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

- $AGoal_i^{CL}\phi \stackrel{\text{def}}{=} Pref_iF\phi \wedge Bel_i\neg\phi$
- $PGoal_i^{CL}\phi \stackrel{\text{def}}{=} AGoal_i^{CL}\phi \wedge (Bel_i\phi \vee Bel_iG\phi) Before \neg Pref_iF\phi$
- $Int_i^{CL}\phi \stackrel{\text{def}}{=} PGoal_i\phi \wedge Pref_iF\exists i{:}\alpha\langle i{:}\alpha\rangle\phi$

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, Horty, 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)

$M = \langle Mom, <, ATM, AGT, Choice, Belief, Preference, v \rangle$

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

A logic of agency, belief and preference (semantics)

$M = \langle \mathit{Mom}, <, \mathit{ATM}, \mathit{AGT}, \mathit{Choice}, \mathit{Belief}, \mathit{Preference}, v \rangle$

- $\langle \mathit{Mom}, < \rangle$ = *branching-time*, *discrete* structure
  - history = maximal $<$-ordered subset of *Mom*
  - *Hist* = set of all histories
  - H_w = set of histories passing through w
  - $\mathit{Ctxt} \stackrel{\text{def}}{=} \{m/h \mid w \in \mathit{Mom}, h \in H_w\}$ = set of contexts
- $\mathit{Choice} : 2^{\mathit{AGT}} \times \mathit{Mom} \longrightarrow 2^{2^{\mathit{Hist}}}$
  - $\mathit{Choice}_a^w(h)$ = a's particular at moment w choice containing history h

A logic of agency, belief and preference (semantics)

$M = \langle Mom, <, ATM, AGT, Choice, Belief, Preference, v \rangle$

- $\langle Mom, < \rangle$ = *branching-time*, *discrete* structure
  - history = maximal $<$-ordered subset of *Mom*
  - *Hist* = set of all histories
  - H_w = set of histories passing through w
  - $Ctxt \stackrel{\text{def}}{=} \{m/h \mid w \in Mom, h \in H_w\}$ = set of contexts
- $Choice : 2^{AGT} \times Mom \longrightarrow 2^{2^{Hist}}$
  - $Choice_a^w(h)$ = a's particular at moment w choice containing history h
- $Belief_i \subseteq Ctxt \times Ctxt$

A logic of agency, belief and preference (semantics)

$M = \langle Mom, <, ATM, AGT, Choice, Belief, Preference, v \rangle$

- $\langle Mom, < \rangle$ = *branching-time*, *discrete* structure
  - history = maximal $<$-ordered subset of *Mom*
  - *Hist* = set of all histories
  - H_w = set of histories passing through w
  - $Ctxt \stackrel{\text{def}}{=} \{m/h \mid w \in Mom, h \in H_w\}$ = set of contexts
- $Choice : 2^{AGT} \times Mom \longrightarrow 2^{2^{Hist}}$
  - $Choice_a^w(h)$ = a's particular at moment w choice containing history h
- $Belief_i \subseteq Ctxt \times Ctxt$
- $Preference_i \subseteq Ctxt \times Ctxt$

A logic of agency, belief and preference (semantics ctd)

- agents' choices are always compatible
  - at least one common history to each possible combination of agent's choices
  - for groups: $Choice_J^w(h) = \bigcap_{i \in J} Choice_i^w(h) \neq \emptyset$
- $Belief_i$ and $Preference_i$
  - serial, transitive and euclidean
  - $Preference_i \subseteq Belief_i$ (**realism**)
  - if $wBelief_i w'$ then $Preference_i(w) = Preference_i(w')$ (**introspection**)

Semantics of operators

- $M, w/h \models \Box\phi$ iff $M, w/h' \models \phi$ for all $h' \in H_w$
- $M, w/h \models Stit_J\phi$ iff $M, w/h' \models \phi$ for every $h' \in Choice^w_J(h)$

Semantics of operators

- $M, w/h \models \Box\phi$ iff $M, w/h' \models \phi$ for all $h' \in H_w$
- $M, w/h \models Stit_J\phi$ iff $M, w/h' \models \phi$ for every $h' \in Choice_J^w(h)$

- $M, w/h \models Bel_i\phi$ iff $M, w'/h' \models \phi$ for every $w'/h' \in Belief_i(w/h)$
- $M, w/h \models Pref_i\phi$ iff $M, w'/h' \models \phi$ for every $w'/h' \in Preference_i(w/h)$

Semantics of operators

- $M, w/h \models \Box\phi$ iff $M, w/h' \models \phi$ for all $h' \in H_w$
- $M, w/h \models Stit_J\phi$ iff $M, w/h' \models \phi$ for every $h' \in Choice_J^w(h)$

- $M, w/h \models Bel_i\phi$ iff $M, w'/h' \models \phi$ for every $w'/h' \in Belief_i(w/h)$
- $M, w/h \models Pref_i\phi$ iff $M, w'/h' \models \phi$ for every $w'/h' \in Preference_i(w/h)$

- $M, m/h \models X\phi$ iff $M, w'/h \models \phi$, w' immediate successor of w in history h
  - $G\phi$ = "from now on, ϕ always true *on this history*"
  - $F\phi \stackrel{\text{def}}{=} \neg G \neg\phi$ = "ϕ is true at some future point *on this history*"

Some validities

(Stit)	S5 axioms for $Stit_J$
(Box)	S5 axioms for $\Box$
(BoxStit)	$\Box\phi \rightarrow Stit_i\phi$
(Monotony)	$Stit_I\phi \rightarrow Stit_J\phi$, for $I \subseteq J$
(LTL)	axioms of LTL
(Bel/Pref)	KD45 axioms for Bel_i and $Pref_i$
(Inclusion)	$Bel_i\phi \rightarrow Pref_i\phi$
(Pos. introspection)	$Pref_i\phi \rightarrow Bel_i Pref_i\phi$
(Neg. introspection)	$\neg Pref_i\phi \rightarrow Bel_i\neg Pref_i\phi$

Future directed intention to be

- $AGoal_i \phi \stackrel{\text{def}}{=} Pref_i F \phi \wedge \neg Bel_i \phi$
  - C&L's negative condition was $Bel_i \neg \phi$

Definition

$Int_i \phi \stackrel{\text{def}}{=} AGoal_i \phi \wedge Bel_i \neg Stit_{AGT \setminus \{i\}} F \phi$

- i has the achievement goal that ϕ
- i believes that ϕ will not be achieved without i's intervention
  - *dependence clause*

Properties of intention

- $Int_i\phi \wedge Int_i\neg\phi$ is satisfiable
  - future-directed intentions: *indeterminate* moment in the future
- $Indep(\phi, i) \stackrel{\mathrm{def}}{=} \phi \rightarrow Stit_{AGT\setminus\{i\}}\phi$
  - $\models Bel_i Indep(F\phi, i) \wedge Int_i\phi \rightarrow \bot$
- $Veto(i, j, \phi) \stackrel{\mathrm{def}}{=} \neg\Diamond Stit_{AGT\setminus\{i\}}F\phi \wedge AGoal_j\phi$
  - $\models Bel_i Veto(i, i, \phi) \rightarrow Int_i\phi$
- intentions to believe persist (under *no forgetting* for Pref)
  - $\models Int_i Bel_i\phi \rightarrow X(Bel_i\phi \vee Int_i Bel_i\phi \vee \neg Bel_i\neg Stit_{AGT\setminus\{i\}}F Bel_i\phi)$

Delegation

- 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

$$PassiveDel(i,j,\phi) \stackrel{\text{def}}{=}$$

$$\neg Bel_i\phi \wedge Pref_i F\, Stit_j\phi \wedge \neg Bel_i \neg Stit_{AGT\setminus\{i\}} F\, Stit_j\phi$$

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

Passive delegation

Definition

$$PassiveDel(i, j, \phi) \stackrel{\text{def}}{=}$$

$$\neg Bel_i \phi \wedge Pref_i F\, Stit_j \phi \wedge \neg Bel_i \neg Stit_{AGT \setminus \{i\}} F\, Stit_j \phi$$

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

Passive delegation

Definition

$$PassiveDel(i, j, \phi) \stackrel{\text{def}}{=}$$

$$\neg Bel_i \phi \wedge Pref_i F\,Stit_j \phi \wedge \neg Bel_i \neg Stit_{AGT \setminus \{i\}} F\,Stit_j \phi$$

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

Properties of Passive Delegation

- $\models \mathit{PassiveDel}(i, j, \phi) \wedge \mathit{Int}_i \phi \rightarrow \bot$
  - passive delegation and intention are incompatible
- $\models \mathit{PassiveDel}(i, j, \phi) \wedge \mathit{Int}_i \mathit{Stit}_j \phi \rightarrow \bot$

Active delegation

Definition

$ActiveDel(i, j, \phi) \stackrel{\text{def}}{=}$
$\neg Bel_i \phi \wedge Pref_i F Stit_j \phi \wedge Bel_i \neg Stit_{AGT \setminus \{i\}} F Stit_j \phi \wedge \neg Bel_i F Stit_{AGT \setminus \{j\}} \phi$

- ***i* does not believe that ϕ is already achieved**
- *i* prefers to achieve to achieve ϕ by exploiting *j*
- *i* believes that *j* will not achieve ϕ independently of *i*'s intervention
- *i* does not believe that the future achievement of ϕ will be independent of *j*'s future choices

Active delegation

Definition

$ActiveDel(i, j, \phi) \stackrel{\text{def}}{=}$
$\neg Bel_i \phi \wedge Pref_i F Stit_j \phi \wedge Bel_i \neg Stit_{AGT \setminus \{i\}} F Stit_j \phi \wedge \neg Bel_i F Stit_{AGT \setminus \{j\}} \phi$

- *i* does not believe that ϕ is already achieved
- *i* **prefers to achieve to achieve ϕ by exploiting** *j*
- *i* believes that *j* will not achieve ϕ independently of *i*'s intervention
- *i* does not believe that the future achievement of ϕ will be independent of *j*'s future choices

Active delegation

Definition

$ActiveDel(i, j, \phi) \stackrel{\text{def}}{=}$
$\neg Bel_i \phi \wedge Pref_i F Stit_j \phi \wedge Bel_i \neg Stit_{AGT \setminus \{i\}} F Stit_j \phi \wedge \neg Bel_i F Stit_{AGT \setminus \{j\}} \phi$

- i does not believe that ϕ is already achieved
- i prefers to achieve to achieve ϕ by exploiting j
- **i believes that j will not achieve ϕ independently of i's intervention**
- i does not believe that the future achievement of ϕ will be independent of j's future choices

Active delegation

Definition

$ActiveDel(i, j, \phi) \stackrel{\text{def}}{=}$
$\neg Bel_i \phi \wedge Pref_i F\,Stit_j \phi \wedge Bel_i \neg Stit_{AGT \setminus \{i\}} F\,Stit_j \phi \wedge \neg Bel_i F\,Stit_{AGT \setminus \{j\}} \phi$

- i does not believe that ϕ is already achieved
- i prefers to achieve to achieve ϕ by exploiting j
- i believes that j will not achieve ϕ independently of i's intervention
- **i does not believe that the future achievement of ϕ will be independent of j's future choices**

Properties of Active Delegation

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

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

Conclusion and perspectives

- Just a general specification
- Towards collective intentionality