}\{i\}} F \text{Stit}_j \phi
\]

- \(i\) does not believe \(\phi\) is already achieved
- \(i\) prefers to achieve \(\phi\) by exploiting \(j\)
- according to \(i\)’s beliefs, it is possible that there will be a moment where \(j\) will ensure \(\phi\), independently of what \(i\) does now
Passive delegation

Definition

\[
\text{PassiveDel}(i, j, \phi) \overset{\text{def}}{=} \\
\neg \text{Bel}_i \phi \land \text{Pref}_i F \text{Stit}_j \phi \land \neg \text{Bel}_i \neg \text{Stit}_{AGT \setminus \{i\}} F \text{Stit}_j \phi
\]

- \(i\) does not believe \(\phi\) is already achieved
- \(i\) prefers to achieve \(\phi\) by exploiting \(j\)
- according to \(i\)'s beliefs, it is possible that there will be a moment where \(j\) will ensure \(\phi\), independently of what \(i\) does now
Passive delegation

**Definition**

\[
\text{PassiveDel}(i, j, \phi) \overset{\text{def}}{=} \\
\neg \text{Bel}_i \phi \land \text{Pref}_i F \text{Stit}_j \phi \land \neg \text{Bel}_i \neg \text{Stit}_{\text{AGT} \setminus \{i\}} F \text{Stit}_j \phi
\]

- \(i\) does not believe \(\phi\) is already achieved
- \(i\) prefers to achieve \(\phi\) by exploiting \(j\)
- according to \(i\)'s beliefs, it is possible that there will be a moment where \(j\) will ensure \(\phi\), independently of what \(i\) does now
Properties of Passive Delegation

- \( \vdash \text{PassiveDel}(i, j, \phi) \land \text{Int}_{i}\phi \rightarrow \bot \)
  - passive delegation and intention are incompatible

- \( \vdash \text{PassiveDel}(i, j, \phi) \land \text{Int}_i\text{Stit}_j\phi \rightarrow \bot \)
Active delegation

Definition

\[ ActiveDel(i, j, \phi) \overset{\text{def}}{=} \neg Bel_i \phi \land Pref_i F Stit_j \phi \land Bel_i \neg Stit_{AGT\{i\}} F Stit_j \phi \land \neg Bel_i F Stit_{AGT\{j\}} \phi \]

- \( i \) does not believe that \( \phi \) is already achieved
- \( i \) prefers to achieve to achieve \( \phi \) by exploiting \( j \)
- \( i \) believes that \( j \) will not achieve \( \phi \) independently of \( i \)'s intervention
- \( i \) does not believe that the future achievement of \( \phi \) will be independent of \( j \)'s future choices
Active delegation

**Definition**

\[
\text{ActiveDel}(i, j, \phi) \overset{\text{def}}{=} \neg \text{Bel}_i \phi \land \text{Pref}_i \text{F Stit}_j \phi \land \text{Bel}_i \neg \text{Stit}_{AGT \setminus \{i\}} \text{F Stit}_j \phi \land \neg \text{Bel}_i \text{F Stit}_{AGT \setminus \{j\}} \phi
\]

- \(i\) does not believe that \(\phi\) is already achieved
- \(i\) prefers to achieve to achieve \(\phi\) by exploiting \(j\)
- \(i\) believes that \(j\) will not achieve \(\phi\) independently of \(i\)’s intervention
- \(i\) does not believe that the future achievement of \(\phi\) will be independent of \(j\)’s future choices
Active delegation

Definition

\[ \text{ActiveDel}(i, j, \phi) \overset{\text{def}}{=} \neg \text{Bel}_i \phi \land \text{Pref}_i F \text{Stit}_j \phi \land \text{Bel}_i \neg \text{Stit}_{\text{AGT}\setminus \{i\}} F \text{Stit}_j \phi \land \neg \text{Bel}_i F \text{Stit}_{\text{AGT}\setminus \{j\}} \phi \]

- \( i \) does not believe that \( \phi \) is already achieved
- \( i \) prefers to achieve to achieve \( \phi \) by exploiting \( j \)
- \( i \) believes that \( j \) will not achieve \( \phi \) independently of \( i \)’s intervention
- \( i \) does not believe that the future achievement of \( \phi \) will be independent of \( j \)’s future choices
Active delegation

**Definition**

\[
\text{ActiveDel}(i, j, \phi) \overset{\text{def}}{=} \neg \text{Bel}_i \phi \land \text{Pref}_i F \text{Stit}_j \phi \land \text{Bel}_i \neg \text{Stit}_{\text{AGT}\{i\}} F \text{Stit}_j \phi \land \neg \text{Bel}_i F \text{Stit}_{\text{AGT}\{j\}} \phi
\]

- **i** does not believe that \( \phi \) is already achieved
- **i** prefers to achieve to achieve \( \phi \) by exploiting \( j \)
- **i** believes that \( j \) will not achieve \( \phi \) independently of \( i \)’s intervention
- **i** does not believe that the future achievement of \( \phi \) will be independent of \( j \)’s future choices
Properties of Active Delegation

- $\vdash ActiveDel(i, j, \phi) \rightarrow Int_i Stit_j \phi$
  - $i$ actively delegates the achievement of $\phi$ to $j$ only if $i$ has the intention that $j$ achieves $\phi$

- $\vdash Bel_i Stit_{AGT \setminus \{i\}} F Stit_k \phi \rightarrow \neg ActiveDel(i, j, \phi) \quad k \neq j$
  - $i$ cannot actively delegate the achievement of his goal that $\phi$ to agent $j$ when he believes that agent $k$ will see to it that $\phi$ independently from what agent $i$ actually does
Conclusion and perspectives

- Just a general specification
- Towards collective intentionality