“Individual and Collective Intentionality”

Introductory course @ ESSLLI 2009

Andreas Herzig and Emiliano Lorini
IRIT-CNRS, University of Toulouse, France
www.irit.fr/~Andreas.Herzig/Esslli09course
herzig@irit.fr, lorini@irit.fr

Bordeaux, July 2009

Course overview

Monday	epistemic logic and its dynamics
Tuesday	doxastic logic and its dynamics
Wednesday	logic of goals and intentions
Thursday	common belief, group belief and group acceptance
Friday	group action, group intention

Wednesday:
Logic of goals and intentions

Plan

- Introduction
- Basic notions
  - action and time
  - belief
  - choice
- Achievement goals, persistent goals, and intentions
- Criticisms and improvements

1. Which mental attitudes?

Basic pro-attitude: desire

- desires = agent's ideal states ('where do I want to be?')

- two cases:
  - beliefs = desires: ☺
  - beliefs ≠ desires: ☹
    ➔ do something!

From desire to action

naïve algorithm:

1. select some desire
2. check whether beliefs imply desire
3. generate plan
   - practical syllogism:

     ‘if you want *p*, and believe action *a* has effect *p*, then do *a*’
4. execute plan

Do desires directly drive our actions?

problems:

1. desires might be inconsistent: *marryAnn*$\wedge$*marryBeth*
   ➔ which should I pursue?
2. desires might be unrealistic = believed to be impossible to satisfy
   ➔ should I keep on trying?
3. the world changes
   ➔ is there a new opportunity to satisfy my desire?
4. agents' resources are bounded
   ➔ I can't replan all the time!

A 2nd pro-attitude: intentions [Bratman 1986]

different from desires:

- temporally stable, only abandoned when believed to be
  1. satisfied, or
  2. impossible, or
  3. superfluous because superior intention achieved

  ➔ *rational balance* with belief
- consistent
  - N.B.: intention to be *eventually* in Bordeaux and intention to be *eventually* in Frankfurt might be satisfied in sequence:

    first *goToBordeaux*, then *goToFrankfurt*
- directly related to planning and action

Kinds of intention

- ‘intention-to-be’ vs. ‘intention-to-do’
  - intention to *be in* Bordeaux ➔ proposition (state) (…exist?)
  - intention to *go to* Bordeaux ➔ action (transition)

- present-directed intention (pdi)
  vs. future-directed intention (fdi)
  - pdi causes immediate attempt to act: ‘switch to next slide’
  - fdi triggers plan generation: ‘go skiing in February’

- maintenance intention vs. achievement intention
  - ‘stay alive’ vs. ‘become rich’

Kinds of intention

- ‘intention-to-be’ vs. ‘intention-to-do’
  - intention to *be in* Bordeaux ➔ proposition (state) (…exist?)
  - intention to *go to* Bordeaux ➔ action (transition)

- present-directed intention (pdi)
  vs. future-directed intention (fdi)
  - pdi causes immediate attempt to act: ‘switch to next slide’
  - fdi triggers plan generation: ‘go skiing in February’

- maintenance intention vs. achievement intention
  - ‘stay alive’ vs. ‘become rich’

➔ here: future-directed achievement intentions-to-be

Intentions ➔ plans & actions

- intentions are high-level plans
- intentions trigger sub-intentions
- sub-intentions finally trigger actions

- ESSLLI approaching:
  1. *{goToBordeaux, goToFrankfurt}*
  2. *… , goToBordeaux ; … ; goToFrankfurt; …*
  3. *goToBordeaux := buyTicket ; takeMetro; takeTrain ; takeTram*
  4. *takeTram := findTramStation ; buyTicket ; getOnTram ; …*
  5. …

Desire ➔ intention

- “desire processing”

  [Paglieri&Castelfranchi, Synthese 2007]

  1. active desire
     - motivating belief
  2. pursuable desire
     - assessment belief (self-realizing? impossible?)
  3. chosen desire (intention)
     - determined by beliefs about costs, preferences, urgency, incompatibility
     - must be consistent
  4. present-directed intention
     - beliefs about preconditions, means-ends

A slogan: “BDI”

- relevant mental attitudes:
  belief + desire + intention = BDI [Bratman]
- successful in AI and multiagent systems:
  - BDI ‘frameworks’
  - BDI logics
  - BDI architectures
  - BDI agent programming languages

2. Major BDI frameworks

Existing BDI approaches

- Cohen&Levesque's reduction of intention to choice
  [Cohen&Levesque 87, 90, Sadek 92]
- Rao&Georgeff's branching-time logic of intention
  [Rao&Georgeff 91, 92, 98]
- KARO framework [Linder, Hoek&Meyer 94, 95]
- [Konolige&Pollack 93]
- [Wooldridge 02, Hoek&Wooldridge 03]

[C&L 1990]

Cohen, Philip R. and Levesque, Hector:

- "Persistence, Intentions, and Commitment"

  In *Intentions in Communication* (Ph. R. Cohen, J. Morgan, M. E. Pollack, eds.), MIT Press, 1990
- "Intention is choice with commitment"

  Artificial Intelligence 42:2-3 (1990)
  - *Influential Paper Award* at AAMAS'06
  - refer to philosophical theory of [Bratman 1987]
  - mandatory reference in AI and MAS papers on intention
  - … but few applied
    - only few 'direct successors': [Perrault 1990; Sadek 1991, 2000]

C&L theory: the bases

- mental attitudes: focus on belief and intention
  - intention non primitive: reduced to choice
    (while intention is primitive in most other approaches: Rao&Georgeff, Konolige&Pollack,…)
  - desires not in focus

- *change*: action and time

From C&L to a logic of intention

- C&L: formally complex
  - first-order
    - quantification over sequences of events
  - part of the semantics only
    - ‘assumptions’ without semantic counterpart

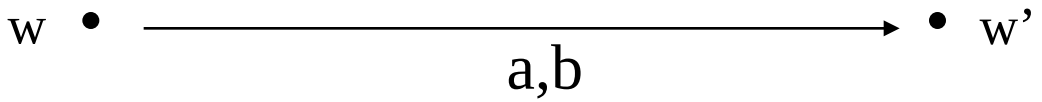
- last part of today’s course: sound and complete logic
  - minimalism: propositional modal logic
  - completeness

Plan

- Introduction
- Basic notions
  - action and time
  - belief
  - choice
- Achievement goals, persistent goals, and intentions
- Criticisms and improvements

1. Action and time

Transition systems



- set of possible worlds: *w, w', v,…*
- set of actions: *a, a_1, a_2,…, b, …*
- acc.rel. is a partial function R_a
  - $R_a(w)$ not defined = action does not happen
  - histories = sequences of states

Action: language

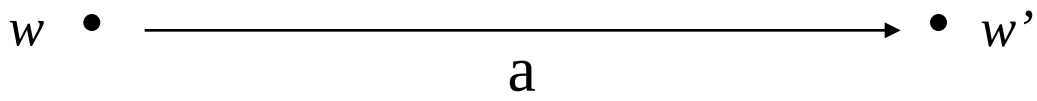
$\langle a \rangle p$ = "*a* is going to happen, and *p* will hold afterwards"

$[a]\, p$ = "if *a* happens then *p* will hold afterwards"

= $\neg \langle a \rangle \neg p$

- $\langle takeTrain \rangle \top$ = *takeTrain* is going to happen
- $[takeTrain] \bot$ = *takeTrain* is not going to happen

Action: semantics



- *w* |= <*a*>*p* iff $R_a(w)$ defined, and $R_a(w)$ |= *p*

Action: axiomatics

weak version of Dynamic Logic (modal logic K)

- standard principles for normal modal logics:
  - ...
- determinism:
  - $\langle a \rangle p \rightarrow [a]\, p$
- at most one transition:
  - $\langle a \rangle p \rightarrow [b]\, p$ (stronger than determinism)

Time

$G\ p$ = "p holds from now on"

- semantics: linear accessibility relation
  - $R_G(w)$ = the history starting from w
  - R_G reflexive, transitive, …
- $F\ p = \neg G \neg p$

Time: axiomatics

modal logic S4.3

- standard principles for normal modal logics:
  - ...
- reflexivity of time:
  - $p \rightarrow F\, p$
- transitivity of time:
  - $F\, p \rightarrow F\, F\, p$
- linearity of time:
  - $F\, p \wedge F\, q \rightarrow F\, (p \wedge Fq)\ v\ F\, (F\, p \wedge q)$
- axiom relating time and action:
  - $G\, p \rightarrow [a]\, p$

Time: the “Before” operator

p Before q = “*p* holds before *q*”

- semantics of LTL
  - *w* |= *p Before q* iff for every *w”* s.th. wR_Gw” & *w”* |= *q*
    there is *w’* s.th. $wR_Gw'R_Gw$” and *w’* |= *p*

- *p Before q* $\leftrightarrow$ $\neg(\neg p$ *Until q)*

2. Belief

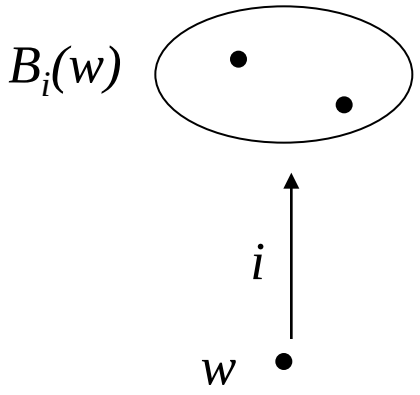
Belief: language

$Bel_i\ p$ = "agent *i* believes that *p*"

Bel_i <*takeTrain*>T =

"*i* believes *takeTrain* is going to happen"

Belief: semantics



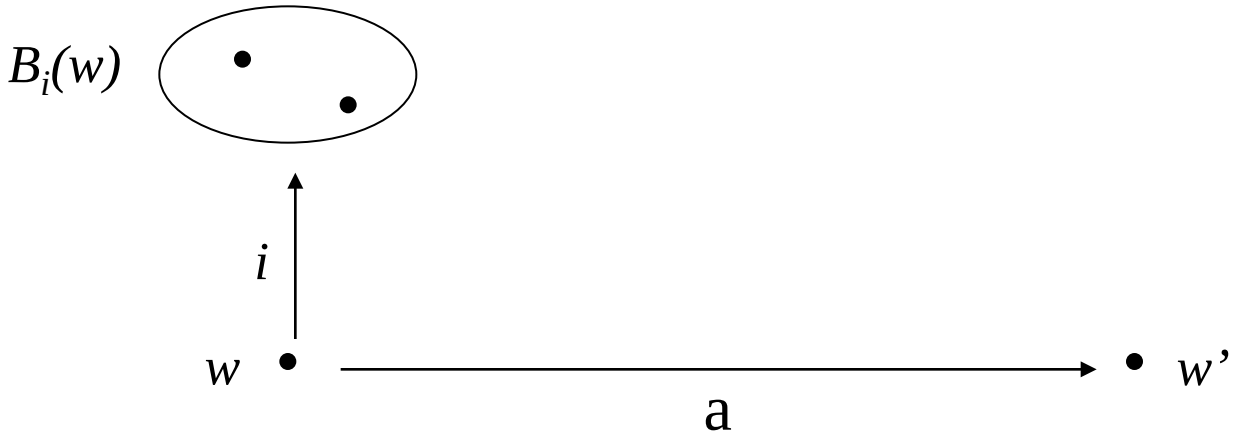
- belief state $B_i(w)$ = set of possible states
- $w \models Bel_i\ p$ iff for every $v \in B_i(w)$, $v \models p$

Belief: axiomatics

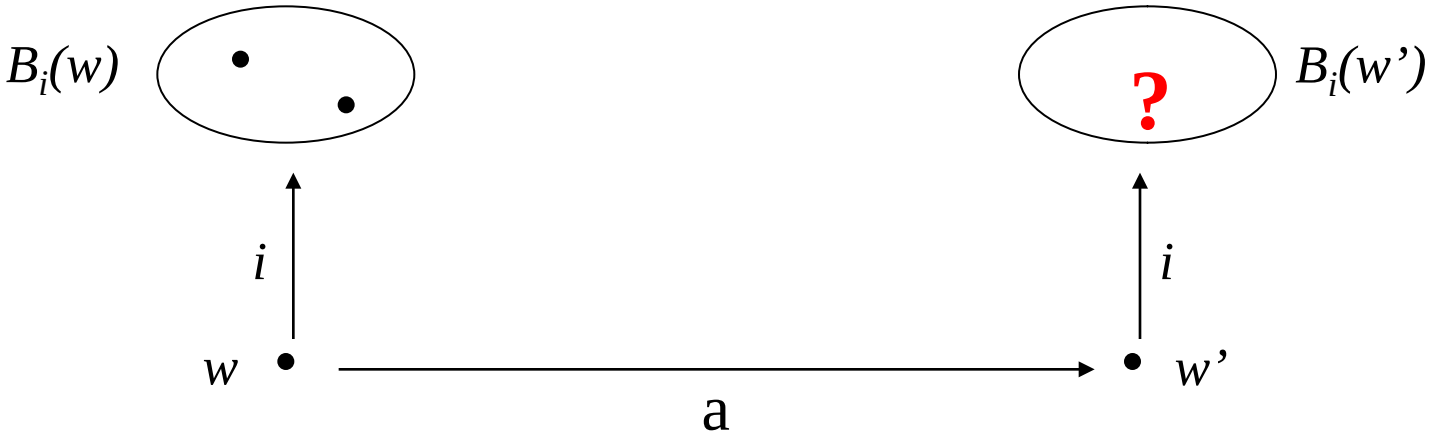
modal logic KD45

- standard principles for normal modal logics:
  - if p is a logical truth then $Bel_i\ p$ (omniscience)
  - ...
- consistency:
  - $Bel_i\ p \rightarrow \neg Bel_i \neg p$
- positive introspection:
  - $Bel_i\ p \rightarrow Bel_i\ Bel_i\ p$
- negative introspection:
  - $\neg Bel_i\ p \rightarrow Bel_i \neg Bel_i\ p$

Belief and action:



Belief and action: which interaction?



3. Choice

Motivation: choice as a base for intention

- C&L: intention reduced to belief, action, and choice (= chosen goal)

- idea: *i* intends that *p* =
  1. among states in $B_i(w)$, *i* prefers states where *Fp* holds;
  2. $\neg Bel_i\, p$: *p* has to be achieved;
  3. *i*'s choice persists until *p* is believed;
  4. *i* is prepared to act to make *p* true.

Choice: language

$Choice_i\ p$ = "agent *i* prefers that *p*"

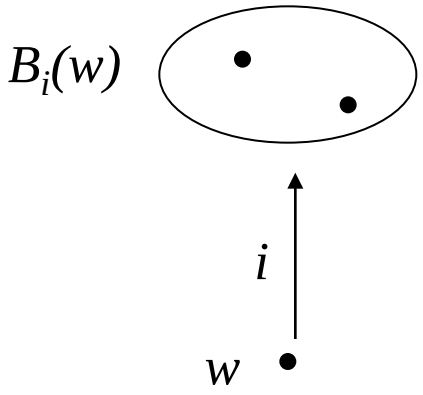
$Bel_i \neg Rich \ \curlywedge \ Choice_i\ F\ Rich$

$Bel_i \neg Rich \ \curlywedge \ Choice_i\ F\ Bel_i\ Rich$

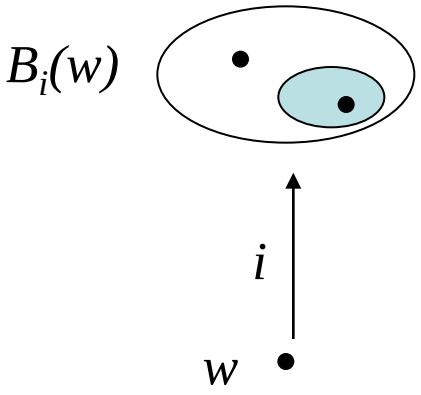
$Choice_i\ [takePlane]\ \bot$

- simple case of choices: no degrees, no orderings

Choice: semantics



Choice: semantics



- $C_i(w)$ = set of preferred states
  - subset of *i*'s belief state $B_i(w)$
- $w \models Choice_i\ p$

  iff *p* holds in all states preferred by *i*

  iff for every $v \in C_i(w)$, $v \models p$

Choice: axiomatics

modal logic KD45

- standard principles for normal modal logics:
  - ...
- consistency:
  - $Choice_i\, p \rightarrow \neg Choice_i \neg p$
- positive introspection:
  - $Choice_i\, p \rightarrow Choice_i\, Choice_i\, p$
- negative introspection:
  - $\neg Choice_i\, p \rightarrow Choice_i \neg Choice_i\, p$

Choice and belief: axiomatics

- realism:
  - $Bel_i\ p \rightarrow Choice_i\ p$
- introspection:
  - $Choice_i\ p \rightarrow Bel_i\ Choice_i\ p$
  - $\neg Choice_i\ p \rightarrow Bel_i \neg Choice_i\ p$

Where are we?

- standard multimodal logic of action, time, belief, and choice
- possible worlds semantics

t.b.d.: define intentions

➔ define persistent goals

➔ define achievement goals

Plan

- Introduction
- Basic notions
  - action and time
  - belief
  - choice
- Achievement goals, persistent goals, and intentions
- Criticisms and improvements

1. Achievement goal

Achievement goal: definition

$$AGoal_i\, p \;=\; \neg Bel_i\, p \wedge Choice_i\, F\, Bel_i\, p$$

- "s.th. must **happen** to obtain *p*"

- N.B.: original definition was

  $$AGoal_i^{C\&L}\, p \;=\; Bel_i\, \neg p \wedge Choice_i\, Fp$$

  ➔ $AGoal_i^{C\&L}(p \wedge \neg Bel_i\, p)$ consistent

  ➔ …but can never be believed to be achieved!

  ➔ will never be abandoned

  'epistemized': $AGoal_i\, p \;=\; AGoal_i^{C\&L}\, Bel_i\, p$

Achievement goal: derivable principles

- goal = goal to believe:
  - $AGoal_i\ p \leftrightarrow AGoal_i\ Bel_i\ p$ (≠ C&L)
- introspection:
  - $AGoal_i\ p \rightarrow Bel_i\ AGoal_i\ p$
  - $\neg AGoal_i\ p \rightarrow Bel_i\ \neg AGoal_i\ p$
- realism:
  - $\neg(AGoal_i\ p \wedge Bel_i\ G\ \neg p)$

2. Persistent goals

Persistent goals: 1st definition

$$PGoal_i\ p \ = \ AGoal_i\ p \wedge$$
$$((Bel_i\ p \vee Bel_i\ G\ \neg p)\ Before\ \neg AGoal_i\ p)$$

- if ever *p* is abandoned, then:
  - either believed to be satisfied,
  - or believed to be impossible

Persistent goals: 1st definition

$$PGoal_i\ p \ = \ AGoal_i\ p \wedge$$
$$((Bel_i\ p \vee Bel_i\ G\ \neg p)\ Before\ \neg AGoal_i\ p)$$

- too strong: I might abandon a goal if a superior goal is achieved (that had triggered the present goal in the past)

Persistent goals: 1st definition

$$PGoal_i\ p = AGoal_i\ p \wedge$$
$$((Bel_i\ p \vee Bel_i\ G\ \neg p)\ Before\ \neg AGoal_i\ p)$$

Persistent goals: 1st definition

$$PGoal_i\ p = AGoal_i\ p \wedge$$
$$((Bel_i\ p \vee Bel_i\ G \neg p\ ) \ Before \ \neg AGoal_i\ p)$$

Persistent goals: 2nd definition

$$PGoal_i\ p = AGoal_i\ p \wedge$$
$$((Bel_i\ p \vee Bel_i\ G \neg p \vee q)\ Before \neg AGoal_i\ p)$$

- unspecified side condition '*q*'

Persistent goals: properties

$PGoal_i\, p \rightarrow [a](PGoal_i\, p \vee Bel_i\, p \vee Bel_i\, G\, \neg p \vee q)$

- wanted: only abandon goal *p* if *a* is surprising
  - event *a* was possible ➔ stay with your Goals
  - event *a* is surprising ➔ goal revision
- v.i. …

3. Intention

Intentions: C&L's definition (roughly)

$$Intend_i\ p \;=\; PGoal_i\ p \wedge Choice_i\ F\ (\text{exists } i{:}a)\ \langle i{:}a \rangle\ Bel_i\ p$$

- *i:a* = action performed by agent *i*
- "s.th. must be **done** by *i* to make *p* believed"

- too strong: *i* may ask *j* to perform *a* [Sadek]
- too weak: lack of causal connection between *i:a* and *p*
  - $Intend_i$ *sunny* as soon as me taking breakfast occurs immediately before sunrise …

Intention: derivable principles

- extensionality
  - if $p \leftrightarrow q$ then $Intend_i\ p \leftrightarrow Intend_i\ q$
- rational balance = equilibrium between an agent's different mental attitudes
  - $Intend_i\ p \rightarrow \neg Bel_i\ p$
  - $Intend_i\ p \leftrightarrow Bel_i\ Intend_i\ p$
  - $\neg Intend_i\ p \leftrightarrow Bel_i\ \neg Intend_i\ p$

Intention: not derivable

- not derivable and unwanted:
  - $Intend_i\ p \rightarrow Bel_i \neg p$ (orig.C&L; too strong)
  - $Intend_i$ T (not to be achieved)
  - $Intend_i\ (p \wedge q) \rightarrow Intend_i\ p \wedge Intend_i\ q$ (take q = T)
  - $Intend_i\ p \wedge Intend_i\ q \rightarrow Intend_i\ (p \wedge q)$ (p and q at different time points)
  - $Intend_i\ p \rightarrow \neg Intend_i \neg p$ (v.s.)
  - $Intend_i\ p \wedge Bel_i\ (p \rightarrow q) \rightarrow Intend_i\ q$ ('side effect problem')
  - $Intend_i\ p \wedge G\ Bel_i\ (p \rightarrow q) \rightarrow Intend_i\ q$ ('side effect problem')

Intention to do

$$Intend\text{-}do_i\ a \ = \ Intend_i <a^{-1}>T$$

- intentions to do a =

  intention to be in a state

  where *a* has just been done

Maintenance intentions

$$Intend\text{-}maintain_i\ p \;=\; Bel_i\ p \wedge Intend_i\ G\ p$$

- maintenance intention that p =

  belief that *p* & intention that *p* be always true

Present-directed intention

$$PDI_i\ i{:}a \;=\; Choice_i\ \langle i{:}a \rangle T$$

- present-directed intention to do *a* =

  i's choice that *a* happen now

Plan

- Introduction
- Basic notions
  - action and time
  - belief
  - choice
- Achievement goals, persistent goals, and intentions
- Criticisms and improvements

1. Beyond C&L:

effects of events on beliefs and choices

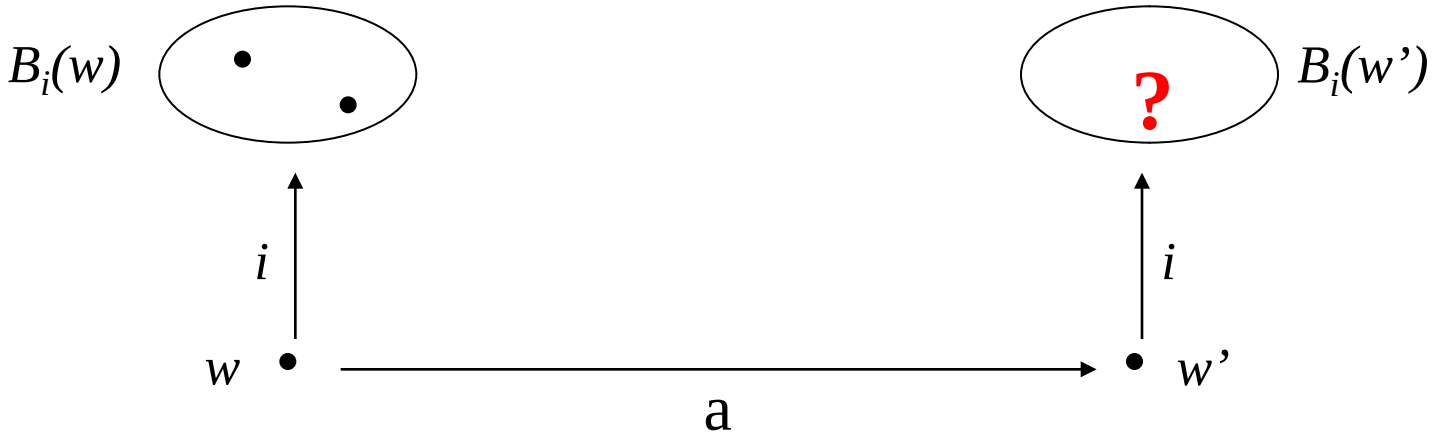
Belief and events: introduction

- no interaction in C&L's original paper

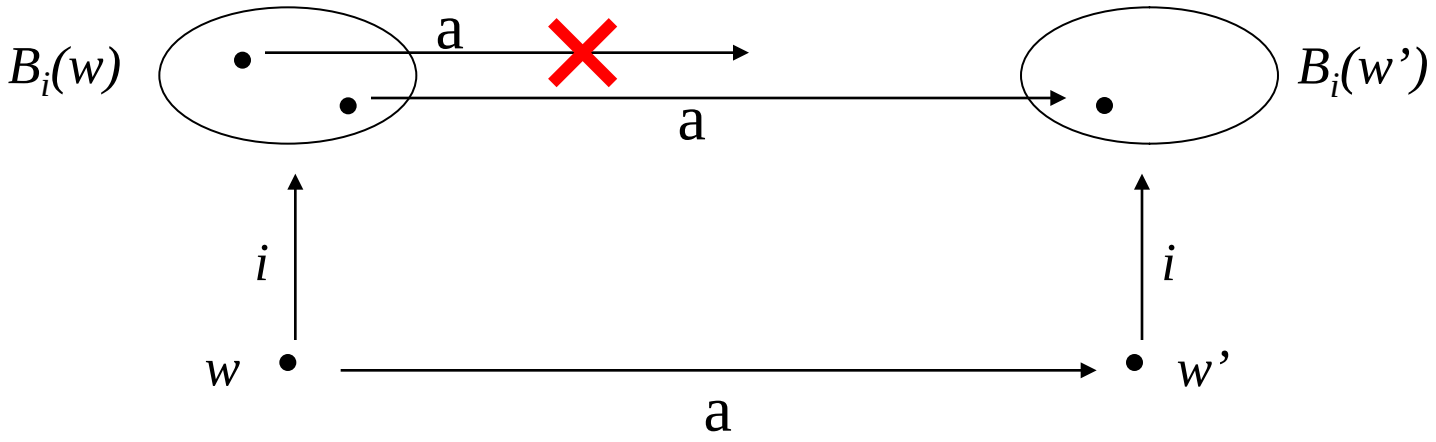
- here: combination with more recent results in dynamic epistemic logics DEL [Plaza, Gerbrandy, van Ditmarsch, Baltag, van Benthem, …]
  - simplest case: publicly announced events (cf. public announcement logic PAL)
  - benefit: principle of persistence of intentions provable ➔ no need to define PGoals

Belief and events: semantics

Belief and events: semantics

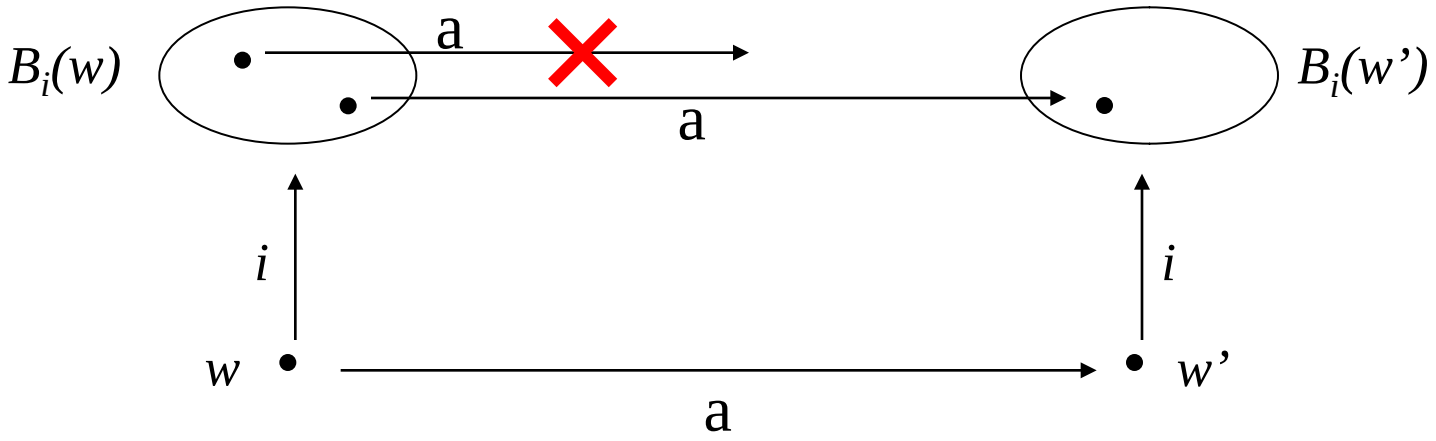


Belief and events: semantics



- $B_i(w') = (R_a \circ B_i)(w) = \bigcup_{v \in B_i(w)} R_a(v)$
  [Moore 85, Scherl&Levesque 03, Baltag et al. 99, …]

Belief and events: semantics



- hyp. (cf. DEL): *a* is public
- hyp. (cf. DEL): *i* only learns that *a* happened
  - i does not learn the outcome:
    - *a* is not *testif(p), informif(p)*
    - *a* might be *observe(p), inform(j,p)*
  - $B_i(w')$ only depends on R_a and $B_i(w)$, but not on *w*

Belief and events: axiomatics

- no learning
  - $[a]\, Bel_i\, p \wedge \neg [a]\bot \rightarrow Bel_i\, [a]\, p$

Belief and events: axiomatics

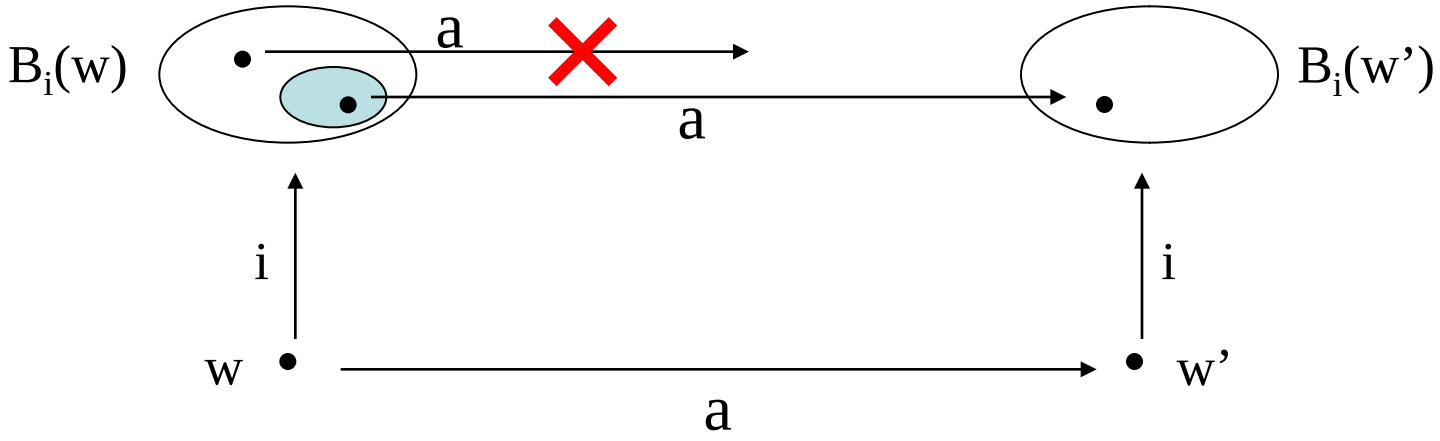
- no learning
  - $[a]\, Bel_i\, p \wedge \neg [a]\bot \rightarrow Bel_i\, [a]\, p$
- no forgetting
  - $Bel_i\, [a]\, p \wedge \neg Bel_i\, [a]\bot \rightarrow [a]\, Bel_i p$

Belief and events: axiomatics

- no learning
  - $[a]\, Bel_i\, p \wedge \neg [a]\bot \rightarrow Bel_i\, [a]\, p$
- no forgetting
  - $Bel_i\, [a]\, p \wedge \neg Bel_i\, [a]\bot \rightarrow [a]\, Bel_i p$

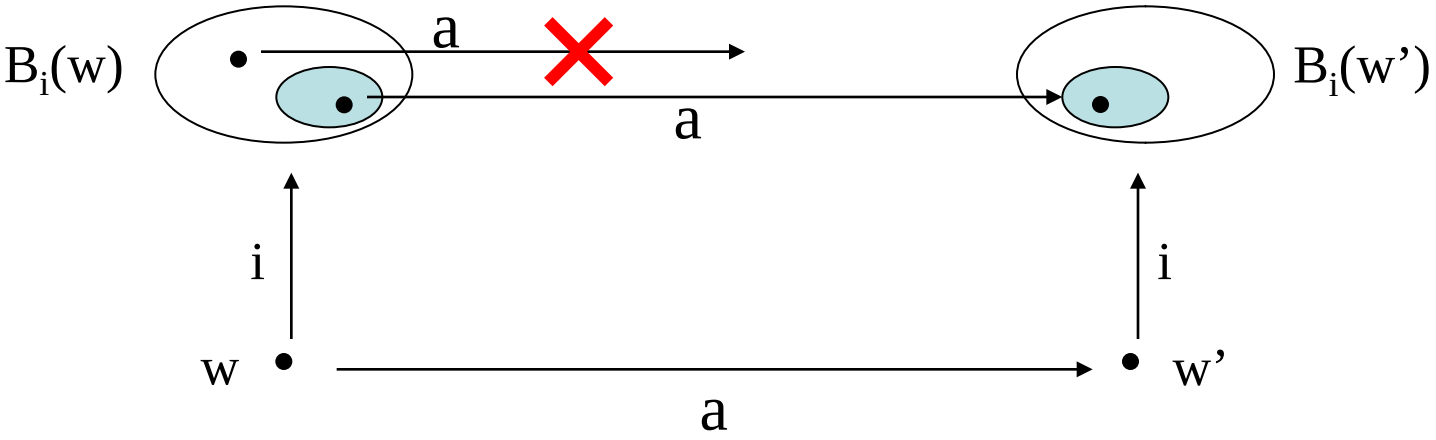
- almost reduction axioms (aka successor state axioms)
- case $Bel_i\, [a]\bot$: belief revision

Choice and events: semantics



- $B_i(w') = R_a(B_i(w))$

Choice and events: semantics



- $B_i(w') = R_a(B_i(w))$
- $C_i(w') = R_a(C_i(w))$

Choice and events: axiomatics

- no forgetting
  - $Choice_i\ [a]\ p \wedge \neg Choice_i\ [a]\bot \rightarrow [a]\ Choice_i\ p$
- no learning
  - ...
- intentional action
  - $\langle i{:}a \rangle \top \rightarrow Choice_i\ \langle i{:}a \rangle \top$

where *i* is the agent of i:*a*

- the case where $Choice_i\ [a]\bot$:
  - if $Bel_i\ [a]\bot$ then belief revision
  - if $\neg Bel_i\ [a]\bot$ then goal revision only

Choice and events: the case of attempts

- variant of logic of action:

  $\langle\langle i{:}a \rangle\rangle p$ = "*i* is going to try to do *a*, and *p* will hold afterwards"

  [Lorini&Herzig 08]

- axioms:
  - $\langle\langle i{:}a \rangle\rangle \top \rightarrow Choice_i \langle\langle i{:}a \rangle\rangle \top$
  - $\neg\langle\langle i{:}a \rangle\rangle \top \rightarrow Choice_i \neg\langle\langle i{:}a \rangle\rangle \top$
- provable:
  - $\langle\langle i{:}a \rangle\rangle p \leftrightarrow Choice_i \langle\langle i{:}a \rangle\rangle p$
- all other principles for successful action transfer

Where are we?

- standard multimodal logic of action, time, belief, and choice
- possible world semantics
  - no forgetting + no learning
  - close to product logics [Gabbay&Shehtman, Marx,…]
- deductive characterization
  - sound and complete (follows from [Sahlqvist 1972])
    - N.B.: no LTL *“Before”*
  - conjecture: EXPSPACE complete (cf. [Marx 2000])

2. Beyond C&L:

deriving persistence of achievement goals

Achievement goals persist!

$$AGoal_i\ p \wedge \neg Choice_i\ [a]\bot \rightarrow [a]\ (Bel_i\ p \vee AGoal_i\ p)$$

- $AGoal_i\ p \wedge \neg Choice_i\ [a]\bot \rightarrow [a]\ \neg Bel_i\ G\ \neg p$

Proof (uses 'no forgetting' for choice in step 3):

1. $Choice_i\ F\ Bel_i\ p \rightarrow Choice_i\ (Bel_i\ p \vee [a]\ F\ Bel_i\ p)$
2. $AGoal_i\ p \rightarrow Choice_i\ [a]\ F\ Bel_i\ p$
3. $Choice_i\ [a]\ F\ Bel_i\ p \wedge \neg\ Choice_i\ [a]\ \bot \rightarrow [a]\ Choice_i\ F\ Bel_i\ p$
4. $AGoal_i\ p \wedge \neg\ Choice_i\ [a]\ \bot \rightarrow [a]\ Choice_i\ F\ Bel_i\ p$
5. $[a]\ Choice_i\ F\ Bel_i\ p \rightarrow Bel_i\ p \vee (\neg Bel_i\ p \wedge [a]\ Choice_i\ F\ Bel_i\ p)$

- N.B.: if *i* is the agent of *a:*
  - $AGoal_i\ p \rightarrow [i:a]\ (Bel_i\ p \vee AGoal_i\ p)$

Persistence goals: differences w.r.t. C&L

C&L: $PGoal_i\, p \rightarrow [a]\,(Bel_i\, p \vee PGoal_i\, p \vee Bel_i\, G\, \neg p \vee q)$

- here: no unspecified side condition '*q*'
- here: only abandon if goal revision
  - $AGoal_i\, p \wedge \neg Choice_i\, [a]\bot \rightarrow [a]\, \neg Bel_i\, G\, \neg p$

3. Beyond C&L:

integrating causality into the definition of intention

Intention = achievement goal involving choice

$$Intend_i\ p = AGoal_i\ p \wedge Bel_i \neg Stit_{AGT-\{i\}}\ F\ p$$

- dependence clause "$Bel_i \neg Stit_{AGT-\{i\}}\ F\ p$"
  - other agents (including nature/environment/god/…) won't do it
  - *i* must act!
- see Friday lecture on group action

What we saw today

- minimal logic of interaction
  - hypotheses: public events
  - sound and complete axiomatization
- intentions = achievement goals that won't obtain alone
  - condition for intention reconsideration

What we are going to see tomorrow

- collective informational attitudes:
  - common belief
  - group belief
  - acceptance

References

- M. E. Bratman.
  Intentions, plans, and practical reason.
  Harvard University Press, MA, 1987.
- Michael Bratman.
  Faces of intention.
  Cambridge, Cambridge, 1999.
- Philip R. Cohen and Hector J. Levesque.
  Intention is choice with commitment.
  Artificial Intelligence J., 42(2–3):213–261, 1990.
- Philip R. Cohen and Hector J. Levesque.
  Persistence, intentions, and commitment.
  In *Intentions in Communication, chapter 3. MIT Press, 1990.*

References

- Andreas Herzig and Dominique Longin.

  C&L intention revisited.

  In Didier Dubois, Chris Welty, and Mary-Anne Williams, editors, *Proc. 9th Int. Conf. on Principles on Principles of Knowledge Representation and Reasoning(KR2004), pages 527–535. AAAI Press, 2004.*
- Emiliano Lorini and Andreas Herzig.

  A logic of intention and attempt.

  Synthese KRA, 163(1):45–77, 2008.