“Logics of Action and Agency” Course at EASSS 2011

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

Girona, July 2011

Introduction to the course

Action

- action: involves an agent, an action, and an outcome
  - “After Ann toggles the switch the light is on.”
- action $\neq$ outcome
  - outcome is a state, or rather: a set of states (‘a proposition’)
  - action is a relation between states [Segerberg]
- plan = composition of actions

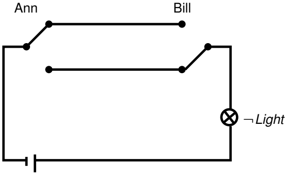
Agency

- agency: involves an agent and an outcome
  - "Ann brings it about that the light is on (by performing some action)" [Belnap]
  - "Ann sees to it that the light is on (by performing some action)" [Kanger, Pörn]
- agentive vs. non-agentive sentences
  - "A sentence φ is agentive for agent i if and only if φ can be reformulated as i *sees to it that* φ" [Belnap]

Capability

- capability = possibility to bring about
  - "Ann is capable to achieve that the light is on (by performing some action)"
  - "Ann has a plan guaranteeing that the light is on (by performing some action)"
- . . . but what do the other agents do?
  - "Ann is capable to make the light on *if the other agents do nothing*"
  - "Ann is capable to make the light on *whatever the other agents do*"
    - cf. strategic games in game theory
- related to multiagent planning

Doing things together



- “Ann is capable to make the light on if Bill does nothing”
- “Ann is not capable to make the light on whatever Bill does”
- “Ann and Bill together are capable to make the light on”

Actions and mental states

- actions are triggered by the agents' *motivational (proactive, teleological) mental states*:
  - desires
  - preferences
  - goals
  - standards, values (internalized norms)
  - future-directed intentions
  - present-directed intentions (plans)
- actions modify not only the world, but also the agents' *informational mental states*
  - knowledge
  - beliefs
  - acceptance ($\neq$ belief)

  ...and may modify their motivational mental states

Course overview

Part I	**logics of knowledge and action (A. Herzig)**
Part II	logics of agency (E. Lorini)

Part I:
Epistemic logics and the dynamics of knowledge

Plan

[0]

Reasoning about knowledge: *de dicto* vs. *de re*

(1) "*there are* irrational x and y such that x^y is rational"
(2) "Hilbert *knows that there are* irrational x, y such that x^y is rational"
(3) "*there are* irrational x, y such that Hilbert *knows* that x^y is rational"

- write these statements in the language of logic
  - abbreviate $\neg Rat(x) \wedge \neg Rat(y) \wedge Rat(x^y)$ by $P(x, y)$
- it follows from the axioms of Peano Arithmetic that $\exists x \exists y P(x, y)$
  - non-constructive proof (5 lines)
- Hilbert knew Peano Arithmetic
- Hilbert knew that $\exists x \exists y P(x, y)$
- there are no x, y of which Hilbert knew that $P(x, y)$
  - although there is a constructive proof ($\sim$20 pages, $\sim$1950)
  - Hilbert was not a perfect, 'omniscient' reasoner

Reasoning about knowledge: muddy children

a famous puzzle:

1. two children come back from the garden, both with mud on their forehead; their father looks at them and says:

   "at least one of you has mud on his forehead"

   then he asks:

   "those who know whether they are dirty, step forward!"
2. nobody steps forward
3. the father asks again:

   "those who know whether they are dirty, step forward!"
4. both simultaneously answer: *"I know!"*

N.B.: can be generalized to an arbitrary number $n \geq 2$ of children

Reasoning about knowledge: muddy children

- use (second-order) predicate $Knows(i, \varphi)$, where $i \in \{1, 2\}$
  - $Knows(i, \varphi)$ = "agent i knows that φ"
- some of child 2's knowledge at the different stages:
  - (S0) background knowledge:
    $Knows(2,\ Knows(1, m_2) \vee Knows(1, \neg m_2))$
    equivalently:
    $Knows(2,\ \neg Knows(1, \neg m_2) \rightarrow Knows(1, m_2))$
  - (S1) learns that at least one of them has mud on his forehead:
    $Knows(2, Knows(1, (m_1 \vee m_2)))$
  - (S2) child 2 does not respond:
    $Knows(2, \neg Knows(1, m_1))$
  - (S3) should follow from (S0)-(S2):
    $Knows(2, m_2)$
- proof?

Reasoning about knowledge: muddy children

deduction of (S3) from (S0), (S1), (S2):

1. $Knows(2,\ Knows(1,(m_1 \vee m_2)))$ hyp. (S1)
2. $Knows(2,\ Knows(1,\neg m_2) \rightarrow Knows(1,m_1))$ conseq. of 1.
3. $Knows(2,\ \neg Knows(1,m_1) \rightarrow \neg Knows(1,\neg m_2))$ equiv. to 2.
4. $Knows(2,\ \neg Knows(1,m_1))$ hyp. (S2)
5. $Knows(2,\ \neg Knows(1,\neg m_2))$ from 3. and 4.
6. $Knows(2,\ \neg Knows(1,\neg m_2) \rightarrow Knows(1,m_2))$ equiv. to hyp. (S0)
7. $Knows(2,\ Knows(1,m_2))$ from 5. and 6.
8. $Knows(2,\ m_2)$ from 7., bec. $Knows(1,m_2) \rightarrow m_2$ ('knowledge implies truth')

informal deduction $\Rightarrow$ formal rules? $\Rightarrow$ deduction in a formal logic?

A second-order theory of the *Knows* predicate

- desirable principles:
  - $\forall i \forall p\, (Knows(i, p) \rightarrow p)$
    - used in step 8.
  - $\forall i \forall p \forall q\, ((Knows(i, p \vee q) \wedge Knows(i, \neg p)) \rightarrow Knows(i, q))$
    - used in step 2.
  - ...
- make up theory of knowledge $\mathcal{T}^2_{Knows}$
  - second-order formulas: "$\forall p$" quantifies over propositions
- reasoning about knowledge in second-order logic (SOL):
  - $\mathcal{T}^2_{Knows} \vdash_{SOL} ((S0) \wedge (S1) \wedge (S2)) \rightarrow (S3)$
  - *SOL* consequence problem
    - undecidable ...

Knows: from second-order to first-order logic

idea [Hin62, FHMV95]:

$Knows(i, \varphi)$ = "φ true in all worlds that are possible for i"

- set of possible worlds W
- ternary 'accessibility' relation $\mathcal{K}(i, w_1, w_2)$
  - i = agent
  - w_1 = actual world
  - w_2 = world that *i cannot distinguish* from w_1
- in first-order logic:

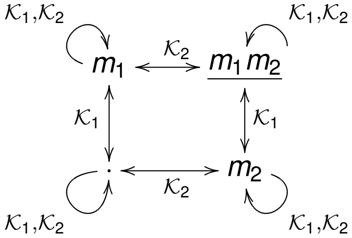
$$Knows(i, \varphi, w) = \text{"at } w, i \text{ knows that } \varphi\text{"}$$

$$\stackrel{\text{def}}{=} \forall w' \, (R(i, w, w') \rightarrow \varphi[w'])$$

Knows: from second-order to first-order logic, ctd.

- muddy children:
  - $Knows(1, m_2, w) \;=\; \forall w' \; (R(1, w, w') \rightarrow m_2(w'))$
  - $\neg Knows(1, m_1, w) \;=\; \exists w' \; (R(1, w, w') \wedge \neg m_1(w'))$

- exercise: draw the set of possible worlds and the accessibility relation in the initial situation



Knows: from second-order to first-order logic, ctd.

- desirable principles for knowledge $\Rightarrow$ properties of $\mathcal{K}$
  - $\forall i \forall p\ (Knows(i, p) \rightarrow p)$ corresponds to: $\forall i \forall w\ \mathcal{K}(i, w, w)$
  - ...
- make up first-order theory $\mathcal{T}^1_{Knows}$
- reasoning about knowledge:
  - $\mathcal{T}^1_{Knows} \vdash_{FOL} \forall w\ (((S0) \wedge (S1) \wedge (S2)) \rightarrow (S3))[w]$
  - consequence problem in first-order logic (FOL)
    - semi-decidable ...

Knows: from first-order to modal logic

idea [Hin62]:

don't use first-order language, but add **modal operators of knowledge** to the language of classical propositional logic *CPL*

- K_i : modal operator
- $K_i\,\varphi$ = "*i* knows that φ"
- epistemic logic
  - *episteme* = $\epsilon\pi\iota\sigma\tau\eta\mu\eta$ = 'know' (Greek)
- N.B.:
  - propositional language; no $\forall$, $\exists$
  - φ might contain modal operator K_j
    - precise definition requires recursive definition of language

Epistemic language: examples

- knowing-whether:
  - $K_1\, m_2 \vee K_1\, \neg m_2$ "child 1 knows whether m_2"
- ignorance:
  - $\neg K_2\, m_2 \wedge \neg K_2\, \neg m_2$ "child 2 does not know whether m_2"
- nesting of modal operators ('higher-order knowledge'):
  - $K_1\, K_2\, (m_1 \vee m_2)$
  - $K_1\, K_2\, K_1\, (m_1 \vee m_2)$
  - …
  - $K_2\, (K_1\, m_2 \vee K_1\, \neg m_2)$
  - $K_2\, (\neg K_1\, m_1 \wedge (K_1\, m_2 \vee K_1\, \neg m_2))$

The propositional logic of knowledge

- extend *CPL* by axiom schemas and inference rules for the modal operator K_i
  - $\vdash \mathrm{K}_i\,\varphi \rightarrow \varphi$
  - if $\vdash \varphi$ then $\vdash \mathrm{K}_i\,\varphi$
  - ...
- reasoning about knowledge:
  - $\vdash \mathrm{K}_2\,\mathrm{K}_1\,m_2 \rightarrow \mathrm{K}_2\,m_2$
  - $\vdash ((S0) \wedge (S1) \wedge (S2)) \rightarrow (S3)$
  - ...
  - reasoning problem: given φ, do we have $\vdash \varphi$?
    - decidable!
    - more details later ...

Reasoning in epistemic logic

- semantics: models? truth conditions?
  - resort to first-order semantics in terms of possible worlds
- models: $M = \langle W, \mathcal{K}, V \rangle$ where
  - W some set ('possible worlds')
  - $\mathcal{K} \subseteq Agts \times W \times W$ ('accessibility relations')
  - V valuation
- truth conditions:
  - $M, w \Vdash \mathrm{K}_i \varphi$ iff $M, w' \Vdash \varphi$ for all w' such that $\mathcal{K}(i, w, w')$
- N.B.: language of epistemic logic less expressive than that of *FOL*
  - $\exists$ different models that give same truth value to all formulas
  - cannot be distinguished by means of a formula
  - bisimulation ...

Recap of basic logic notions

- **logic** Λ = language $\mathcal{L}_\Lambda$ + *particular subset* of $\mathcal{L}_\Lambda$ (called theorems or validities)
- *particular subset* of $\mathcal{L}_\Lambda$ can be characterized in two ways:
  - semantically: using models $\Rightarrow$ validities
  - syntactically: using axioms and inference rules $\Rightarrow$ theorems

Recap of basic logic notions: axiomatics

- requires:
  1. **axiom schemas** = basic theorems of the logic
     - in an axiom schema, we can perform *uniform substitutions*: $K_i\,\varphi \to \varphi$ instantiates to: $K_1\,(m_2 \vee m_1) \to (m_2 \vee m_1)$
     - N.B.: the φ are *meta-variables* over the language
  2. **inference rules** = generate new theorems from existing theorems
     - notation: $\{\varphi_1, \ldots, \varphi_m\} \;/\; \varphi$, or: $\frac{\varphi_1, \ldots, \varphi_m}{\varphi}$
- a **proof** of φ in Λ is a sequence of formulas $\langle \varphi_1, \ldots, \varphi_n \rangle$ such that $\varphi_n = \varphi$, and for every $i \leq n$:
  - φ_i is an (instance of) some axiom schema for Λ, or
  - there are formulas $\varphi_{i_1}, \ldots, \varphi_{i_m}$, such that $i_j < i$, and $\frac{\varphi_{i_1}, \ldots, \varphi_{i_m}}{\varphi_i}$ is (an instance of) some inference rule for Λ
- φ is a **theorem** of Λ iff φ is provable in Λ
  - notation: $\vdash_\Lambda \varphi$
- φ is **consistent** in Λ iff $\not\vdash_\Lambda \neg\varphi$
- **deductions** $\Gamma \vdash_\Lambda \varphi$ iff ... (several options in modal logic)

Recap of basic logic notions: semantics

- requires:
  1. a **class of models** M for Λ
  2. **truth conditions**: when is φ true in M?
     - notation in general: $M \Vdash \varphi$
     - in modal logic: $M, w \Vdash \varphi$ ‘φ is true in $\langle M, w \rangle$’
- φ is **valid** in Λ iff $M, w \Vdash \varphi$, for every model M for Λ and world w in M
  - notation: $\models_\Lambda \varphi$
- φ is **satisfiable** in Λ iff $\not\models_\Lambda \neg\varphi$
- **logical consequence** $\Gamma \models_\Lambda \varphi$ iff . . . (several options in modal logic)

Recap of basic logic notions: soundness and completeness

syntactic and semantic characterizations should coincide!

- **soundness**: for every formula φ, if $\vdash_{\Lambda} \varphi$ then $\models_{\Lambda} \varphi$
  - proof by induction on the length of the proof of φ
- **completeness**: for every formula φ, if $\models_{\Lambda} \varphi$ then $\vdash_{\Lambda} \varphi$
  - actually proved: 'if φ is consistent in Λ then φ is satisfiable in Λ'
  - non-constructive proofs: canonical models [Henkin]
  - constructive proofs: via tableau method

Plan

[0]

Language

- primitive symbols:
  - countable set of propositional atoms *Atms*
  - finite set of agent symbols *Agts*
- BNF:

$$\varphi \ ::= \ p \mid \neg\varphi \mid \varphi \land \varphi \mid \mathsf{K}_i\, \varphi$$

where *p* ranges over *Atms* and *i* over *Agts*

- abbreviations:
  - $\varphi \lor \psi \stackrel{\text{def}}{=} \neg(\neg\varphi \land \neg\psi)$
  - $\varphi \rightarrow \psi \stackrel{\text{def}}{=} \ldots$
  - $\varphi \leftrightarrow \psi \stackrel{\text{def}}{=} \ldots$
  - $\hat{\mathsf{K}}_i\, \varphi \stackrel{\text{def}}{=} \neg\mathsf{K}_i\, \neg\varphi$ = "φ is possible for *i*"

Language (ctd.)

- 3 possible *epistemic attitudes* w.r.t. a formula φ:

$\mathrm{K}_i\,\varphi$	$\hat{\mathrm{K}}_i\,\varphi \wedge \hat{\mathrm{K}}_i\,\neg\varphi$	$\mathrm{K}_i\,\neg\varphi$

  - φ should be *contingent*: neither theorem nor inconsistent
  - what if φ of the form $\mathrm{K}_i\,\psi$?
  - what if φ of the form $\hat{\mathrm{K}}_i\,\psi$?

- 4 possible *epistemic situations* w.r.t. a formula φ:

$\varphi \wedge \mathrm{K}_i\,\varphi$	$\varphi \wedge \hat{\mathrm{K}}_i\,\varphi \wedge \hat{\mathrm{K}}_i\,\neg\varphi$	
	$\neg\varphi \wedge \hat{\mathrm{K}}_i\,\varphi \wedge \hat{\mathrm{K}}_i\,\neg\varphi$	$\neg\varphi \wedge \mathrm{K}_i\,\neg\varphi$

  - ... for φ contingent and non-epistemic
  - why are situations $\varphi \wedge \mathrm{K}_i\,\neg\varphi$ and $\neg\varphi \wedge \mathrm{K}_i\,\varphi$ missing?

Plan

[0]

Motivation and introduction

Multiagent epistemic logic $S5_n$

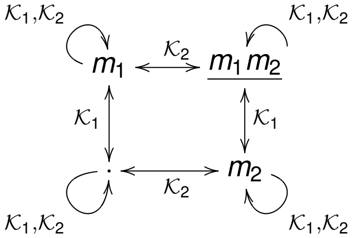
- Introduction
- Language
- Semantics
- Axiomatics
- Discussions

Public announcement logic *PAL*

Dynamic logic of propositional assignments *DL–PA*

Semantics of $S5_n$: Kripke models

- 'Saul **K**ripke' [Kri59]
- ***$S5_n$*-model** = labeled graph $\langle W, \mathcal{K}, V \rangle$ where:
  - W nonempty set 'possible worlds', 'states'
  - $\mathcal{K} : Agts \longrightarrow 2^{W \times W}$ such that every $\mathcal{K}_i$ is an *equivalence relation*
    - equivalence relation = reflexive, transitive, and symmetric relation
    - write $\mathcal{K}_i$ instead of $\mathcal{K}(i)$ 'accessibility relation for i'
  - $V : Atms \longrightarrow 2^W$ 'valuation'
    - $V(p) \subseteq W$
- muddy children:

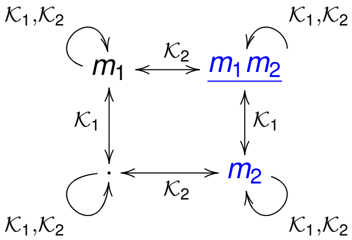


Semantics of $S5_n$: models

- equivalence relation = indistinguishability

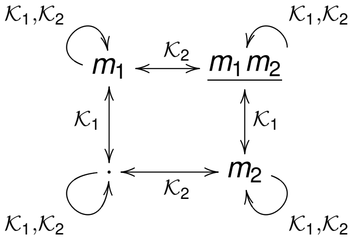
$$\begin{array}{rcl} \mathcal{K}_i(m_1 m_2) & = & \{w \ : \ \langle m_1 m_2, w\rangle \in \mathcal{K}_i\} \\ & = & \text{“set of worlds } i \text{ cannot distinguish from } m_1 m_2\text{”} \\ & = & \text{“set of worlds compatible with } i\text{'s knowledge”} \\ & = & \text{“\textit{knowledge state} of agent } i \text{ at } m_1 m_2\text{”} \end{array}$$

- muddy children:



Semantics of $S5_n$: truth conditions

- truth in a pointed model:
  - $M, w \Vdash p$ iff $w \in V(p)$
  - $M, w \Vdash \neg\varphi$ iff $M, w \nVdash \varphi$
  - $M, w \Vdash \varphi \wedge \psi$ iff $M, w \Vdash \varphi$ and $M, w \Vdash \psi$
  - $M, w \Vdash \mathrm{K}_i\,\varphi$ **iff** $M, w' \Vdash \varphi$ **for every** $w' \in \mathcal{K}_i(w)$
    - hence: $M, w \Vdash \hat{\mathrm{K}}_i\,\varphi$ iff $M, w' \Vdash \varphi$ for *some* $w' \in \mathcal{K}_i(w)$
- muddy children:



$$M, (m_1 m_2) \Vdash m_1 \wedge m_2 \wedge \mathrm{K}_1\, m_2 \wedge \hat{\mathrm{K}}_1\, m_1 \wedge \hat{\mathrm{K}}_1\, \neg m_1$$

Semantics of $S5_n$: satisfiability and validity

- φ is $S5_n$**-satisfiable** iff $M, w \Vdash \varphi$ for *some* $S5_n$-model $M = \langle W, \mathcal{K}, V \rangle$ and *some* possible world $w \in W$
- φ is $S5_n$**-valid** ($\models_{S5_n} \varphi$) iff $M, w \Vdash \varphi$ for *every* $S5_n$-model $M = \langle W, \mathcal{K}, V \rangle$ and *every* possible world $w \in W$

Plan

[0]

Axiomatics of $S5_n$

- axiom schemas for $S5_n$:
  - every theorem schema of *classical propositional logic* (*CPL*)
  - $(K_i\,\varphi \land K_i\,\psi) \to K_i\,(\varphi \land \psi)$ — conjunction $C(K_i)$
  - $K_i\,\top$ — necessity $N(K_i)$
  - $K_i\,\varphi \to \varphi$ — truth $T(K_i)$
  - $K_i\,\varphi \to K_i\,K_i\,\varphi$ — pos. introspection $4(K_i)$
  - $\neg K_i\,\varphi \to K_i\,\neg K_i\,\varphi$ — neg. introspection $5(K_i)$
- inference rules for $S5_n$:
  - $\frac{\varphi, \quad \varphi\to\psi}{\psi}$ — modus ponens (MP)
  - $\frac{\varphi\to\psi}{K_i\,\varphi\to K_i\,\psi}$ — rule of monotony $RM(K_i)$
- N.B.: in axiom schemas and rules, φ, ψ and i are meta-variables
- $S5_n$-proof, $S5_n$-theorem: as usual
- we say:
  - "*CPL*+C(K_i)+N(K_i)+RM(K_i)+T(K_i)+4(K_i)+5(K_i) *axiomatizes* $S5_n$"

Axiomatics of $S5_n$: some useful theorems

- **Kripke's axiom** $K(K_i)$: $\vdash_{S5_n} K_i(\varphi \to \psi) \to (K_i\varphi \to K_i\psi)$
  - proof:
    1. $(K_i\varphi \land K_i(\varphi \to \psi)) \to K_i(\varphi \land (\varphi \to \psi))$ — $C(K_i)$
    2. $(\varphi \land (\varphi \to \psi)) \to \psi$ — (CPL)
    3. $K_i(\varphi \land (\varphi \to \psi)) \to K_i\psi$ — from 2. by $RM(K_i)$
    4. $(K_i\varphi \land K_i(\varphi \to \psi)) \to K_i\psi$ — from 1. and 3. by (CPL)
    5. $K_i(\varphi \to \psi) \to (K_i\varphi \to K_i\psi)$ — from 4. by (CPL)
- $\vdash_{S5_n} (K_i\varphi \land \hat{K}_i\psi) \to \hat{K}_i(\varphi \land \psi)$
  - proof: ... — hint: use (REq) and $K(K_i)$

Plan

[0]

Motivation and introduction

Multiagent epistemic logic $S5_n$

Introduction

Language

Semantics

Axiomatics

Discussions

Public announcement logic *PAL*

Dynamic logic of propositional assignments $DL\text{-}PA$

Knowledge: omniscience

knowledge set of agent i = set of formulas known by i

- i's knowledge set is…
  - closed under theorems:
    - $\frac{\varphi}{K_i\,\varphi}$ rule RN(K)
  - closed under logical implication:
    - $\frac{\varphi \rightarrow \psi}{K_i\,\varphi \rightarrow K_i\,\psi}$ rule RM(K)
  - closed under material implication:
    - $(K_i\,\varphi \land K_i\,(\varphi \rightarrow \psi)) \rightarrow K_i\,\psi$ axiom K(K)
- *omniscience problem*
  - if I know the axioms and inference rules of Peano Arithmetic then I know whether every even integer greater than 2 can be written as the sum of two prime numbers
    - Goldbach's conjecture; still unproved!
  - $S5_n$ is an idealization: rational agent, perfect reasoner
  - inadequate for human agents
  - widely accepted in AI
  - negative introspection criticized

The logic of knowledge: properties

- sound and complete: $\vdash_{S5_n} \varphi$ iff $\models_{S5_n} \varphi$
- decidable
- complexity of $S5_n$-satisfiability is
  - NP-complete if $n = 1$
  - PSPACE-complete if $n > 1$
- there exists a simple normal form for $n = 1$ (single agent)
  - modal depth ≤ 1

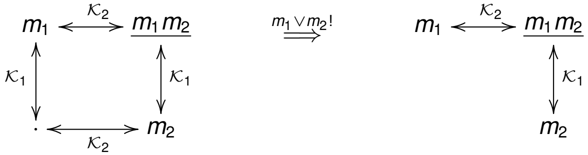
Public announcement logic
PAL

Plan

[0]

Epistemic logic: what if the agents learn something new?

- observe: after the children have heard father's announcement that $m_1 \vee m_2$, they eliminate all those worlds where $m_1 \vee m_2$ is false
- idea: public announcements transform the model ('epistemic update')
- example of muddy children puzzle: father says "$m_1 \vee m_2$!"



(reflexive arrows omitted)

Public announcement logic *PAL*: language

- $\varphi!$ = announcement of truth of φ
- modal operators of public announcement logic (roughly):
  $\{\mathrm{K}_{i_1}, \ldots, \mathrm{K}_{i_{\mathrm{card}(Agts)}}\} \cup \{[\varphi!] \;:\; \varphi \text{ is a formula }\}$
  - either circular definition of formulas
  - or would not allow complex announcements
    - $[([p!]q)!]\mathrm{K}_i\, q$
- BNF:

$$\varphi \;::=\; p \mid \neg\varphi \mid \varphi \wedge \varphi \mid \mathrm{K}_i\,\varphi \mid [\varphi!]\varphi$$

where p ranges over *Atms* and i over *Agts*

- reading:
  $[\varphi!]\psi$ = "ψ is true after every possible execution of the announcement of φ"
  $\langle\varphi!\rangle\psi$ = $\neg[\varphi!]\neg\psi$

Plan

[0]

Motivation and introduction

Multiagent epistemic logic $S5_n$

Public announcement logic *PAL*

- Introduction, motivation, language
- Semantics and axiomatics
- Applications of *PAL*
- From public announcements to private announcements

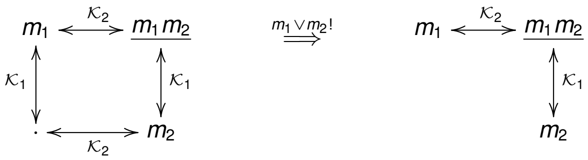
Dynamic logic of propositional assignments $DL{-}PA$

Public announcement logic *PAL*: models

- *PAL*-model = $S5_n$-model
- truth conditions:

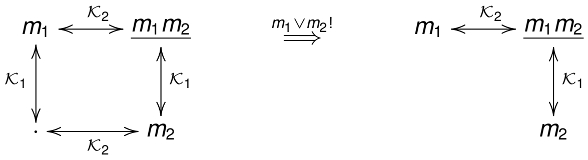
$M, w \Vdash p$	iff	$w \in V(p)$
$M, w \Vdash \neg\varphi$	iff	$\ldots$
$M, w \Vdash \varphi \wedge \psi$	iff	$\ldots$
$M, w \Vdash \mathrm{K}_i\, \varphi$	iff	$M, w' \Vdash \varphi$ for all $w' \in \mathcal{K}_i(w)$
$M, w \Vdash [\varphi!]\psi$	**iff**	$M, w \nVdash \varphi$ **or** $M^{\varphi!}, w \Vdash \psi$

- $M^{\varphi!}$ = "update of M by φ"



(reflexive arrows omitted)

Public announcement logic *PAL*: models (ctd.)



(reflexive arrows omitted)

- $M^{\varphi!} = \langle W^{\varphi!}, \mathcal{K}^{\varphi!}, V^{\varphi!} \rangle$, where

$$\begin{array}{rcl} W^{\varphi!} & = & \{w' \in W \;:\; M, w' \Vdash \varphi\} \\ \mathcal{K}_i^{\varphi!} & = & \mathcal{K}_i \cap (W^{\varphi!} \times W^{\varphi!}) \\ V^{\varphi!}(p) & = & V(p) \cap W^{\varphi!} \end{array}$$

- *Remarks.*
  - announcements have to be truthful
    - else satisfaction relation $\Vdash$ would be ill-defined
  - if there is $w \in W$ such that $M, w \Vdash \varphi$ then $M^{\varphi!}$ is an $S5_n$-model
- *PAL*-validity ($\models_{PAL} \varphi$), *PAL*-satisfiability: defined as usual

Public announcements: non-validities!

- public announcements do not always preserve knowledge:
  $\not\models_{PAL} \mathsf{K}_i \psi \to [\varphi!]\mathsf{K}_i \psi$
  - consider $\psi = \neg \mathsf{K}_i p \ldots$

- public announcements are not always successful:
  $\not\models_{PAL} [\varphi!]\mathsf{K}_i \varphi$
  - consider $\varphi = p \land \neg \mathsf{K}_i p$ ('Moore sentence'),
    and remember: $\mathsf{K}_i (p \land \neg \mathsf{K}_i p)$ is $S5_n$-unsatisfiable!

Reducing *PAL* to $S5_n$

- useful *PAL* validities:

$$\begin{array}{lcll} [\varphi!]\psi & \leftrightarrow & (\neg\varphi \vee \psi) & \text{if } \psi \text{ is atomic} \\ [\varphi!]\neg\psi & \leftrightarrow & (\neg\varphi \vee \neg[\varphi!]\psi) & \\ [\varphi!](\psi_1 \wedge \psi_2) & \leftrightarrow & ([\varphi!]\psi_1 \wedge [\varphi!]\psi_2) & \\ [\varphi!]\mathrm{K}_i\,\psi & \leftrightarrow & (\neg\varphi \vee \mathrm{K}_i\,[\varphi!]\psi) & \end{array}$$

- idea: use equivalences as reduction axioms (rewriting from left to right)
  - ‘push down’ announcement operators
  - eliminate when a Boolean formula is attained
  - $red(\varphi)$ = result of reduction of φ
- exercises:
  - $red([p!]\mathrm{K}_1\, p) = ?$
  - $red([p!]\mathrm{K}_1\, \mathrm{K}_2\, p) = ?$
  - $red([(p \wedge \neg \mathrm{K}_1\, p)!]\mathrm{K}_1\, p) = ?$
- reduction axioms also provide axiomatics (together with rule of substitution of equivalents)
  - while the other axiom schemas of *K* are *PAL*-valid, too, reduction axioms suffice to prove all valid formula instances

Reducing *PAL* to $S5_n$, ctd.

Reduction Theorem.
for every *PAL*-formula φ:

1. $red(\varphi)$ is an $S5_n$-formula
2. $\vdash_{PAL} \varphi \leftrightarrow red(\varphi)$

Sketch of proof.

- equivalences are theorems
- substitution of proved equivalents (REq) preserves *PAL*-theoremhood
- define a decreasing counter (sum of the number of announcements governing subformulas)
  $\Rightarrow$ rewriting terminates

PAL: properties

- satisfiability in *PAL* is decidable
  - apply *red* + decision procedure for $S5_n$
- reduction to $S5_n$ leads to suboptimal decision procedure
- N.B.: rule of uniform substitution not *PAL*-valid:
  - $\vdash_{PAL} [p!]K_1\, p$ (v.s.; p formula!)
  - $\nvdash_{PAL} [\varphi!]K_i\, \varphi$ (v.s.; φ schema!)

Plan

[0]

Muddy children reloaded

- *positive formula* π:
$$\pi ::= \beta \mid \varphi \wedge \varphi \mid \varphi \vee \varphi \mid \mathrm{K}_i\, \varphi$$
where β ranges over Boolean formulas
- prove that $\vdash_{PAL} \pi \to [\varphi!]\pi$ if π is a positive formula
  - induction step for $\pi = \mathrm{K}_i\, \pi_1$:
    1. $\pi_1 \to [\varphi!]\pi_1$ — by induction hyp.
    2. $\mathrm{K}_i\, \pi_1 \to \mathrm{K}_i\, [\varphi!]\pi_1$ — by rule RM(K_i)
    3. $\mathrm{K}_i\, [\varphi!]\pi_1 \to [\varphi!]\mathrm{K}_i\, \pi_1$ — no forgetting
    4. $\mathrm{K}_i\, \pi_1 \to [\varphi!]\mathrm{K}_i\, \pi_1$ — from 2. and 3. by *CPL*
- prove that $\vdash_{PAL} [\pi!]\pi$ if π is a positive formula
  - $\vdash_{PAL} \pi \to [\pi!]\pi$ because ...
  - $\vdash_{PAL} \neg\pi \to [\pi!]\pi$ because ...
- show:
  - $\vdash_{PAL} [(m_1 \vee m_2)!]\mathrm{K}_1\, \mathrm{K}_2\, (m_1 \vee m_2)$
  - $\vdash_{PAL} [\neg \mathrm{K}_2\, m_2!]\mathrm{K}_1\, \neg \mathrm{K}_2\, m_2$
  - $\vdash_{S5_n} (\mathrm{K}_1\, \mathrm{K}_2\, (m_2 \vee m_1) \wedge \mathrm{K}_1\, \neg \mathrm{K}_2\, m_2 \to \mathrm{K}_1\, \neg \mathrm{K}_2\, \neg m_1$
  - $\vdash_{S5_n} (\mathrm{K}_1\, \neg \mathrm{K}_2\, \neg m_1 \wedge \mathrm{K}_1\, (\mathrm{K}_2\, \neg m_1 \vee \mathrm{K}_2\, m_1)) \to \mathrm{K}_1\, \mathrm{K}_2\, m_1$
- conclude that
$\vdash_{PAL} \mathrm{K}_1\, (\mathrm{K}_2\, \neg m_1 \vee \mathrm{K}_2\, m_1) \to [(m_1 \vee m_2)!][\neg \mathrm{K}_2\, m_2!]\mathrm{K}_1\, m_1$

Excursion: the Russian Cards problem [Dit03]

Moscow Mathematics Olympiad in 2000:

From a pack of seven known cards Ann and Bill each draw three cars and Cath gets the remaining card.
How can Ann and Bill publicly and truthfully inform each other about their cards, without Cath learning from any of their cards who holds it?

- cards are 0,1,...,6; Ann holds 012 and Bill holds 345
- some bad solutions:
  - Ann says: “Cath holds 6”
    - Ann can only announce what she knows!
  - Ann says: “I don’t hold 6”
    - Ann should know that Cath doesn’t learn anything!
  - Ann says: “I our Bill hold 012” (and Bill: “I our Ann hold 345”)
    - Cath learns that Ann has 012!
  - Ann says: “either I hold 012, or I hold none of 0, 1, 2”
    - Cath doesn’t learn any card,
    - Ann knows that,
    - but Cath does not know *that*!

Excursion: the Russian Cards problem [Dit03]

- solutions:
  - Ann says: “My cards are among 012, 034, 056, 135 and 246”, and then Bill says: “Cath has 6”
  - ...
- can be modeled in *PAL*
- does not work for any number and any distribution of cards
  - for which numbers there is a solution? (open problem)
- perspective: unconditionally sure cryptographic protocols (perfect reasoners, public communication)
  - RSA algorithm presupposes non-omniscience (decomposition into prime factors not feasible)

Excursion: the paradox of knowability [Fitch]

- add a new modal operator quantifying over announcements:
  - $M, w \Vdash \Diamond\varphi$ iff there is ψ such that $M, w \Vdash \langle\psi\rangle\varphi$
    - N.B.: ψ should have no occurrence of $\Diamond$ (why?)
- allows to reason about plan existence (epistemic actions only)
  - $\models^{?}$ $\mathit{Init} \to \Diamond \mathit{Goal}$

  example: $\models \Diamond(\mathrm{K}_i\, p \vee \mathrm{K}_i\, \neg p)$
- verificationist thesis:
  - $\varphi \to \Diamond\mathrm{K}_i\, \varphi$ should be valid for every φ
- paradox of knowability:
  - $\not\models (p \wedge \neg\mathrm{K}_i\, p) \to \Diamond\mathrm{K}_i\, (p \wedge \neg\mathrm{K}_i\, p)$

Plan

[0]

Dynamic epistemic logic *DEL*

- *PAL*: announcements are perceived by every agent:
  - $[p!](K_1\, p \wedge K_2\, p \wedge K_3\, p \wedge \ldots)$
- idea: $S5_n$ models the agents' uncertainty about current state by means of *possible states*
  $\Rightarrow$ model uncertainty about current event by *possible events*

static uncertainty	dynamic uncertainty
possible worlds	possible events
indistinguishability of worlds	indistinguishability of events

- example: suppose $p \wedge \neg K_1\, p \wedge \neg K_1\, \neg p \wedge \neg K_2\, p \wedge \neg K_2\, \neg p$
  - agent 2 learns that p
  - various possible perceptions of 1:
    - 1 also learns that p, and 2 knows that, etc. $\Rightarrow$ *PAL*
    - 1 sees that 2 learns whether p, but does learn it himself (and 2 knows that, etc.)
    - 1 does not sees this (and 2 knows that, etc.)
    - 1 *suspects this*
    - ...

DEL: event models

- static epistemic logic: static model $M^s = \langle W^s, \mathcal{K}^s, V^s \rangle$
- dynamic epistemic logic: dynamic model $M^d = \langle W^d, \mathcal{K}^d, V^d \rangle$, where
  - W^d is a nonempty set of events
  - $\mathcal{K}^d : Agts \longrightarrow W^d \times W^d$
    - every $\mathcal{K}_i^d$ is an equivalence relation
    - $e\mathcal{K}_i e'$ = "i perceives occurrence of e as occurrence of e'"
  - $V^d : W^d \longrightarrow Fmls$
    - precondition of event w^d
- exercise: find dynamic models for the above examples

DEL: product construction

- given:
  - a static model $M^s = \langle W^s, \mathcal{K}^s, V^s \rangle$
  - a dynamic model $M^d = \langle W^d, \mathcal{K}^d, V^d \rangle$

  what is the resulting static model?

- $M = M^s \otimes M^d = \langle W, \mathcal{K}, V \rangle$ where
  - $W = \{\langle w^s, w^d \rangle \ : \ w^s \in W^s, w^d \in W^d$, and $M, w^s \Vdash V^d(w^d)\}$
  - $\mathcal{K}_i = \{\langle \langle w^s, w^d \rangle, \langle v^s, v^d \rangle \rangle \ : \ w^s \mathcal{K}_i^s v^s$ and $w^d \mathcal{K}_i^d v^d\}$
  - $V(\langle w^s, w^d \rangle) = V^s(w^s)$

restricted product

- exercise: build outcome models for the above examples

DEL: properties

- reduction axioms
- completeness (via reduction axioms)
- applications
  - Cluedo
  - cryptographic protocols
  - ...

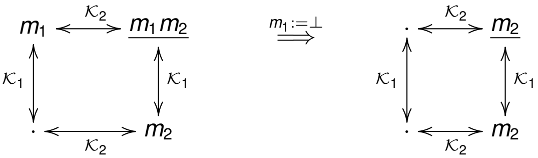
Dynamic logic of propositional assignments $DL-PA$

Plan

[0]

Epistemic logic: what if the facts change?

- suppose the first child publicly cleans his forehead: m_1 becomes false
- idea: public assignment $m_1 := \bot$ transforms the model ('ontic update')



(reflexive arrows omitted)

DL−PA: language

- non-epistemic version
- programs (π) and formulas (φ):

$$\pi \quad ::= \quad p{:=}\top \mid p{:=}\bot \mid \pi;\pi \mid \pi\cup\pi \mid \pi^* \mid \varphi?$$

$$\varphi \quad ::= \quad p \mid \neg\varphi \mid \varphi\wedge\varphi \mid [\pi]\varphi$$

where p ranges over *Atms*

- reading:

$[\pi]\varphi$ = "φ is true after every possible execution of the program π"

$\langle\pi\rangle\varphi$ = $\neg[\pi]\neg\varphi$

Plan

[0]

$DL\!-\!PA$: models

- $DL\!-\!PA$-model = set of propositional variables $w \subseteq Atms$
  - = model of classical propositional logic
  - = a single possible world
- interpretation of formulas:

$$\begin{array}{lll} w \Vdash p & \text{iff} & p \in w \\ w \Vdash \neg\varphi & \text{iff} & \ldots \\ \ldots & & \\ w \Vdash [\pi]\psi & \textbf{iff} & w' \Vdash \varphi \textbf{ for every } w' \textbf{ such that } wR_\pi w' \end{array}$$

- interpretation of programs:

$$\begin{array}{lll} R_{\pi;\pi'} & = & R_\pi \circ R_{\pi'} \\ R_{\pi\cup\pi'} & = & R_\pi \cup R_{\pi'} \\ R_{\pi^*} & = & (R_\pi)^* \\ R_{\varphi?} & = & \{\langle w, w\rangle \ : \ w \Vdash \varphi\} \\ R_{p:=\top} & = & \{\langle w, w'\rangle \ : \ w' = w \cup \{p\}\} \\ R_{p:=\bot} & = & \{\langle w, w'\rangle \ : \ w' = w \setminus \{p\}\} \end{array}$$

- $DL\!-\!PA$-validity ($\models_{DL-PA} \varphi$), $DL\!-\!PA$-satisfiability: defined as usual

DL−PA: properties

- reduction procedure: ...
- satisfiability in *DL−PA* is decidable
  - apply *red* + decision procedure for classical logic
  - suboptimal
- complexity: just as standard *PDL*
  - full *DL−PA* is EXPTIME complete
  - star-free *DL−PA* is PSPACE complete

Plan

[0]

Motivation and introduction

Multiagent epistemic logic $S5_n$

Public announcement logic *PAL*

Dynamic logic of propositional assignments $DL-PA$

- Introduction, motivation, language
- Semantics and axiomatics
- Applications of $DL-PA$

Applications of $DL\text{-}PA$

- logical account of multiagent simulation [Gaudou et al., MABS 2011]
- coalition logic: special propositional variables of ability [Herzig et al., IJCAI 2011]
  - $A_i(p{:=}\top)$ = "i is able to perform $p{:=}\top$"
  - $A_i(p{:=}\bot)$ = ...
- normative systems: special propositional variables of 'deontic ability' [Herzig et al., IJCAI 2011]
  - $P_i(p{:=}\top)$ = "i is permitted to perform $p{:=}\top$"
  - $P_i(p{:=}\bot)$ = ...
- epistemic extension
  - assignments are public:
    $[p{:=}\top]\mathrm{K}_i\,\varphi \leftrightarrow \mathrm{K}_i\,[p{:=}\top]\varphi$
  - reduction

Summary

- What we saw in Part I
  - standard logic of knowledge: $S5_n$
    - criticisms: omniscience
    - static
  - dynamics of knowledge
    - public announcement logic
    - dynamic epistemic logic
- Part II: logics of agency
  - coalition logic
  - the logic of “seeing-to-it-that” (STIT)
  - epistemic extension of STIT
    - ‘I know that there is a plan’ vs. ‘I know a plan’

Hans P. van Ditmarsch.
The russian cards problem.
Studia Logica, 75:31–62, 2003.
Special issue on Dynamics of Knowledge, eds. M. Wooldridge and W. van der Hoek.

Ronald Fagin, Joseph Y. Halpern, Yoram Moses, and Moshe Y. Vardi.
Reasoning about Knowledge.
MIT Press, 1995.

Jaakko K. K. Hintikka.
Knowledge and belief.
Cornell University Press, Ithaca, N.Y., 1962.

Saul Kripke.
A completeness theorem in modal logic.
Journal of Symbolic Logic, 24(1):1–14, 1959.

“Logics of Action and Agency” Course at EASSS’2011 (Part II)

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

Girona, July 2011

Logics of group action and social interaction: an overview

Coalition Logic

STIT logic with agents and coalitions

Epistemic extension of STIT and the concept of uniform strategy

Logics of group action and social interaction: an overview

Examples of group action

- Bill and Bob are painting a house together.
- Brazil soccer team can win FIFA world cup.
- Ann and Mary could have avoided the accident (if they were more cautious).

Logics of joint and group action

- Coalition Logic (CL): Pauly (2001, 2002)
- Alternating-time Temporal Logic (ATL): Alur and Henzinger (2002); van der Hoek and Wooldridge (2003)
- Logic of "Seeing To It That" (STIT logic): Belnap et al. (2001), Horty (2001)

Coalition logic (Pauly 2001, 2002)

- Social software: logics for modelling procedures involving the interactions between multiple agents
- E.g., voting procedure
- It enables to express what a coalition of agents can ensure by doing a joint action

Coalition Logic: main operator

$$\begin{aligned}\langle\!\lbrack J \rbrack\!\rangle\varphi &= \text{“the agents } J \text{ can act together to ensure that } \varphi \\ &\quad \text{is true in the next state”} \\ &= \text{“}\exists \text{ a joint action of } J \text{ such that} \\ &\quad \forall \text{ outcome } \varphi \text{ holds in the next state”}\end{aligned}$$

Alternating-time Temporal Logic (Alur and Henzinger, 2002)

- Logics for multi-agent systems
- Reasoning about dynamic systems with multiple agents
- Temporal extension of Coalition Logic (Goranko 2003)

Alternating-time Temporal Logic (ATL): main operators

$\langle\langle J\rangle\rangle \mathtt{X}\varphi$ = "the agents J can act together to ensure that φ is true in the next state"
= "$\exists$ a joint action of J such that $\forall$ outcome φ holds in the next state"

$\langle\langle J\rangle\rangle \mathtt{G}\varphi$ = "the agents J can act together to ensure that φ is always true"
= "$\exists$ a joint action of J such that $\forall$ outcome φ is always true"

$\langle\langle J\rangle\rangle \varphi \mathtt{Until} \psi$ = "the agents J can act together to ensure that φ remains true until ψ is true"
= "$\exists$ a joint action of J such that $\forall$ outcome φ remains true until ψ is true"

Relation between ATL and CL (Goranko 2003)

$$\langle\![J]\!\rangle\varphi \approx \langle\langle C\rangle\rangle \mathtt{X}\varphi$$

STIT logic (Belnap et al. 2001, Horty 2001)

- Philosophical Logic
- Formalization of concepts in Action theory: *agency*, *ability*, *responsibility*, etc.
- It makes a distinction between *doing* (seeing to it that) and *being able to do* (being able to see to it that)
  - what the agents in a coalition ensure by acting together (not in CL and ATL)
  - what the agents in a coalition can ensure by acting together (also in CL and ATL)

STIT logic: main operators

Formulas in STIT logic are built by means of modal operator $\Box$ which quantifies over the set of all possible outcomes of the current social interaction, whose dual is $\Diamond$, and the so-called *Chellas* STIT operator $[J]$

$$
\begin{aligned}
[J]\varphi &= \text{“coalition } J \text{ sees to it that } \varphi\text{”} \\
&= \text{“the agents } J \text{ ensure that } \varphi \text{ by acting together”} \\
\Box\varphi &= \text{“}\varphi \text{ is necessarily true”} \\
\Diamond\varphi &= \text{“}\varphi \text{ is possibly true”} \\
\Diamond[J]\varphi &= \text{“coalition } J \text{ can see to it that } \varphi\text{”} \\
&= \text{“the agents } J \text{ can act together to ensure that } \varphi\text{”}
\end{aligned}
$$

Relation between STIT and CL (Broersen et al. 2006)

$$\langle\!\langle J \rangle\!\rangle \varphi \approx \Diamond [J] \mathtt{X} \varphi$$

where $\mathtt{X}$ is the operator *next* of linear temporal logic (LTL)
($\mathtt{X}\varphi$ means "φ will be true in the next state")

Coalition Logic

Coalition logic (Pauly 2001, 2002)

- Social software: logics for modelling procedures involving the interactions between multiple agents
- E.g., voting procedure

- It enables to express what a coalition of agents can ensure by doing a joint action

Coalition logic (CL): language

- $AGT = \{1, \ldots, n\}$: a countable set of agents;
- ATM: a countable set of atomic propositions.

Coalition logic (CL): language (cont.)

The language $\mathcal{L}^{CL}$ of CL with agents and groups is defined by the following BNF:

$$\varphi ::= p \mid \varphi \wedge \varphi \mid \neg\varphi \mid \langle\!\![J]\!\!\rangle\varphi$$

where p ranges over ATM and $J \subseteq AGT$.

$\Rightarrow$ We write $\langle\!\![i]\!\!\rangle\varphi$ instead of $\langle\!\![\{i\}]\!\!\rangle\varphi$

- $\langle\!\![J]\!\!\rangle\varphi$: the coalition J can ensure φ at the next time point by acting together, whatever the others agents do.
  - $\exists$ a collective choice of J s.t. $\forall$ next state φ holds.
- $\langle\!\![\emptyset]\!\!\rangle\varphi$: φ is necessarily true at the next time point.

Example: coordinated attack

- Two agents i and j are trying to move an attack against a common enemy. The enemy will be defeated iff i and j move a coordinated attack (both i and j attack the enemy).
  - i has two actions available: *attack* and *skip* (do nothing).
  - j has two actions available: *attack* and *skip* (do nothing).

In formulas:
$\langle\!\langle \{i,j\} \rangle\!\rangle \mathit{defeatEnemy} \land \neg \langle\!\langle i \rangle\!\rangle \mathit{defeatEnemy} \land \neg \langle\!\langle j \rangle\!\rangle \mathit{defeatEnemy}$

Effectivity functions

- S: a set of states
- $e : 2^{AGT} \longrightarrow 2^{2^S}$: effectivity function
  - $X \in e(J)$ iff is a set of possible outcomes for which J is effective (or J can force the world to be in some state of X at the next step)

Example: coordinated attack (cont.)

- $S = \{s_{Defeat}, s_{Undefeat}\}$
- $e(i) = e(j) = \{\{s_{Defeat}, s_{Undefeat}\}\}$
- $e(\{i, j\}) = \{\{s_{Defeat}, s_{Undefeat}\}, \{s_{Defeat}\}, \{s_{Undefeat}\}\}$

Playable effectivity function

e is **playable** iff:

1. $\emptyset \notin e(J)$;
2. $S \in e(J)$;
3. $S \setminus X \notin e(\emptyset)$ then $X \in e(AGT)$, for all $X \subseteq S$ (AGT-*maximality*);
4. if $X_1 \in e(J)$ then $X_1 \cup X_2 \in e(J)$ (*Outcome monotonicity*);
5. if $J \cap I = \emptyset$ then if $X_1 \in e(J)$ and $X_2 \in e(I)$ then $X_1 \cap X_2 \in e(J \cup I)$ (*Superadditivity*).

CL models

A CL-model is a tuple $M = ((S, E), V)$ where:

- S is a set of states;
- $E : S \longrightarrow (2^{AGT} \longrightarrow 2^{2^S})$ associates an effectivity function E_s to every state s in S;
- V is a valuation function, that is, $V : S \longrightarrow 2^{ATM}$.

Truth conditions

- Truth conditions for Boolean constructions are entirely standard.
- $M, s \models \langle\!\langle J \rangle\!\rangle \varphi$ iff $\{s' | M, s' \models \varphi\} \in E_s(J)$.

Validity, satisfiability are defined as usual

A complete axiomatization of CL

(RE)	If $\varphi \leftrightarrow \psi$ then $\langle\!\![J]\!\!\rangle\varphi \leftrightarrow \langle\!\![J]\!\!\rangle\psi$
(M)	$\langle\!\![J]\!\!\rangle(\varphi \wedge \psi) \rightarrow (\langle\!\![J]\!\!\rangle\varphi \wedge \langle\!\![J]\!\!\rangle\psi)$
($\bot$)	$\neg\langle\!\![J]\!\!\rangle\bot$
($\top$)	$\langle\!\![J]\!\!\rangle\top$
(AGT)	$\neg\langle\!\![\emptyset]\!\!\rangle\varphi \rightarrow \langle\!\![AGT]\!\!\rangle\neg\varphi$
(S)	$(\langle\!\![J]\!\!\rangle\varphi \wedge \langle\!\![I]\!\!\rangle\psi) \rightarrow \langle\!\![J \cup I]\!\!\rangle(\varphi \wedge \psi)$ if $J \cap I = \emptyset$

Some CL validities

$\langle\![\emptyset]\!\rangle$ and $\langle\![AGT]\!\rangle$ are normal modalities

- $\vdash (\langle\![\emptyset]\!\rangle\varphi \wedge \langle\![\emptyset]\!\rangle\psi) \rightarrow \langle\![\emptyset]\!\rangle(\varphi \wedge \psi)$
- $\vdash (\langle\![AGT]\!\rangle\varphi \wedge \langle\![AGT]\!\rangle\psi) \rightarrow \langle\![AGT]\!\rangle(\varphi \wedge \psi)$
- $\vdash \langle\![AGT]\!\rangle\varphi \leftrightarrow \neg\langle\![AGT]\!\rangle\neg\varphi$

$\langle\![\emptyset]\!\rangle$ and $\langle\![AGT]\!\rangle$ are inter-definable

- $\vdash \langle\![\emptyset]\!\rangle\varphi \leftrightarrow \neg\langle\![AGT]\!\rangle\neg\varphi$

Two disjoint coalitions cannot bring about conflicting effects

- $\vdash (\langle\![J]\!\rangle\varphi \wedge \langle\![I]\!\rangle\neg\varphi) \rightarrow \bot$ if $I \cap J = \emptyset$

STIT logic with agents and coalitions

STIT logic with agents and coalitions

- Non-standard semantics for STIT in terms of *moments* and *histories* (Belnap et al., 2001; Horty, 2001; Horty & Belnap, 1995).
- STIT can be 'simulated' in a standard Kripke semantics (Herzig & Schwarzentruber, 2008).
  - We use this for today's presentation.

STIT logic with agents and coalitions (Horty, 2001)

- $AGT = \{1, \ldots, n\}$: a countable set of agents;
- ATM: a countable set of atomic propositions;
- $2^{AGT*} = 2^{AGT} \setminus \emptyset$: the set of non-empty *coalitions*.

STIT logic with agents and coalitions

The language $\mathcal{L}^{STIT}$ of STIT with agents and coalitions is defined by the following BNF:

$$\varphi ::= p \mid \varphi \wedge \varphi \mid \neg\varphi \mid [J]\varphi \mid \Box\varphi$$

where p ranges over ATM and J over 2^{AGT*}.

$\Rightarrow$ If J has more than one element: $[J]\varphi \approx$ "coalition J sees to it that φ no matter what the other agents in $AGT \setminus J$ do".
$\Rightarrow$ If J is a singleton $\{i\}$: $[\{i\}]\varphi \approx$ "agent i sees to it that φ no matter what the other agents in $AGT \setminus \{i\}$ do".
$\Rightarrow$ $\Box\varphi \approx$ "φ is necessarily true" (historic necessity).

Further notations

$$\langle J \rangle \varphi \stackrel{\text{def}}{=} \neg [J] \neg \varphi$$

$$\Diamond \varphi \stackrel{\text{def}}{=} \neg \Box \neg \varphi$$

- We write $[i]$ instead of $[\{i\}]$;
- $\Diamond \varphi \approx$ "φ is possibly true";
- $\Diamond [J] \varphi \approx$ "J can see to it that φ whatever the other agents in $AGT \setminus J$ do".

STIT models

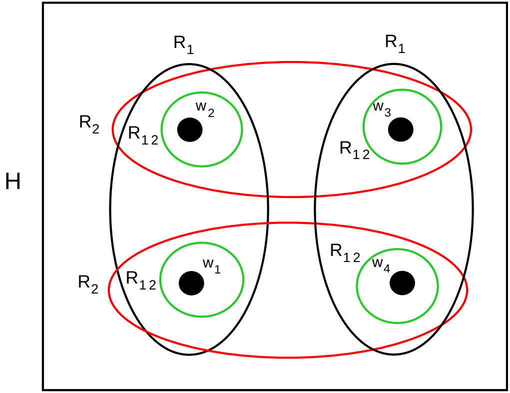
A STIT-model is a tuple $M = (W, \{R_J\}_{J \subseteq AGT}, H, V)$ where:

- W is a non-empty set of possible worlds or states;
- For all $J \subseteq AGT$, R_J is an equivalence relation over W:
  1. **Reflexive**: $(w, w) \in R_J$;
  2. **Transitive**: if $(w, v) \in R_J$ and $(v, u) \in R_J$ then $(w, u) \in R_J$;
  3. **Symmetric**: if $(w, v) \in R_J$ then $(v, w) \in R_J$.
- H is an equivalence relation over W.
- V is a valuation function, that is, $V : W \longrightarrow 2^{ATM}$.

$\Rightarrow R_J(w) = \{v \in W | (w, v) \in R_J\}$ is the set of outcomes that are forced by the action chosen by coalition J at w
$\Rightarrow H(w) = \{v \in W | (w, v) \in H\}$ is the set of all possible outcomes at w

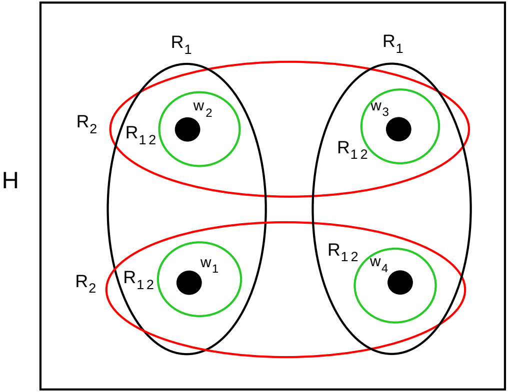
Constraints on STIT models

Action Outcomes: $R_J \subseteq H$



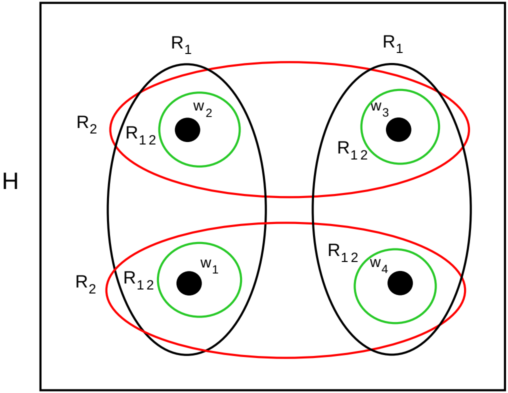
Constraints on STIT models (cont.)

Group Action: $R_J = \bigcap_{j \in J} R_{\{j\}}$



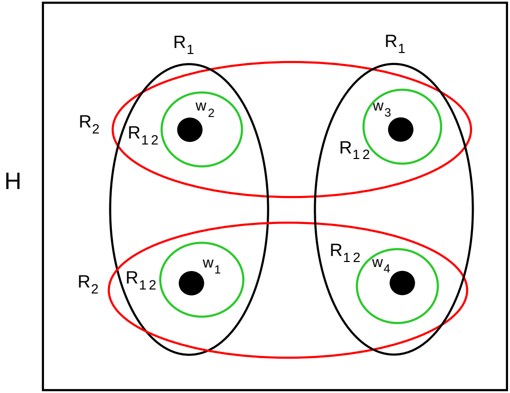
Constraints on STIT models (cont.)

Independence of Agents:
for all $w \in W$, for all $\langle w_j \rangle_{j \in AGT} \in H(w)^n$, $\bigcap_{j \in AGT} R_{\{j\}}(w_j) \neq \emptyset$



Constraints on STIT models (cont.)

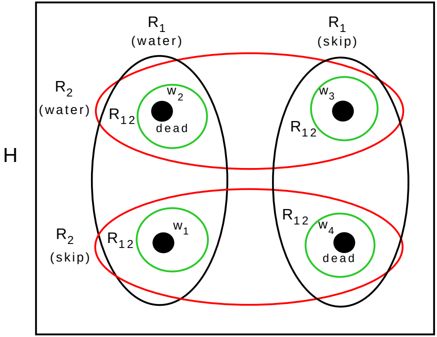
AGT-**Determinism**: $R_{AGT} = id_W$



Truth conditions

- $M, w \models p$ iff $p \in V(w)$.
- $M, w \models \varphi \wedge \psi$ iff $M, w \models \varphi$ and $M, w \models \psi$.
- $M, w \models \neg\varphi$ iff not $M, w \models \varphi$.
- $M, w \models [J]\varphi$ iff $M, v \models \varphi$ for all $(w, v) \in R_J$.
- $M, w \models \Box\varphi$ iff $M, v \models \varphi$ for all $(w, v) \in H$.

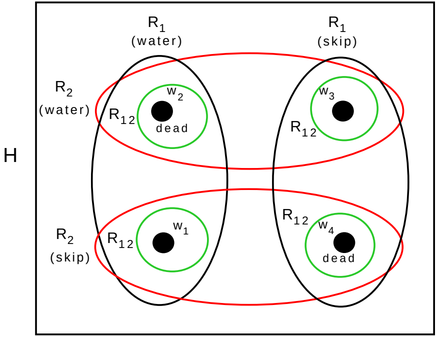
Example 1: two agents taking care of a plant



Two agents $AGT = \{1, 2\}$ have to take care of a plant. If both of them water the plant (resp. do nothing) the plant will die.

- $[\{1, 2\}]$*alive* is true at w_1 and w_3.

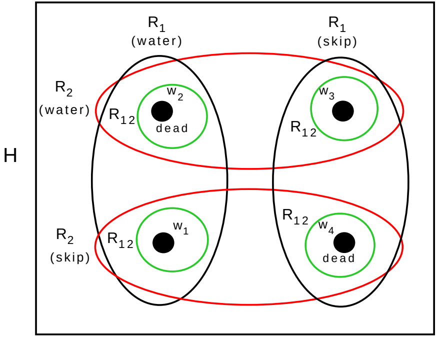
Example 1: two agents taking care of a plant



Two agents $AGT = \{1, 2\}$ have to take care of a plant. If both of them water the plant (resp. do nothing) the plant will die.

- $\Diamond[\{1, 2\}]$*alive* is true at w_1, w_2, w_3, w_4.

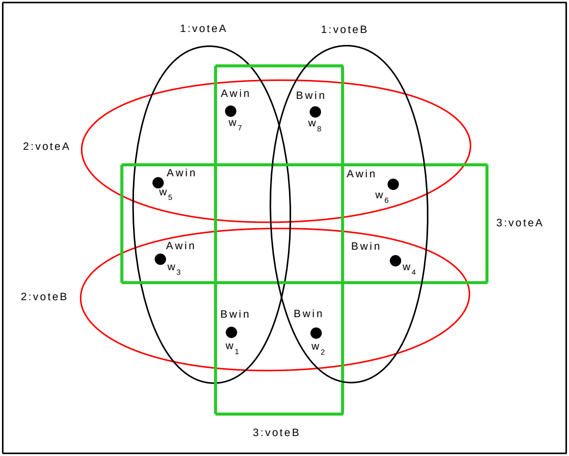
Example 1: two agents taking care of a plant



Two agents $AGT = \{1, 2\}$ have to take care of a plant. If both of them water the plant (resp. do nothing) the plant will die.

- $\neg\Diamond[1]alive$ and $\neg\Diamond[2]alive$ are true at w_1, w_2, w_3, w_4.

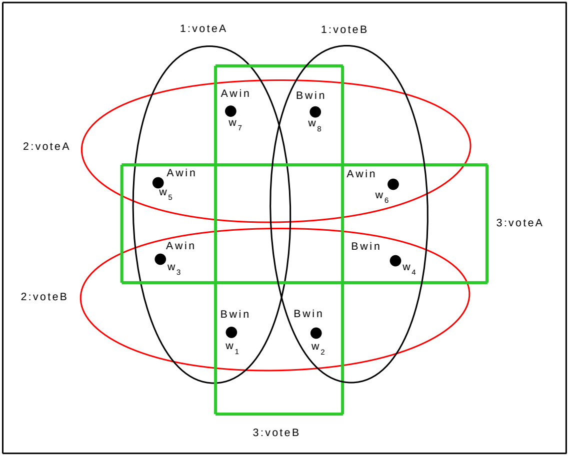
Example 2: voting for a candidate



A and B are the two candidates and $1, 2, 3$ are the three voters.

- $[\{1, 2\}]Awin$ holds at w_5 and w_7.

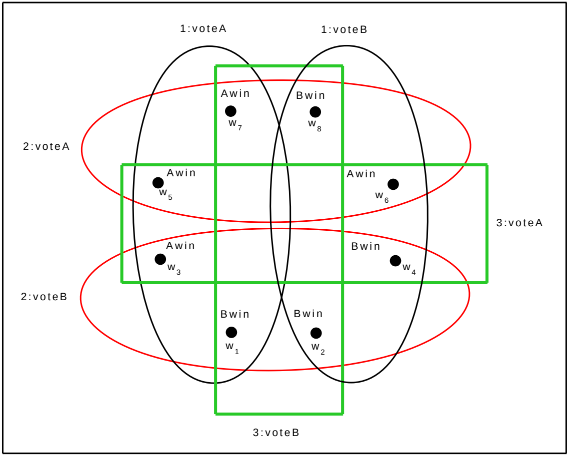
Example 2: voting for a candidate



A and B are the two candidates and 1, 2, 3 are the three voters.

- $\Diamond[\{1,2\}]Awin$ holds at w_1-w_8.

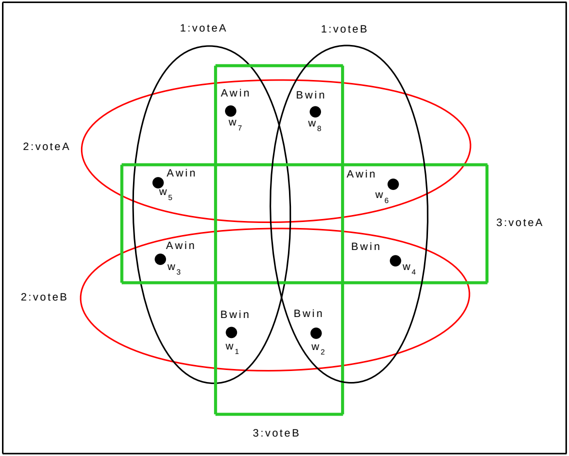
Example 2: voting for a candidate



A and B are the two candidates and 1, 2, 3 are the three voters.

- $\neg\Diamond[\{1\}]Awin \wedge \neg\Diamond[\{1\}]Bwin$ holds at w_1-w_8.

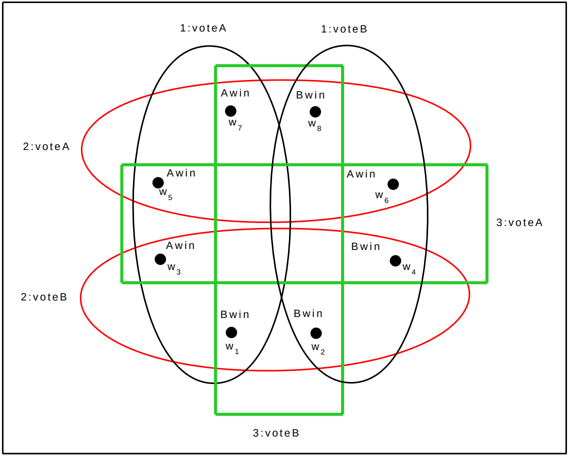
Example 2: voting for a candidate



A and B are the two candidates and 1, 2, 3 are the three voters.

- $\bigwedge_{i\in\{1,2,3\}}(\neg\Diamond[\{i\}]\mathit{1win} \wedge \neg\Diamond[\{i\}]\mathit{2win})$ holds at w_1-w_8.

Example 2: voting for a candidate



A and B are the two candidates and 1, 2, 3 are the three voters.

- $\bigwedge_{J\subseteq\{1,2,3\},|J|\geq 2}(\Diamond[J]Awin \wedge \Diamond[J]Bwin)$ holds at w_1-w_8.

Some STIT validities

$\Box$ and $[J]$ are S5 modalities

1. $\models ([J]\varphi \wedge [J]\psi) \rightarrow [J](\varphi \wedge \psi)$
2. $\models [J]\varphi \rightarrow \varphi$
3. $\models [J]\varphi \rightarrow [J][J]\varphi$
4. $\models \langle J \rangle\varphi \rightarrow [J]\langle J \rangle\varphi$
5. $\models (\Box\varphi \wedge \Box\psi) \rightarrow \Box(\varphi \wedge \psi)$
6. $\models \Box\varphi \rightarrow \varphi$
7. $\models \Box\varphi \rightarrow \Box\Box\varphi$
8. $\models \Diamond\varphi \rightarrow \Box\Diamond\varphi$

Some STIT validities (cont.)

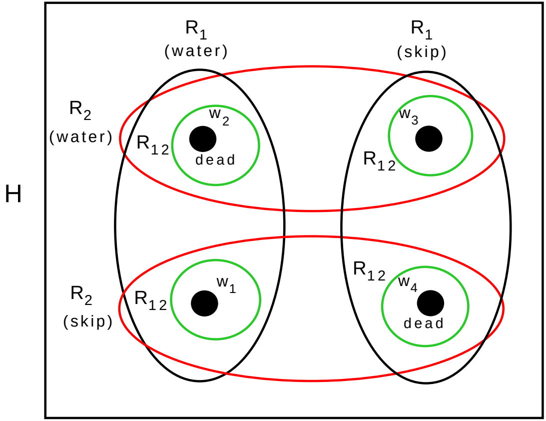
1. $\models \Box\varphi \rightarrow [J]\varphi$
2. $\models \Box\varphi \rightarrow \Box[J]\varphi$
3. $\models \langle AGT\rangle\varphi \leftrightarrow [AGT]\varphi$
4. $\models [J]\varphi \rightarrow [J \cup I]\varphi$
5. $\models [J]\varphi \rightarrow \Diamond[J]\varphi$
6. $\models \Diamond[J]\varphi \rightarrow \Diamond[J \cup I]\varphi$
7. $\models (\Diamond[1]\varphi_1 \wedge \ldots \wedge \Diamond[n]\varphi_n) \rightarrow \Diamond[AGT](\varphi_1 \wedge \ldots \wedge \varphi_n)$
8. $\models (\Diamond[J]\varphi \wedge \Diamond[I]\psi) \rightarrow \Diamond[J \cup I](\varphi \wedge \psi)$ if $I \cap J = \emptyset$
9. $\models (\Diamond[J]\varphi \wedge \Diamond[I]\neg\varphi) \rightarrow \bot$ if $I \cap J = \emptyset$

Discussion: the notion of responsibility

$\chi \wedge \neg[AGT \setminus J]\chi$ expresses a basic notion of **responsibility** of the form "coalition J *could have prevented* a certain state of affairs χ to be true now".

$$\texttt{Resp}_J\chi \stackrel{\text{def}}{=} \chi \wedge \neg[AGT \setminus J]\chi$$

Discussion: the notion of responsibility (cont.)



- $\texttt{Resp}_1 \neg alive$ and $\texttt{Resp}_2 \neg alive$ are true at w_2 and w_4.
- $\texttt{Resp}_{\{1,2\}} \neg alive$ and $\texttt{Resp}_{\{1,2\}} \neg alive$ are true at w_1 and w_4.

Mathematical properties of STIT

Definition

A logic $\mathcal{L}$ is *finitely axiomatizable* if there is a finite set Ax of formula schemas such that $\varphi \in \mathcal{L}$ iff there is a sequence $(\varphi_1, \ldots, \varphi_n)$ of formulas such that $\varphi_n = \varphi$ and for $1 \leq i \leq n$:

- φ_i is a tautology of propositional calculus or an instance of an axiom in Ax or,
- φ_i is obtained by necessitation from φ_j with $j < i$ or,
- φ_i is obtained by modus ponens from φ_j and φ_k with $j, k < i$.

Theorem (Herzig & Schwarzentruber, 2008)

STIT with agents and coalitions is undecidable and not finitely axiomatizable for $AGT \geq 3$.

A decidable fragment of STIT

The following fragment of STIT called DF^{STIT} is decidable:

$$\chi ::= \bot \mid p \mid \chi \wedge \chi \mid \neg\chi \text{ (propositional formulas)}$$
$$\psi ::= [J]\chi \mid \psi \wedge \psi \text{ ("see-to-it" formulas)}$$
$$\varphi ::= \chi \mid \psi \mid \varphi \wedge \varphi \mid \neg\varphi \mid \Diamond\psi \text{ ("see-to-it" and "can" formulas)}$$

where p ranges over ATM and J over 2^{AGT*}.

For instance, $[\{1\}][\{1,2\}]p$ is in STIT but is not in DF^{STIT}

Theorem (Lorini & Schwarzentruber, 2009)

The satisfiability problem of DF^{STIT} is NP-complete.

Epistemic extension of STIT and the concept of uniform strategy

A STIT extension with knowledge

We add modalities for knowledge and a temporal operator to the STIT language:

$\Rightarrow$ $\mathtt{K}_i\varphi$: agent i knows that φ is true

$\Rightarrow$ $\mathtt{X}\varphi$: φ will be true in the next state

Epistemic STIT models

An epistemic STIT-model with discrete time is a tuple $M = (W, \{R_J\}_{J \subseteq AGT}, H, \{E_i\}_{i \in AGT}, F_X, V)$ where:

- W, R_J, H and V are defined as in STIT models;
- For every $i \in AGT$, E_i is an equivalence (epistemic) relation over W;
- $F_X : W \longrightarrow W$.

$\Rightarrow E_i(w) = \{v | (w, v) \in E_i\}$ are the worlds that agent i considers possible at w

$\Rightarrow F_X(w)$ is the successor of world w

Truth conditions for knowledge

$$M, w \models \mathtt{K}_i\varphi \text{ iff } M, v \models \varphi \text{ for all } (w, v) \in E_i$$

$$M, w \models \mathtt{X}\varphi \text{ iff } M, F_X(w) \models \varphi$$

Interactions between knowledge and historic necessity: discussion

$\Rightarrow$ *Semantic constraint* S1: For every $w \in W$, $E_i(w) \subseteq H(w)$

$\Rightarrow$ *Corresponding axiom* **PerfectInfo**: $\Box\varphi \rightarrow \mathtt{K}_i\varphi$

Perfect information about the situation of interaction

Interactions between knowledge and historic necessity: discussion (cont.)

⇒ *Semantic constraint* S2: For every $w \in W$, $E_i(w) \subseteq R_i(w)$
⇒ *Corresponding axiom* **PerfectInfo**+**ActAware**: $[i]\varphi \rightarrow \mathtt{K}_i\varphi$

Perfect information about the structure of interaction + agents' knowledge about their current choice (the only uncertainty is about choices of others)

Theorem
If S2 then S1.

Interactions between knowledge and historic necessity: discussion (cont.)

More realistic principle:

⇒ *Semantic constraint (**KChoiceDet**)* S3: For every $w \in W$, if $v \in H(w)$ and $u \in E_i(v)$ then $u \in E_i(w)$
⇒ *Corresponding axiom*: $\texttt{K}_i\varphi \rightarrow [i]\texttt{K}_i\varphi$

Knowledge is choice determinate

The concept of uniform strategy

For an agent i to have the **power of** ensuring φ, i must have both: the objective capability to achieve φ and, the discretion (awareness) over his capability (Castelfranchi 2003, Barnes 1988)

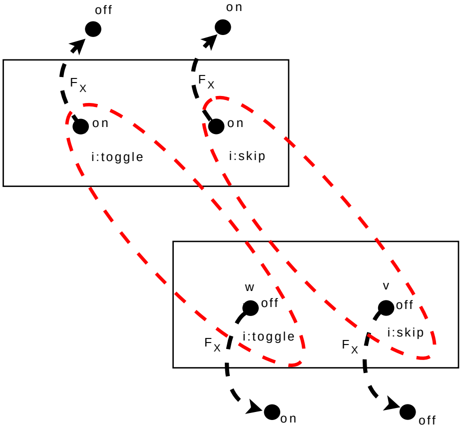
The concept of uniform strategy (cont.)

- *DE DICTO* sentence: "i knows that there is some action such that if he chooses it, he will ensure φ in the next state"
- *DE RE* sentence: "there is some action such that if agent i chooses it, he knows that he will ensure φ in the next state"

$\Rightarrow$ DE DICTO: $\mathtt{K}_i \Diamond [i] \mathtt{X} \varphi$
$\Rightarrow$ DE RE: $\Diamond \mathtt{K}_i [i] \mathtt{X} \varphi$

Only formula $\Diamond \mathtt{K}_i [i] \mathtt{X} \varphi$ captures a proper concept of agent i's *power of* ensuring φ or agent i's *uniform strategy* over φ

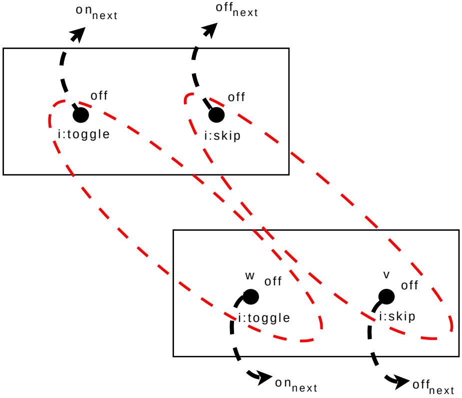
Example: an agent trying to switch on a light



Agent *i* can either *toggle* or *skip* (do nothing) but he is uncertain about the current state of the light (*on* or *off*)

- $\mathtt{K}_i\Diamond[i]\mathtt{X}on$ is true at w and v.
- $\Diamond\mathtt{K}_i[i]\mathtt{X}on$ is false at w and v.

Example: an agent trying to switch on a light (cont.)



Agent i can either *toggle* or *skip* (do nothing) and he knows the current state of the light

- $\mathtt{K}_i \Diamond [i] \mathtt{X} on$ is true at w and v.
- $\Diamond \mathtt{K}_i [i] \mathtt{X} on$ is true at w and v.

References

- Alur, R., Henzinger, T., Kupferman, O. (2002). *Alternating-time temporal logic*. Journal of the ACM, 49:672-713.
- Belnap, N. Perloff, M., Xu, M. (2001). *Facing the future: agents and choices in our indeterminist world*. Oxford University Press.
- Broersen, J. Herzig, A., Troquard, N. (2006). *Embedding Alternating-time Temporal Logic in strategic STIT logic of agency*. Journal of Logic and Computation, 16(5):559-578.
- Goranko, V. (2001). *Coalition games and alternating temporal logics*. In Proceedings of the 8th conference on Theoretical aspects of rationality and knowledge (TARK'01), Morgan Kaufmann Publishers Inc, p. 259-272.
- Goranko, V., Jamroga, W. (2004). *Comparing Semantics for Logics of Multi-agent Systems*. Synthese, 139(2):241-280.

References

- Herzig, A., Schwarzentruber, F. (2008). Properties of logics of individual and group agency. In *Proceedings of Advances in Modal Logic 2008 (AiML 2008)*.
- Horty, J. F., Belnap, N. (1995). *The deliberative STIT: A study of action, omission, and obligation*. Journal of Philosophical Logic, 24(6):583-644.
- van der Hoek, W., Wooldridge, M. (2003). *Cooperation, Knowledge, and Time: Alternating-Time Temporal Epistemic Logic and its Applications*. Studia Logica, 75(1):125-157.
- Horty, J. F. (2001). *Agency and Deontic Logic*, Oxford University Press.
- Lorini, E., Schwarzentruber, F. (2011). *A Logic for Reasoning about Counterfactual Emotions*. Artificial Intelligence, 175(3-4): 814-847.
- Pauly, M. (2002). *A modal logic for coalitional power in games*. Journal of Logic and Computation, 12(1):149-166.
- Pauly, M. (2001). *Logic for Social Software*. PhD thesis, University of Amsterdam, The Netherlands.