# Multi-Agent Systems

Albert-Ludwigs-Universität Freiburg

UNI FREIBURG

Bernhard Nebel, Felix Lindner, and Thorsten Engesser

Summer Term 2017

# Course outline

1. Introduction
2. Agent-Based Simulation
3. Agent Architectures
4. Beliefs, Desires, Intentions
   - The GOAL Agent Programming Language
   - Introduction to Modal Logics
   - Epistemic Logic
   - BDI Logic
5. Norms and Duties
6. Communication and Argumentation
7. Coordination and Decision Making

# Applications of Logics in MAS

- Specification
  - The intended behavior of a MAS can be specified using a logical specification language. The concrete program is derived from the specification (manually, in most cases).
- Verification
  - Once a program $\mathcal{P}$ is built, one wishes to be able to proof that it behaves according to its specification $\varphi_p$, i.e., $\mathcal{P} \models \varphi_p$.
- Agent programming
  - Agents themselves can be realized deductive reasoners: What an agent knows is represented as formulae of a formal language. The agent can reason about these formulae to derive new formulae, or to determine what to do next.

# Model Checking

**Definition**

Model checking is an automated technique that, given a finite-state model of a system and a formal property, systematically checks whether this property holds for (a given state in) that model.

- Model of the system $\Rightarrow$How the system actually behaves.
- Formal properties $\Rightarrow$How the system should behave.
  - Safety: something bad never happen
  - Liveness: something good eventually happens
  - Fairness: if something may happen frequently, it will happen

# Runtime Verification

**Definition**

Runtime verification is the discipline of computer science that deals with the study, development, and application of those verification techniques that allow checking whether a run of a system under scrutiny satisfies or violates a given correctness property.

$\Rightarrow$ Testing using formal methods.

# Requirements

- Question: Does a given BDI agent act right (viz., according to some specified properties)?
- Required
  - Representation of the agent's execution.
  - Language to specify the wanted properties.
  - Algorithm to check if some given properties hold in some represention of an execution.

# Cohen & Levesque, 1990 [1]

- Ingredients:
  - Action and time
  - Belief and preference
  - Definition of intention

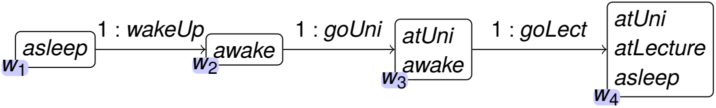
[1] The following notations are according to Meyer, Broersen, Herzig (2015). They slightly deviate from the original notations in Cohen, Levesque (1990).

# Semantics for BDI: Kripkean Model

A BDI Kripke model is a tuple $M = (W, R, B, P, V)$, where:

- $W$ is a set of possible worlds.
- $R : I \times A \to W \times W$
  - Accessibility relations $R_{i:\alpha} \subseteq W \times W$ for each action $i : \alpha$.
  - $(W, R)$ is a linear transition system.
- $B : I \to W \times W$
  - Accessibility relations $B_i \subseteq W \times W$ for each agent $i$.
  - Every $B_i$ is serial, transitive, Euclidean (**KD45**).
- $P : I \to W \times W$
  - Accessibility relations $P_i \subseteq B_i \subseteq W \times W$ for each agent $i$.
  - Every $P_i$ is serial (**KD**).
- $V : P \to 2^W$ m
  - Maps atomic propositions to their extension $V(p) \subseteq W$.

# Actions: Example I

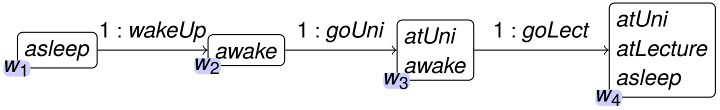


# Actions: Operators

- $M, w \models Happ_{i:\alpha}\varphi$ iff. there is a $w' \in W$ s.th. $(w, w') \in R_{i:\alpha}$ and $M, w' \models \varphi$ ($\Rightarrow$diamond operator).
- $M, w \models IfHapp_{i:\alpha}\varphi$ iff. $M, w \models \neg Happ_{i:\alpha}\neg\varphi$ ($\Rightarrow$box operator).
- $M, w \models \exists\alpha Happ_{i:\alpha}\varphi$ iff. there are agent $i$, action type $\alpha$ and $w'$ s.th. $(w, w') \in R_{i:\alpha}$ and $M, w' \models \varphi$.

- Linearity is characterised by the axiom $(Happ_{i:\alpha}\top \wedge Happ_{j:\alpha'}\varphi) \rightarrow IfHapp_{i:\alpha}\varphi$. ("If $i : \alpha$ is executable and $j : \alpha'$ brings about $\varphi$, then also $i : \alpha$ brings about $\varphi$.")

# Actions: Example II

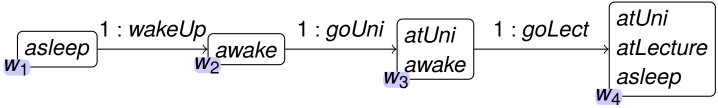



- $M, w_1 \models \mathit{Happ}_{1:\mathit{wakeUp}}\mathit{awake}$
- $M, w_2 \models \exists\alpha \mathit{Happ}_{1:\alpha}\exists\beta \mathit{Happ}_{1:\beta}\mathit{atLecture}$

# Time

- $M,w \models X\varphi$ iff. $M,w' \models \varphi$ for some $w'$ s.th. $(w,w') \in R_{i:\alpha}$ for some $i:\alpha$.
- $M,w \models F\varphi$ iff. $M,w \models \varphi$ or $M,w \models XF\varphi$.
- $M,w \models G\varphi$ iff. $M,w \models \neg F \neg \varphi$.
- $M,w \models \psi U \varphi$ iff. $M,w \models \varphi$ or ($M,w \models \psi$ and $M,w' \models \psi U \varphi$) for some $w'$ s.th. $(w,w') \in R_{i:\alpha}$ for some $i:\alpha$.

# Time: Example




- $M, w_1 \models X(awake \, U \, atLecture)$
- $M, w_1 \models atSleep \wedge XFatSleep$
- $M, w_1 \models G(atSleep \leftrightarrow \neg awake)$
- $M, w_1 \models F \exists \alpha Happ_{1:\alpha} atLecture$

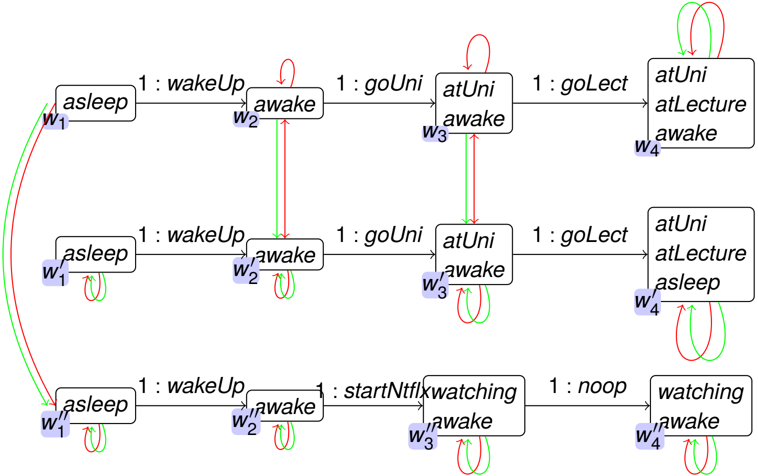
# Belief and Preference

- $M, w \models Bel_i \varphi$ iff. for all $w'$ s.th. $(w, w') \in B_i$: $M, w' \models \varphi$.
  - $Know_i \varphi \stackrel{\text{def}}{=} \varphi \wedge Bel_i \varphi$.
- $M, w \models Pref_i \varphi$ iff. for all $w'$ s.th. $(w, w') \in P_i$: $M, w' \models \varphi$.
  - In the original *Pref* is called *Goal*. Some authors call it *Choice*. It is meant to be a “chosen desire” (consistent!).

## Properties

- For $Bel_i$ all properties for **KD45** operators.
- For $Pref_i$ all properties for **KD** operators.
- $\models Bel_i \varphi \rightarrow Pref_i \varphi$(Realism)
- $\models (Pref_i \varphi \wedge Bel_i(\varphi \rightarrow \psi)) \rightarrow Pref_i \psi$.

# Belief and Preference: Example



# Preferences alone are too weak

- Because of realism, all believed propositions are preferred propositions. But it only makes sense for an agent to adopt some goal $\varphi$ if $\varphi$ is believed to be false (compare to the GOAL programming language).

# Achievement Goal

- Agent $i$ has the achievement goal that $\varphi$ iff $i$ prefers that $\varphi$ is eventually true and believes that $\varphi$ is currently false:

$$AGoal_i\varphi \stackrel{\text{def}}{=} Pref_i F\varphi \wedge Bel_i \neg\varphi$$

**Example**

In the Netflix-vs.-Lecture dilemma:

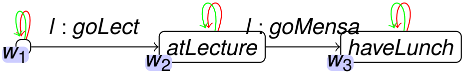
- $M, w_1 \not\models AGoal_1(asleep)$
- $M, w_1 \models AGoal_1(watching)$

# Achievement Goal: Properties

- $\models AGoal_i \neg\varphi \rightarrow \neg AGoal_i \varphi$.
  - Check that $AGoal_i \neg\varphi \wedge AGoal_i \varphi$ is unsatisfiable, because the achievement goal that $\neg\varphi$ implies to believe $\varphi$, and the achievement goal that $\varphi$ implies to believe $\neg\varphi$. This contradicts axiom D ($Bel_i \varphi \rightarrow \neg Bel_i \neg\varphi$). $\square$
- $\not\models AGoal_i(\varphi \wedge \psi) \rightarrow AGoal_i \varphi \wedge AGoal_i \psi$ (for exercise).
- $\not\models AGoal_i \varphi \wedge AGoal_i \psi \rightarrow AGoal_i(\varphi \wedge \psi)$.
- $\not\models AGoal_i(\varphi \vee \psi) \rightarrow AGoal_i \varphi \vee AGoal_i \psi$.
- $\not\models AGoal_i \varphi \vee AGoal_i \psi \rightarrow AGoal_i(\varphi \vee \psi)$.

# $\not\models AGoal_i \varphi \wedge AGoal_i \psi \rightarrow AGoal_i(\varphi \wedge \psi)$

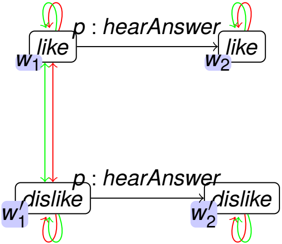
"Lisa has the goal to listen to the lecture and she has the goal to have dinner" vs. "Lisa has the goal to listen to the lecture and to have dinner"




- $M, w_1 \models AGoal_l(atLecture) \wedge AGoal_1(haveLunch)$
- $M, w_1 \not\models AGoal_1(atLecture \wedge haveLunch)$

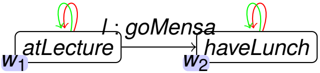
# $\not\models AGoal_i(\varphi \vee \psi) \rightarrow AGoal_i \varphi \vee AGoal_i \psi.$

"Paul asks Lisa whether she likes him." (Paul does not prefer any of the two possible answers.)




- $M, w_1 \models AGoal_p(Know_p like \vee Know_p dislike)$
- $M, w_1 \not\models AGoal_p(Know_p like)$
- $M, w_1 \not\models AGoal_p(Know_p dislike)$

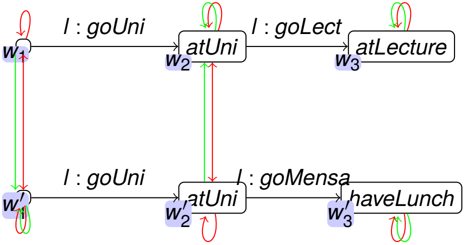
# $\not\models AGoal_i\varphi \vee AGoal_i\psi \rightarrow AGoal_i(\varphi \vee \psi)$




- $M, w_1 \models AGoal_1(haveLunch)$
- $M, w_1 \not\models AGoal_1(atLecture)$
  - Reason: $M, w_1 \not\models Bel_l \neg atLecture$
- $M, w_1 \not\models AGoal_1(atLecture \vee haveLunch)$
  - Reason: $M, w_1 \not\models Bel_l(\neg(atLecture \vee haveLunch))$

# Achievement Goal: Too weak for Intention

- Agents can change their preferences whenever they like: Lack of commitment!




- $M, w_1 \models AGoal_1(haveLunch)$
- $M, w_2 \models \neg AGoal_1(haveLunch)$

# The Nell problem

Say a problem solver is confronted with the classic situation of a heroine, called Nell, having been tied to the tracks while a train approaches. The problem solver, called Dudley, knows that “If Nell is going to be mashed, I must remove her from the tracks.” When Dudley deduces that he must do something, he looks for, and eventually executes, a plan for doing it. This will involve finding out where Nell is, and making a navigation plan to get to her location. Assume that he knows where she is, and he is not too far away; then the fact that the plan will be carried out will be added to Dudley’s world model. Dudley must have some kind of database consistency maintainer to make sure that the plan is deleted if it is no longer necessary. Unfortunately, as soon as an apparently successful plan is added to the world model, the consistency maintainer will notice that “Nell is going to be mashed” is no longer true. But that removes any justification for the plan, so it goes too. But that means “Nell is going to be mashed” is no longer contradictory, so it comes back in. And so forth.

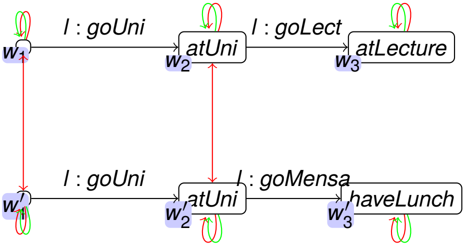
# Commitment

- Cohen & Levesque's intentions involve commitment. Having a commitment means having a persistent goal, viz., a goal the agent only abandons if s(he) comes to believe that the goal is fulfilled or unreachable. This is called single-minded commitment.
- Other forms of commitment:
  - Blind commitment: The agent maintains its intention until it is actually achieved.
  - Open-minded commitment: The agent maintains its intention as long as it is still believed possible. It may e.g. be rendered impossible by adapting new intentions.

# Persistent Goal

- Agent $i$ has the persistent goal that $\varphi$ iff $i$ has the achievement goal that $\varphi$ and will keep that goal until it is either fulfilled or believed to be out of reach:

$$PGoal_i\varphi \stackrel{\text{def}}{=} AGoal_i\varphi \wedge (AGoal_i\varphi)U(Bel_i\varphi \vee Bel_iG\neg\varphi)$$




- $M, w_1 \models PGoal_l(atLecture)$

# Intention

- Agent $i$ has the intention that $\varphi$ iff $i$ has the persistent goal that $\varphi$ and believes that (s)he can achieve $\varphi$ by an action.

$$Intend_i \varphi \stackrel{\text{def}}{=} PGoal_i \varphi \wedge Bel_i F \exists \alpha Happ_{i:\alpha} \varphi$$

- Intending is acting! An agent 1 cannot intend that some other agent 2 does something. However, 1 may intend to make 2 do something.
- Viz., $Intend_1 Happ_{2:act} \top$ expands to $PGoal_1 Happ_{2:act} \top \wedge Bel_1 F \exists \alpha Happ_{i:\alpha} Happ_{2:act} \top$

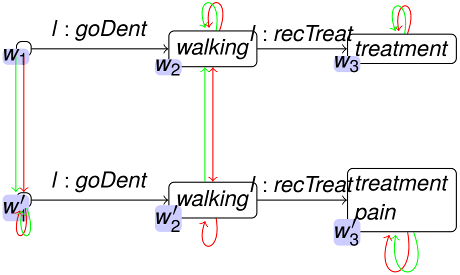
# Intention: Property

- $\not\models (Intend_i \varphi \wedge Bel_i G(\varphi \rightarrow \psi)) \rightarrow Intend_i \psi$.

**Proof**

We provide a model for $Intend_i \varphi \wedge Bel_i G(\varphi \rightarrow \psi) \wedge \neg Intend_i \psi$: John intends to go to the dentist. He believes that going to the dentist always implies pain. At the dentist, John gets some painkiller.

# Dentist Example




- $M, w_1 \models Intend_I(treament) \wedge Bel_I G(treament \rightarrow pain)$, but:
- $M, w_2 \not\models AGoal_I(pain)$, thus:
- $M, w_1 \not\models PGoal_I(pain)$, thus:
- $M, w_1 \not\models Intend_I(pain)$

# Remark: Runtime Verification

- Sketch
  1. Observe the execution of the system to be verified (e.g., log state of the environment, mental state of the agents, the agents' actions).
  2. Represent the execution log using the semantics of Cohen & Levesque.
  3. Model check representation against the agents' specification, e.g.:
     - $G(goldNear \rightarrow Intend(hasGold))$
     - $G(Bel(goldNear) \rightarrow Intend(hasGold))$
     - $G(battLow \rightarrow Intend(\neg battLow))$
  4. Find time points where the specification evaluates false $\Rightarrow$Fault detection.

# Recap and Outlook

- We have studied an integrated logical framework that captures many aspects of agent behavior taking belief and knowledge, preferences, goals, and intentions into account, as well as how these mental attitutes change through time as progressed by actions.
- Next time, we'll look at another important notion, obligations and permissions, and we'll briefly discuss a practical framework (BOID) that deals with decision making in light of conflicts between beliefs, desires, intentions, and obligations.

# Course outline

# Literature


- M. Wooldridge, An Introduction to MultiAgent Systems, 2nd Edition, John Wiley & Sons, 2009.
- Bratman, M. (1987). Intention, plans, and practical reason. Harvard University Press.
- Cohen, P. R., & Levesque, H. J. (1990). Intention is choice with commitment. Artificial intelligence, 42(2-3), 213–261.
- Meyer, J.-J. Ch., Broersen, J., Herzig, A. (2015). BDI Logics. In H. van Ditmarsch, J. Y. Halpern, W. van der Hoek, B. Kooi (Eds.) Handbook of Epistemic Logic. College Publications.