ILIKS Interdisciplinary Laboratory on Interacting Knowledge Systems

Agents and Mental Attitudes Session

chairs: Andreas Herzig & Alessandro Oltramari

Trento, 30 nov. 2006

ILIKS scientific programme for the “agents and mental attitudes” area

- understand the nature of agents’ mind and mental attitudes, their coherence and dynamics
  - ontology of mind, mental states and agency
  - logics of belief and knowledge
  - belief revision
  - models of intentions, goals, plans and commitments

➔ synergies between cognitive, social, philosophical and logic approaches

Session overview

- Introduction and summary of work done

  A. Herzig

- Sample joint work. ***Action, attempt and intention.***

  E. Lorini, A. Herzig, C. Castelfranchi

- Position talk. ***The role of beliefs in goal dynamics: prolegomena to a constructive theory of intentions.***

  F. Paglieri

- Discussion

  A. Oltramari et all

Agents and Mental Attitudes: Overview of the ILIKS Area

Andreas Herzig

ILIKS meeting, Trento, 30 nov. 2006

A common view on agents

- motivations
  - agent-oriented software engineering (U. Trento, DIS)
    - modeling of organizations
    - importance of agents' goals
  - modeling of cognitive processes (ISTC-IAMCI)
    - interaction between desires and goals
    - generation of intentions
  - ontology of interaction (ISTC-LOA)
  - logics of interaction (IRIT)
    - Bratman / Cohen&Levesque theory of intention

➔ BDI-like agent model

Mental attitudes: connection to other areas

- action area
  - logics combining agency and knowledge
  - logics combining actions and belief
  - logic of intentional action

- ‘social’ area
  - agents’ goals in organizations
  - group belief
  - common belief

➔ activities not easy to separate

Common work within ILIKS

- definition of intention
  - based on STIT logic [Herzig et al., Torun’06]
  - applied to strong and weak delegation

    ➔ Nicolas Troquard’s talk in the ‘social’ area tomorrow
  - intention generation

    ➔ Fabio Paglieri’s talk

- intentions, goals, and their relation to action
  - via the notion of attempt [Lorini,Herzig&Castelfranchi, JELIA’06]

    ➔ next talk

Parallel work within ILIKS (1)

- ontology of mental states & attitudes [Ferrario&Oltramari, FOIS’04]
  - intentional agents (intentional stance)
  - aboutness of mental attitudes
- goals of agents (‘actors’) in organizations
  - dependence of agent1 on agent2 w.r.t. goal1 (cf. Castelfranchi’s notion of dependence)
  - agent-oriented software engineering → TROPOS methodology [Mylopolos, Giorgini, Massacci et col.]

Parallel work within ILIKS (2)

- individual vs. group beliefs
  - group attitudes [Tummolini]
  - formalization of Tuomela's group belief [Gaudou, Herzig&Longin, KR'06]
  - application to agent communication languages [Gaudou, Herzig, Longin&Nickles, ECAI'06]

- belief revision
  - foundations [Paglieri]
  - relation with argumentation [Paglieri&Castelfranchi, 06]
  - relation with doxastic logic [Aucher, JANCL'07; Laverny&Lang, 05, 06; Ditmarsch et al., ongoing]

Parallel work within ILIKS (3)

- formalization of emotions
  - surprise [Lorini&Castelfranchi, 07]
    - based on expectations; propositional logic + probability; abduction to the best explanation
    - the more improbable the input, the greater the surprise
    - relation with belief revision
  - [Oltramari, 06]
    - ➔ discussion part of this session
  - Ortony, Clore&Collins's theory (OCC) in a BDI logic
    - ➔ …

Formalization of OCC emotions in a BDI logic

[Adam et al., AIMSA’06]

- based on Ortony, Clore&Collins’ theory [1988]
  - standard reference in AI and agent community
- 3 classes of stimuli ➔ 3 kinds of appraisals
  - event ➔ agreement
    - joy, hope, fear, disappointment, fear-confirmed, …, gloating, …
  - action ➔ approval
    - pride, shame, admiration, …
  - object ➔ attraction
    - … *(not considered)*

OCC: 12 event-triggered emotions

- event stimulus
  - **joy** = being pleased about a desirable event
  - ***distress*** *= being displeased about an undesirable event*
  - **hope** = being pleased about the prospect of a desirable event
  - ***fear*** *= being displeased about the prospect of an undesirable event*
  - **satisfaction** = being pleased about the confirmation of the prospect of a desirable event
  - ***fear-confirmed*** *= being displeased about the confirmation of the prospect of an undesirable event*
  - **relief** = being pleased about the disconfirmation of the prospect of an undesirable event
  - ***disappointment*** *= being displeased about the disconfirmation of the prospect of a desirable event*
  - **happy-for** = being pleased about an event presumed to be desirable for someone else
  - ***pity*** *= being displeased about an event presumed to be undesirable for someone else*
  - **gloating** = being pleased about an event presumed to be undesirable for someone else
  - ***resentment*** *= being displeased about an event presumed to be desirable for someone else*

OCC: 8 action-triggered emotions

- action stimulus
  - **pride** = approving of one's own praiseworthy action
  - ***shame*** *= disapproving of one's own blameworthy action*
  - **admiration** = approving of someone else's praiseworthy action
  - ***reproach*** *= disapproving of someone else's blameworthy action*
  - **gratitude** = approval of an agent's action + pleasure at the desirable outcome
  - ***anger*** *= disapproval of an agent's action + displeasure at the undesirable outcome*
  - **gratification** = approval of one's own action + pleasure at the desirable outcome
  - ***remorse*** *= disapproval of one's own action + displeasure at the undesirable outcome*

- object stimulus
  - …

OCC in a BDI logic: event-triggered emotions

- Well-being emotions

  $\mathrm{Joy}_i\ A\ =\ \mathrm{Bel}_i\ A\ \&\ \mathrm{Des}_i\ A$

  $\mathit{Sadness}_i\ A\ =\ \mathit{Bel}_i\ A\ \&\ \mathit{Des}_i\ {\sim}A$

- Prospect-based emotions

  $\mathrm{Hope}_i\ A\ =\ \mathrm{Prob}_i\ \mathbf{F}\ A\ \&\ {\sim}\mathrm{Bel}_i\ \mathbf{F}\ A\ \&\ \mathrm{Des}_i\ A$

  $\mathit{Fear}_i\ A\ =\ \mathit{Prob}_i\ \boldsymbol{F}\ A\ \&\ {\sim}\mathit{Bel}_i\ \boldsymbol{F}\ A\ \&\ \mathit{Des}_i\ {\sim}A$

- Confirmation emotions

  $\mathrm{Satisfaction}_i\ A\ =\ \mathrm{Bel}_i\ \mathbf{P}\ \mathrm{Expect}_i\ A\ \&\ \mathrm{Des}_i\ A\ \&\ \mathrm{Bel}_i\ A$

  $\mathit{FearConfirmed}_i\ A\ =\ \mathit{Bel}_i\ \boldsymbol{P}\ \mathit{Expect}_i\ A\ \&\ \mathit{Des}_i\ {\sim}A\ \&\ \mathit{Bel}_i\ A$

  $\mathrm{Relief}_i\ A\ =\ \mathrm{Bel}_i\ \mathbf{P}\ \mathrm{Expect}_i\ {\sim}A\ \&\ \mathrm{Des}_i\ {\sim}A\ \&\ \mathrm{Bel}_i\ A$

  $\mathit{Disappointment}_i\ A\ =\ \mathit{Bel}_i\ \boldsymbol{P}\ \mathit{Expect}_i\ {\sim}A\ \&\ \mathit{Des}_i\ {\sim}A\ \&\ \mathit{Bel}_i\ A$

- Fortunes-of-others emotions

  $\mathrm{HappyFor}_{i;j}\ A\ =\ \mathrm{Bel}_i\ A\ \&\ \mathrm{Bel}_i\ \mathbf{F}\ \mathrm{Bel}_j\ A\ \&\ \mathrm{Bel}_i\ \mathrm{Des}_j\ A\ \&\ \mathrm{Des}_i\ \mathrm{Bel}_j\ A$

  $\mathit{SorryFor}_{i;j}\ A\ =\ \mathit{Bel}_i\ A\ \&\ \mathit{Bel}_i\ \boldsymbol{F}\ \mathit{Bel}_j\ A\ \&\ \mathit{Bel}_i\ \mathit{Des}_j\ {\sim}A\ \&\ \mathit{Des}_i\ {\sim}\mathit{Bel}_j\ A$

  $\mathrm{Resentment}_{i;j}\ A\ =\ \mathrm{Beli}\ A\ \&\ \mathrm{Bel}_i\ \mathbf{F}\ \mathrm{Bel}_j\ A\ \&\ \mathrm{Bel}_i\ \mathrm{Des}_j\ A\ \&\ \mathrm{Des}_i\ {\sim}\mathrm{Bel}_j\ A$

  $\mathit{Gloating}_{i;j}\ A\ =\ \mathit{Bel}_i\ A\ \&\ \mathit{Bel}_i\ \boldsymbol{F}\ \mathit{Bel}_j\ A\ \&\ \mathit{Bel}_i\ \mathit{Des}_j\ {\sim}A\ \&\ \mathit{Des}_i\ \mathit{Bel}_j\ A$

OCC in a BDI logic: action-triggered emotions

- Attribution emotions

  Pride$_{i}$ (i:a) = Bel$_{i}$ Done$_{i:a}$ (~Prob$_{i}$ Happens$_{i:a}$ T & Bel$_{i}$ Idl$_{i}$ Happens$_{i:a}$ T)

  Shame$_{i}$ (i:a) = Bel$_{i}$ Done$_{i:a}$ (~Prob$_{i}$ Happens$_{i:a}$ T & Bel$_{i}$ Idl$_{i}$ ~Happens$_{i:a}$ T)

  Admiration$_{i;j}$ (j:a) = Bel$_{i}$ Done$_{j:a}$ (~Prob$_{i}$ Happens$_{j:a}$ T & Bel$_{i}$ Idl$_{j}$ Happens$_{j:a}$ T)

  Reproach$_{i;j}$ (j:a) = Bel$_{i}$ Done$_{j:a}$ (~Prob$_{i}$ Happens$_{j:a}$ T & Bel$_{i}$ Idl$_{j}$ ~Happens$_{i:a}$T)

- Composed emotions

  Gratification$_{i}$ (i:a; A) = Pride$_{i}$ (i:a) & Bel$_{i}$ Resp$_{i;i:a}$ A & Joy$_{i}$ A

  Remorse$_{i}$ (i:a; A) = Shame$_{i}$ (i:a) & Bel$_{i}$ Resp$_{i;i:a}$ A & Sadness$_{i}$ A

  Gratitude$_{i;j}$ (j:a; A) = Admiration$_{i;j}$ (j:a) & Bel$_{i}$ Resp$_{j;j:a}$ A & Joy$_{i}$ A

  Anger$_{i;j}$ (j:a; A) = Reproach$_{i;j}$ (j:a) & Bel$_{i}$ Resp$_{j;j:a}$ A & Sadness$_{i}$ A

Outlook

- role of emotions in rationality
- study interaction between desires and goals
  - generation of intentions
- rework belief revision in BDI framework
  - goal revision
  - relation with argumentation