Course “Logics for Agency and Multi-Agent Systems”

- Thursday -
Logical Frameworks for Multi-Agent Systems: Cohen & Levesque

Andreas Herzig

Outlook

- Tuesday: What agents can do
  - Logic of ability: Coalition Logic
- Wednesday: What agents do, What agents know they (can) do
  - Logic of agency: STIT
- Thursday: What agents want
  - logic of intention: Cohen and Levesque's
- Friday: What agents can plan
  - ATL, Strategic STIT

Modeling MAS

- model the agents' mental attitudes
  - beliefs: epistemic logic, doxastic logic
  - goals, preferences: logic of preferences
  - hopes, fears, regrets, …
- model change
  - events: dynamic logic PDL, dynamic epistemic logics DEL
  - abilities: coalition logic CL, alternating-time temporal logic ATL
  - actions, agency: STIT theory, 'bringing it about' logics
  - belief updates, belief revisions: KM theory, AGM theory
- model interaction and communication
  - …

Today: Cohen&Levesque's logic

- model the agents' mental attitudes
  - beliefs: epistemic logic, doxastic logic
  - goals, preferences: logic of preferences
  - hopes, fears, …
- model change
  - events: dynamic logic PDL, dynamic epistemic logics DEL
  - abilities: coalition logic CL, alternating-time temporal logic ATL
  - actions, agency: STIT theory, 'bringing it about' logics
  - belief updates, belief revisions: KM theory, AGM theory
- model interaction and communication
  - …

Today's outline

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

Today's outline

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

1. Which mental attitudes?

Most basic mental attitude: belief

- agent's internal representation of environment
- reasoning
  - resource-bounded
- basic doxastic logic = KD45 (cf. Monday)
  - belief vs. knowledge
- *omniscience* problem of KD45: agent's beliefs closed under...
  1. logical truth: $p \;/\; Bel_i\, p$
  2. logical equivalences: $p \leftrightarrow q \;/\; Bel_i\, p \leftrightarrow Bel_i\, q$
  3. conjunction: $Bel_i\, p \wedge Bel_i\, q \rightarrow Bel_i\, (p \wedge q)$
  4. material implication: $Bel_i\, p \wedge Bel_i\, (p \rightarrow q) \rightarrow Bel_i\, q$
  5. positive introspection: $Bel_i\, p \rightarrow Bel_i\, Bel_i\, q$
  6. negative introspection: $\neg Bel_i\, p \rightarrow Bel_i \neg Bel_i\, q$

➔ conflicts with resource-boundedness

Belief is not enough: reactive vs. proactive

- reactive agent = performs actions upon stimuli
  - ‘if input = *p* then do *a*’
- *proactive* agent = plans and performs actions to pursue a goal
  - ‘do *a* to achieve p’

➔ *proactive* mental attitudes (pro-attitudes): desires / preferences / goals / intentions / …

From pro-attitudes to action

- a naïve algorithm:
  1. select some pro-attitude
  2. compute difference between pro-attitude and beliefs
  3. generate plan
     - practical syllogism: ‘if you want *p*, and believe action *a* has effect *p*, then do *a*’
  4. perform action

Basic pro-attitude: desire

- beliefs = agent's representation of the world ('where am I?')
- desires = agent's representation of ideal states ('where do I want to be?')
  - … sometimes called goals

- two cases:
  - beliefs = desires: ☺
  - beliefs ≠ desires: ☹

    ➔ do something!

Do desires directly drive our actions?

problems:

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

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

…sometimes called goals

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 Dublin and intention to be *eventually* in Frankfurt might be satisfied in sequence: first *goToDublin*, then *goToFrankfurt*
- directly related to planning and action

Intentions ➔ plans & actions

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

- ESSLLI approaching:
  1. *{goToDublin, goToFrankfurt}*
  2. *… , goToDublin ; … ; goToFrankfurt; …*
  3. *goToDublin ::= buyTicket ; driveToAirport ; takePlane ; takeShuttleBus*
  4. *take shuttle bus ::= findBusStation ; buyTicket ; getOnBus ; getOffBus*
  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

Kinds of intention

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

- present-directed intention (pdi)
  vs. future-directed intention (fdi)
  - pdi causes immediate action: ‘switch to next slide’
  - fdi triggers plan generation: ‘write a book on MAS’

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

Kinds of intention

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

- present-directed intention (pdi)
  vs. future-directed intention (fdi)
  - pdi causes immediate action: ‘switch to next slide’
  - fdi triggers plan generation: ‘write a book on MAS’

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

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

2. Major BDI frameworks

Existing BDI approaches

- Cohen&Levesque's reduction of intention to preference
- Rao&Georgeff's branching-time logic of intention
- Konolige&Pollack
- van Linder, van der Hoek & Meyer's KARO framework

[C&L 1990]

Cohen, Philip R. and Levesque, Hector:
"Intention is choice with commitment"
Artificial Intelligence 42:2-3 (1990)

- *Influential Paper Award* at AAMAS'06
- based on philosophical theory of [Bratman 1987]
- often cited
  - 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
  - actions = entity different from propositions

From C&L to a logic of intention

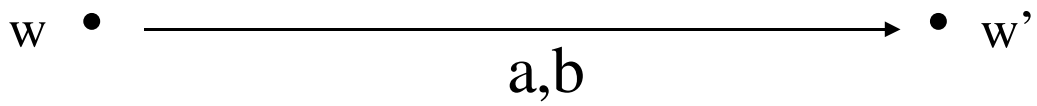
- C&L: formally complex
  - first-order
    - quantification over sequences of events
  - part of the semantics only
    - ‘assumptions’ without semantic counterpart
- last part of talk: sound and complete logic
  - minimalism: propositional modal logic
  - completeness

Today's outline

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

1. Action and time

Transition systems



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

Action: language

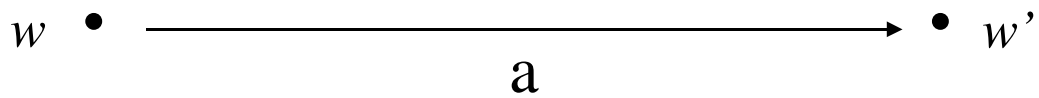
$<a>p$ = "*a* is going to happen, and *p* will hold afterwards"

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

= $\neg <a> \neg p$

- $<takePlane>T$ = *takePlane* is going to happen
- $[takePlane]\,\bot$ = *takePlane* will not happen

Action: semantics



- $w \models \langle a \rangle p$ iff $R_a(w)$ defined & $R_a(w) \models p$

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:
  - $F p \wedge F q \rightarrow F (p \wedge Fq) \vee F (F p \wedge q)$
- axiom relating time and action:
  - $G p \rightarrow [a] p$

Time: the “Before” operator

p Before q = “*p* holds before *q*”

- semantics of LTL
  - $w \models p \textit{ Before } q$ iff for every w'' s.th. $wR_G w''$ & $w'' \models q$
    there is w' s.th. $wR_G w' R_G w''$ and $w' \models p$

- $p \textit{ Before } q \leftrightarrow \neg(\neg p \textit{ Until } q)$

2. Belief

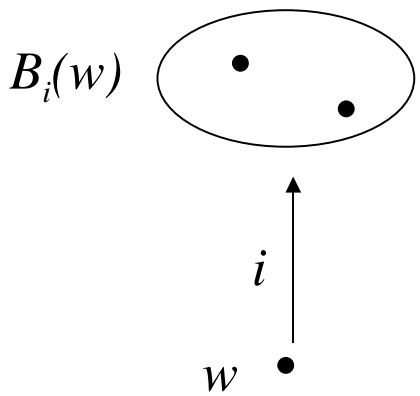
Belief: language

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

Bel_i <*takePlane*>*T* =

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

Belief: semantics



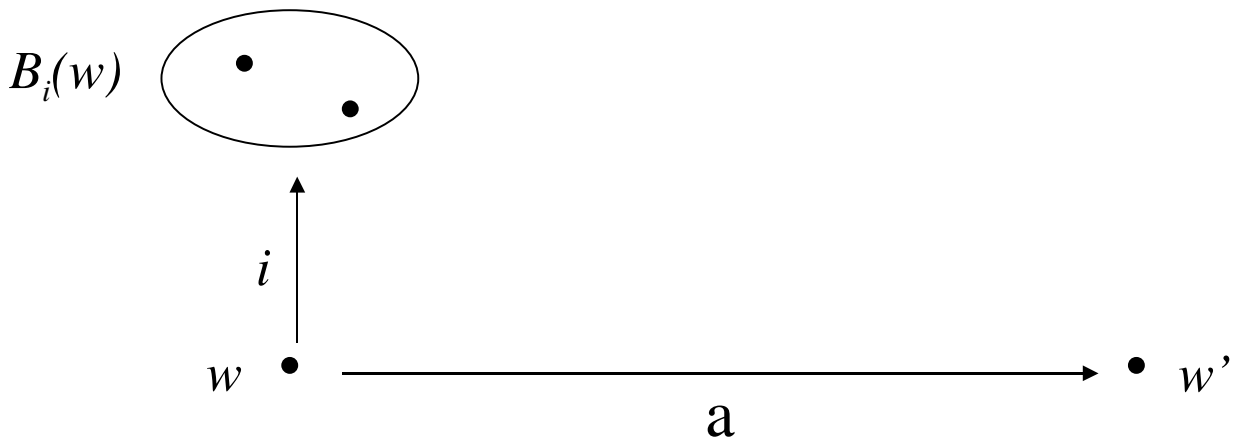
- belief state $B_i(w)$ = set of possible states
- $w \models Bel_i\, p$ iff for every $v \in B_i(w)$, $v \models p$

Belief: axiomatics

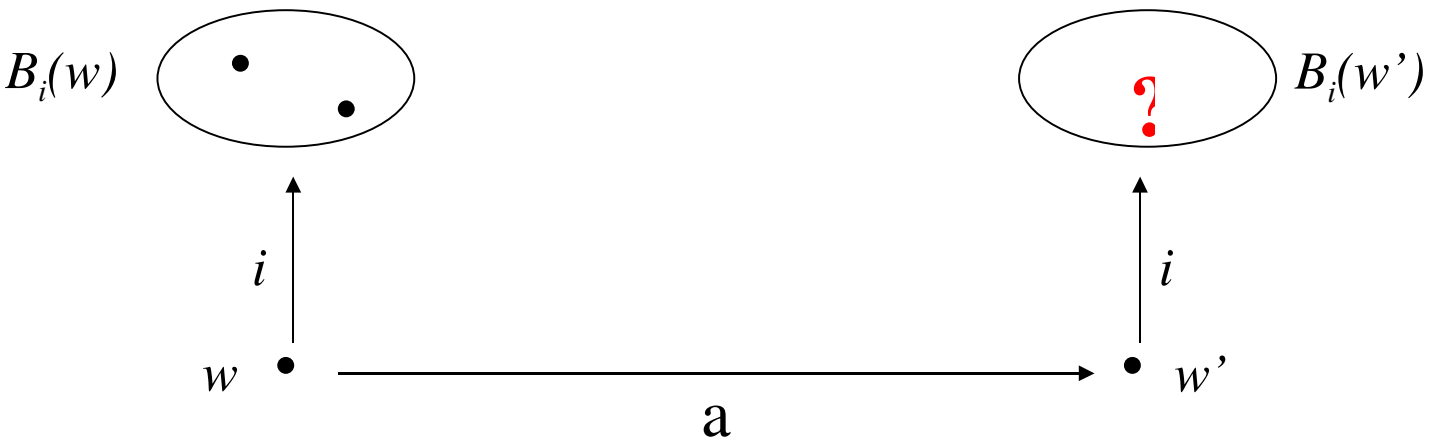
modal logic KD45

- 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. Preference

Motivation: preference as a base for intention

- C&L: intention reduced to belief, action, and preference
  - N.B.: called "choice" by C&L; $\neq$ STIT choices!

- 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 preference persists until *p* is believed;
  4. *i* is prepared to act to make *p* true.

Preference: language

$Pref_i\ p$ = "agent i prefers that p"

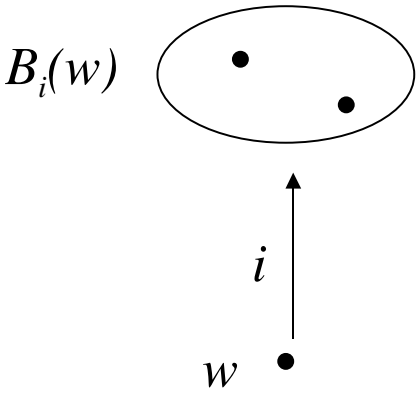
$Bel_i \neg Rich \wedge Pref_i\ F\ Rich$

$Bel_i \neg Rich \wedge Pref_i\ F\ Bel_i\ Rich$

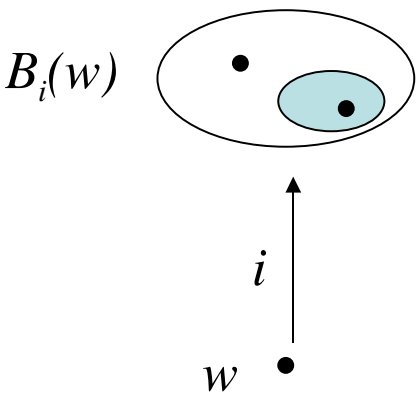
$Pref_i\ [takePlane]\ \bot$

- simple case of preferences: no degrees, no orderings

Preference: semantics



Preference: semantics



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

  iff p holds in all states preferred by i

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

Preference: axiomatics

modal logic KD45

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

Preference and belief: axiomatics

- realism:
  - $Bel_i\, p \rightarrow Pref_i\, p$
- introspection:
  - $Pref_i\, p \rightarrow Bel_i\, Pref_i\, p$
  - $\neg Pref_i\, p \rightarrow Bel_i\, \neg Pref_i\, p$

Where are we?

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

t.b.d.: define intentions

➔ define as particular persistent goals

➔ define as particular achievement goals

Today's outline

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

1. Achievement goal

Achievement goal: definition

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

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

- N.B.: original definition was

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

  ➔ $AGoal_i^{C\&L}(p \wedge \neg Bel_i\, p)$ 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:
  - $AGoal_i\ p \leftrightarrow AGoal_i\ Bel_i\ p$ ($\neq$ 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

$$PGoal_i\, p \;=\; AGoal_i\, p \wedge$$

$$((Bel_i\, p \vee Bel_i\, G \neg p \quad)\ Before\ \neg AGoal_i\, p)$$

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)

$$Intend_i\ p \ = \ PGoal_i\ p \wedge \ Pref_i\ F\ (E\ i{:}a)\ \langle i{:}a\rangle\ Bel_i\ p$$

- *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*!
  - $Intend_i\ sunny$

➔ STIT will help: v.i. …

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

$$\textit{Intend-maintain}_i\ p \ = \ \textit{Bel}_i\ p \wedge \textit{Intend}_i\ G\ p$$

- maintenance intention that p =

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

Present-directed intention

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

- present-directed intention to do *a* =
  preference that *a* happen now

Today's outline

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

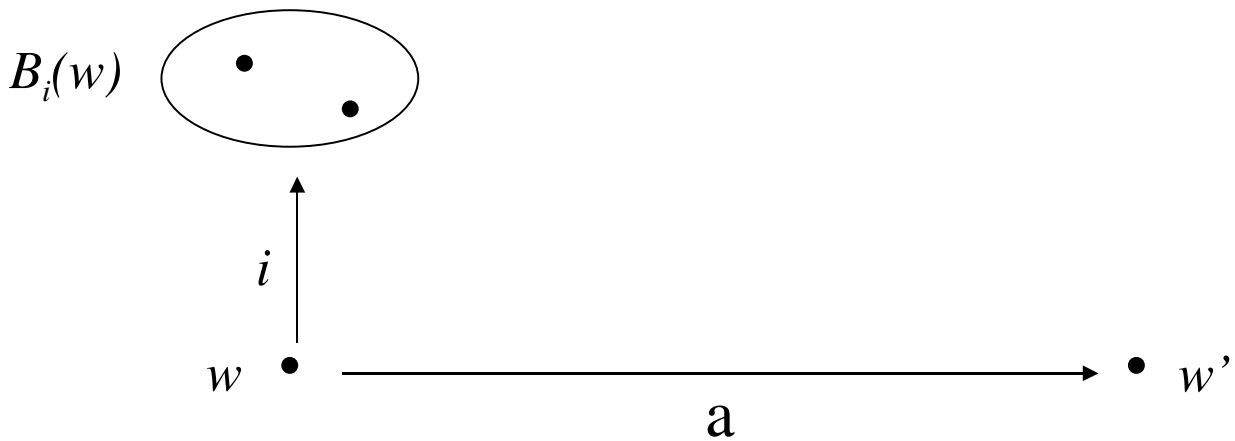
1. Beyond C&L: effects of events on beliefs and preferences

Belief and events: introduction

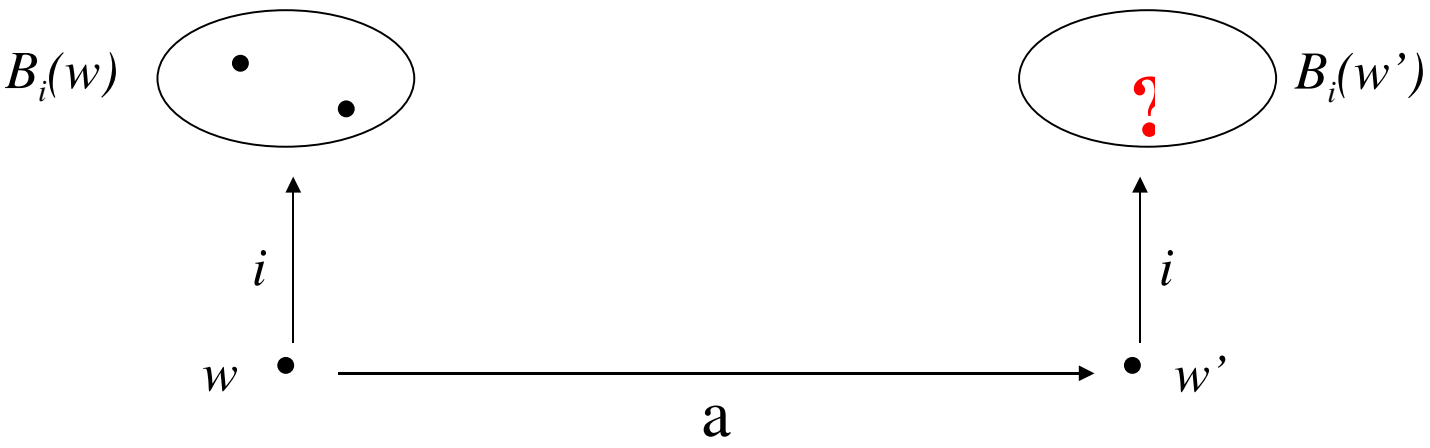
- no interaction in C&L's original paper

- 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

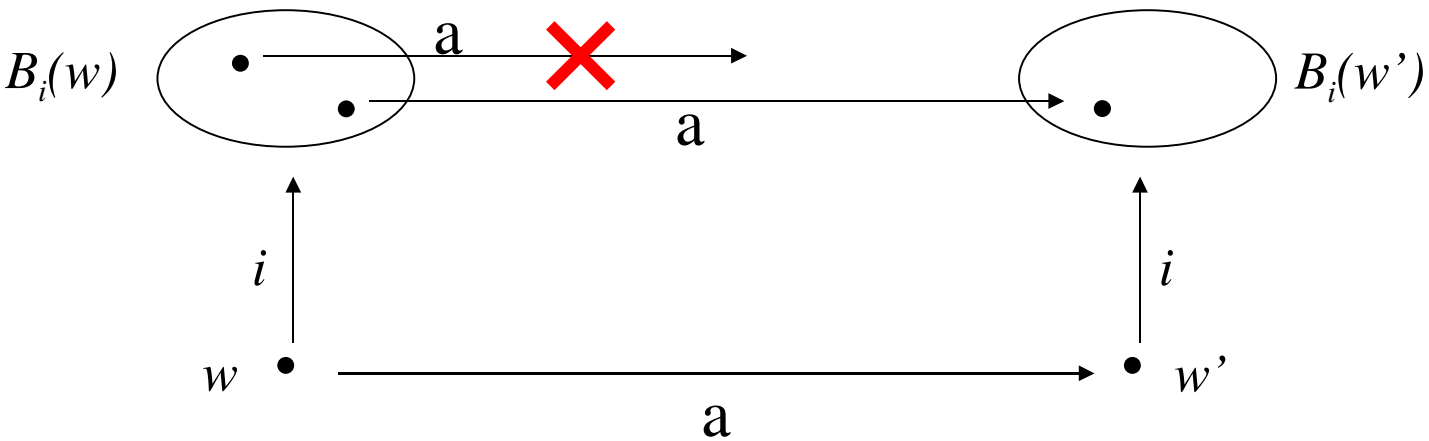
Belief and events: semantics



Belief and events: semantics



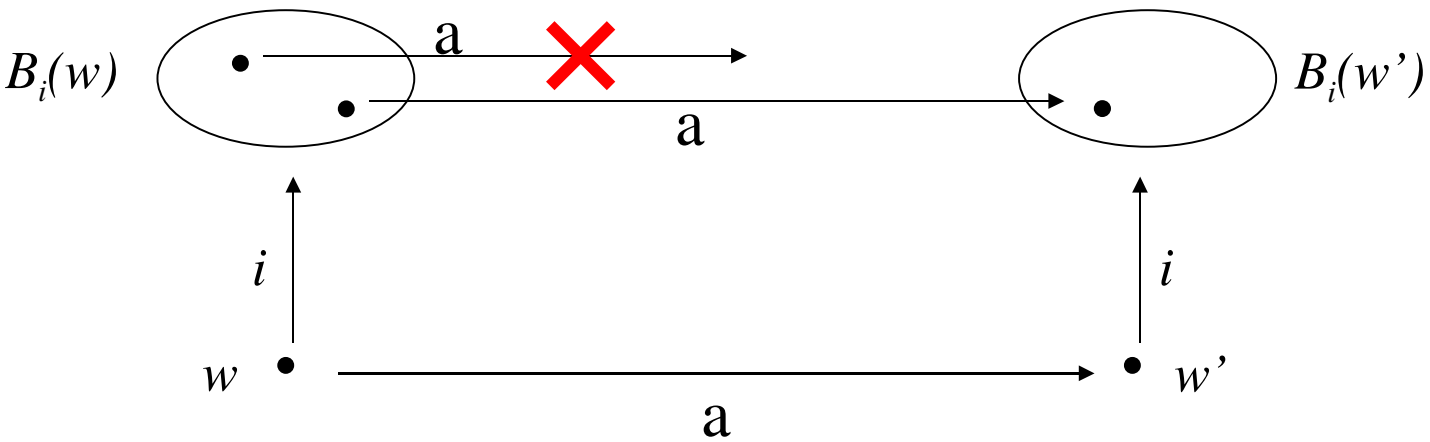
Belief and events: semantics



- $B_i(w') = R_a(B_i(w)) = \bigcup_{v \in B_i(w)} R_a(v)$

  [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

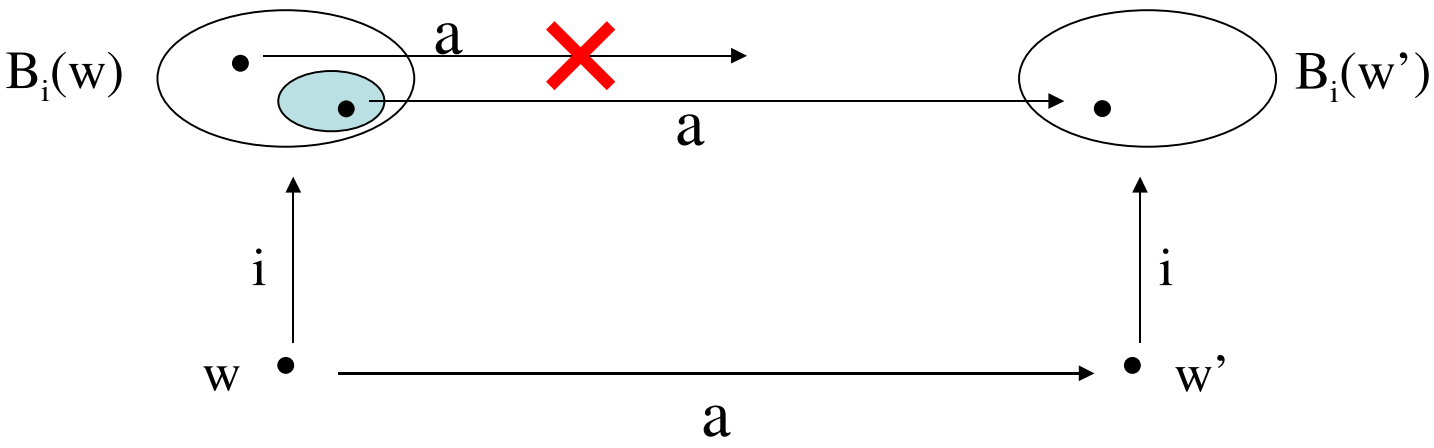
- no learning
  - $[a]\, Bel_i\, p \wedge \neg[a]\bot \;\rightarrow\; Bel_i\, [a]\, p$
- no forgetting
  - $Bel_i\, [a]\, p \wedge \neg Bel_i\, [a]\bot \;\rightarrow\; [a]\, Bel_i p$

Belief and events: axiomatics

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

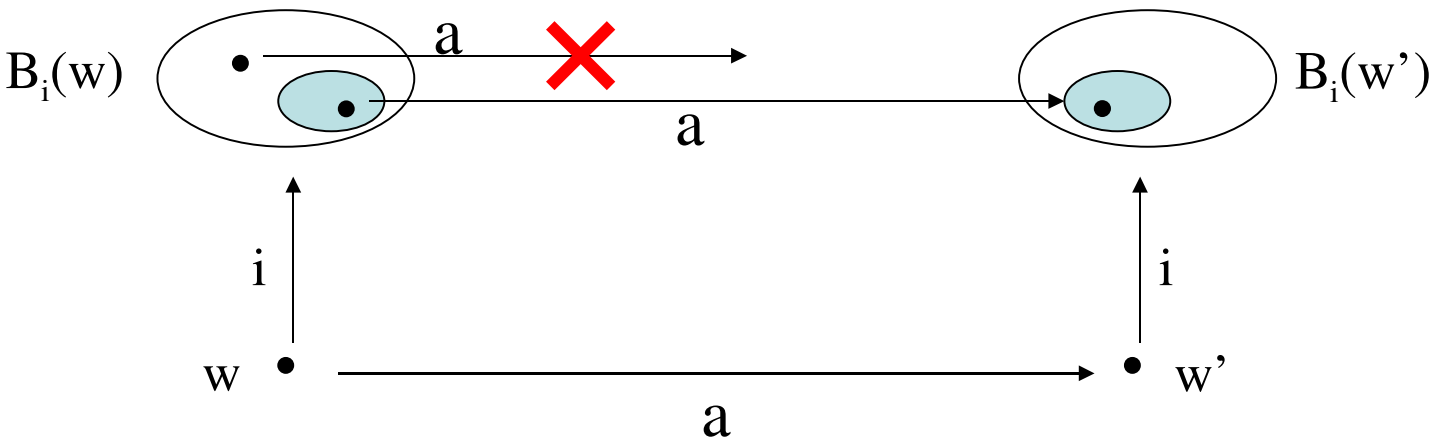
- almost reduction axioms (aka successor state axioms)
- case $Bel_i\, [a]\bot$: belief revision

Preference and action: semantics



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

Preference and action: semantics



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

Preference and action: axiomatics

- no forgetting
  - $Pref_i\ [a]\ p \wedge \neg Pref_i\ [a]\bot\ \rightarrow\ [a]\ Pref_i\ p$
- no learning
  - …
- intentional action
  - $Pref_i\ [i{:}a]\bot\ \rightarrow\ [i{:}a]\bot$ **where *i* is the agent of i:*a***

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

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!

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

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

Proof:

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

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

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
  - $AGoal_i\, p \wedge \neg Pref_i\, [a]\bot \;\rightarrow\; [a]\, \neg Bel_i\, []\, \neg p$

3. Beyond C&L:

integrating causality into the definition of intention

Intention = achievement goal involving preference

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

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

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

- up to now: events, actions take 1 unit of time
- tomorrow: from single-step actions to complex strategies
  - I have a strategy to answer all your questions …