Intelligent Agents
Epistemic Logic

Özgür L. Özçep
Universität zu Lübeck
Institut für Informationssysteme
Today's lecture (and the following five) based on

- The AAMAS 2019 Tutorial „EPISTEMIC REASONING IN MULTI-AGENT SYSTEMS“
MOTIVATION

DIE WISSEN WIRKLICH ALLES ÜBER UNS...

SIE SIND HIER
The need for knowledge

- Many multi-agent systems require to model knowledge of others’ due to imperfect information
  - Agents have local view of environment
  - Agents communicate
  - Agents act -> Decisions taken w.r.t. knowledge
- Similar: Reasoning about other agents „knowledge“ in game theory (second half of this lecture thread Agents, Mechanism, and Collaboration (lecture))
  - Speculation about other one’s strategies/values of things (and about their speculations on our values...)?
  - Collaborative agents (negotiation, communication)
  - Imperfect information
Interaction relies on knowledge

- if I know it is safe then I go
- if I know you are at the market place then I join you
- if (I know it is safe) and (I know you do not know it is safe) then I tell you it is safe
- if I know you know it is safe then I do not tell you it is safe
- if I know you know I know it is safe or not then I do not wait for a message from you
Reminder: Human-compatible AI
Towards XAI?

• XAI = explainable AI: Need to built understandable (human comprehensible) AI systems

• XAI for multi-agent systems?
  – Example: Robots interacting with humans
  – Legal issues in case of failures

Example (Explanations of a robot exploring a world)

• I turned left because \( x = 0 \) and \( y > 5 \)
  \( \Rightarrow \) not human understandable

• I turned left because my neuron 53 was activated.
  \( \Rightarrow \) not human understandable

• I turned left because I *knew* this area was not explored.
  \( \Rightarrow \) human understandable
The need for reasoning

• Given
  – what agents sense
  – The actions and communications that occurred
• What does each agent know?
Once upon a time ... In 2011

- Schwarzentruber (2. presenter of AAMAS 2019) says: „I explained epistemic logic to other researchers in logic/AI/verification ..."

- ... but nobody understood me ..."
Possible worlds

• „... But, since 2017, everybody understood me with comics ...“

• Have a look at http://hintikkasworld.irisa.fr/
Semantics of knowing something

- **Agent a** knows that **agent b** is dirty
- Instance of the famous „muddy children puzzle“
Muddy children Puzzle

“Three children (a,b,c) are playing in the mud. Father calls the children to the house, arranging them in a semicircle so that each child can clearly see every other child. “At least one of you has mud on your forehead”, says Father. The children look around, each examining every other child’s forehead. Of course, no child can examine his or her own. Father continues, “If you know whether your forehead is dirty, then step forward now”. No child steps forward. Father repeats himself a second time, “If you know whether your forehead is dirty, then step forward now”. Some (a,b) but not all of the children step forward. Father repeats himself a third time, “If you know whether your forehead is dirty, then step forward now”. All of the remaining children step forward. Explain why a,b stepped forward after two requests. (In general show: if m children are muddy then after m requests of the father those will step forward”

We will reconsider this puzzle (in the context of dynamic epistemic logic (epistemic logic with operators changing epistemic models))
Epistemic state = pointed Kripke structure

- Comics correspond to unravelling of a pointed Kripke structure

Actual word

$ma = \text{agent a has muddy forehead}$

$mb = \text{agent b has muddy forehead}$
Explaining these in many communities

The Hintikka’s World project

Motivation 1: face the difficulties in explaining possible worlds
Motivation 2: disseminating in many communities

Logic Verification AI

Psychology Cryptography

Distributed systems

Robotics

Games

Philosophy

Open software
Open-source project Hintikka’s world

- http://hintikkasworld.irisa.fr/
- https://gitlab.inria.fr/ fschwarz/hintikkasworld

- Web app
- Modular source code in Typescript
- Easy to add examples
- Several contributors
EPISTEMIC LOGICS (SYNTAX AND SEMANTICS)
Epistemic states

- $AP = \{p, p_1, p_2, \ldots\}$ countable set of atomic propositions
- $AGT = \{a, b, c, \ldots\}$ finite set of agents

**Definition**

An epistemic model $M = (W, (R_a)_{a \in AGT}, V)$ is a tuple where

- $W = \{w, u, \ldots\}$ is a non-empty set of possible worlds
- $R_a \subseteq W \times W$ is an accessibility relation for agent $a$
- $V: W \rightarrow 2^{AP}$ is a valuation function

A pair $(M, w)$ is called an epistemic state, where $w$ represents the actual world. A frame is an epistemic model without the valuation.
Example of an epistemic state

- Muddy children in Hintikka’s world

\[ W = \{ w, u, v, s \} \]
\[ R_a = \{ (w, w), (w, u), (u, w), (u, u), (v, v), (v, s), (s, v), (s, s) \} \]
\[ R_b = \{ (w, w), (w, v), (v, w), (v, v), (u, u), (u, s), (s, u), (s, s) \} \]
\[ V(w) = \{ m_a, m_b \}; V(u) = \{ m_b \}; V(v) = \{ m_a \}; V(s) = \emptyset \]
Syntax of $\mathcal{L}_{EL}$

**Definition**

- The syntax of $\mathcal{L}_{EL}$ (concretely, its set of well-formed formulae) is given by the following grammar:
  \[ \phi ::= p \mid \neg \phi \mid (\phi \lor \phi) \mid K_a \phi \]
  where $p$ ranges over $AP$ and $a$ ranges over $AGT$

- Other operators are defined as follows:
  - $\widehat{K}_a$ abbreviates $\neg K_a \neg \phi$
  - ...

- $K_a \phi$ read as „agent $a$ knows/believes that $\phi$ is true“
- $\widehat{K}_a \phi$ read as „agent $a$ considers $\phi$ as possible“
Syntax of $\mathcal{L}_{EL}$

**Definition**
- Other operators are defined as follows:
  - $(\phi \land \psi)$ abbreviates $\neg(\neg \phi \lor \neg \psi)$
  - $(\phi \rightarrow \psi)$ abbreviates $(\neg \phi \lor \psi)$
  - $\bot$ abbreviates $p \land \neg p$
  - $T$ abbreviates $\neg \bot$
Length and Depth

Definition

The size/length and the modal depth of formulae are defined as follows:

- $|p| = 1$, $d(p) = 0$
- $|\neg \phi| = |\phi| + 1$, $d(\neg \phi) = d(\phi)$
- $|\phi \land \psi| = |\phi| + |\psi| + 1$, $d(\phi \land \psi) = \max\{d(\phi), d(\psi)\}$
- $|K_a \phi| = \phi + 1$, $d(K_a \phi) = 1 + d(\phi)$
Semantics of $\mathcal{L}_{EL}$

**Definition**

The semantics of $\mathcal{L}_{EL}$ (the modelling/satisfaction relation $\models$) is defined recursively by:

- $\mathcal{M}, w \models p$ if $p \in V(w)$
- $\mathcal{M}, w \models \neg \phi$ if not $\mathcal{M}, w \models \phi$
- $\mathcal{M}, w \models \phi \lor \psi$ if $\mathcal{M}, w \models \phi$ or $\mathcal{M}, w \models \psi$
- $\mathcal{M}, w \models K_a \phi$ if for all $u$ s.t. $wR_a u$: $\mathcal{M}, u \models \phi$

Wording: $\mathcal{M}, w$ models $\phi$; $\mathcal{M}, w$ satisfies $\phi$; $\phi$ is true in $\mathcal{M}, w$; $\phi$ holds in world $w$ in $\mathcal{M}$

If $\mathcal{M}, w \models \phi$ for all worlds, then we write $\mathcal{M} \models \phi$ and say $\phi$ is true/valid in $\mathcal{M}$
Semantics of dual operators $\hat{K}_a$

- $\mathcal{M}, w \models K_a \phi$ if for all $u$ s.t. $wR_a u$: $\mathcal{M}, u \models \phi$
- $\mathcal{M}, w \models \bar{K}_a \phi$ if there is $u$ s.t. $wR_a u$: $\mathcal{M}, u \models \phi$

$\mathcal{M}, w \models K_a m_b$

$\mathcal{M}, w \models \hat{K}_a m_a$
Common knowledge

Definition

The **syntax** of $\mathcal{L}_{ELCK}$ is given by the following grammar:

$$\phi ::= p \mid \neg \phi \mid (\phi \lor \phi) \mid K_a \phi \mid C_G \phi$$

where $p \in AP, a \in AGT, G \in 2^{AGT}$

Definition

The **semantics** of $\mathcal{L}_{ELCK}$ is that of $\mathcal{L}_{EL}$ extended by:

$$\mathcal{M}, w \models C_G \phi \iff \text{ for all } u \in W: wR_G u \text{ entails } \mathcal{M}, u \models \phi$$

Here $R_G$ denotes the transitive closure of $\bigcup_{a \in G} R_a$
MODEL CHECKING
Model checking problem

• Input:
  – An epistemic state $\mathcal{M}, w$
  – A formula $\phi$
• Output: yes if $\mathcal{M}, w \vDash \phi$, no otherwise

Theorem

*Model checking (both: with and without common knowledge operators) is P-complete*
(Vanilla) Model checking algorithm

- Input: a Kripke model $\mathcal{M}$, a formula $\phi$
- Output: set of worlds of $\mathcal{M}$ in which $\phi$ holds

**function** $mc(\mathcal{M}, \phi)$

_match $\phi$ do

  case $p$: 
    return \( \{ w \mid \mathcal{M}, w \models p \} \)

  case $\neg \psi$: 
    return \( W \setminus mc(\mathcal{M}, \psi) \)

  case \( \psi_1 \lor \psi_2 \): 
    return \( mc(\mathcal{M}, \psi_1) \cup mc(\mathcal{M}, \psi_2) \)

  case $K_a \psi$: 
    return \( \{ w \mid R_a(w) \subseteq mc(\mathcal{M}, \psi) \} \)
State explosion problem

Example

Minesweeper

- $8 \times 8$ with 10 bombs:
  $> 10^{12}$ possible worlds

- $10 \times 12$ with 20 bombs:
  $> 10^{25}$ possible worlds
State explosion problem

- See (Benthem et al. 2015), (Benthem et al. 2018)

- Also see: (Charrier/S. 2017), (Charrier/S. 2018)
  - Succinct representations of epistemic states; and actions (\(\Rightarrow\) Dynamic Epistemic Logic);
  - Easy to specify by means of accessibility programs;
  - Succinct model checking Pspace-complete (and so stays in Pspace as for non-succinct case).
CALCULI
Satisfiability and validity

Definition

• A formula $\phi$ is **satisfiable** iff there is an epistemic state $\mathcal{M}, w$ s.t. $\mathcal{M}, w \models \phi$
• A formula $\phi$ is **valid** iff for all epistemic states $\mathcal{M}, w : \mathcal{M}, w \models \phi$

Clearly $\phi$ is valid iff $\neg \phi$ is not satisfiable

Example

• $Ka p$ is satisfiable but not valid
• $(Ka p \land Ka (p \rightarrow q)) \rightarrow Ka q$ is valid
Axiomatization

• Checking validity directly not trivial
• Solution: Calculus (with axioms and rules)
  – Axiom (should be valid); rule = „small“ correct inference
  – Derivation/inference: Finite sequence of formulae where
    • each formula is an axiom (instance) or
    • results from applying rule to formulae appearing before.

Definition (calculus $K$)
The basic calculus $K$ is given by the following:
• All classical tautologies (and their uniform substitutions)
• Axiom $K$: $K_a(\phi \rightarrow \psi) \rightarrow (K_a\phi \rightarrow K_a\psi)$
• Rule modus ponens: From $\phi$ and $\phi \rightarrow \psi$ infer $\psi$
• Rule of necessitation: From $\phi$ infer $K_a\phi$
Axiomatization

**Theorem**

A formula is valid (in the class of all epistemic states) iff it is provable in calculus $K$.

*Other wording:* $K$ is correct and complete for the class of all epistemic states.

**Example**

To show: $K_a(\phi \land \psi) \rightarrow K_a\phi$ is valid (by derivation in $K$)

1. $(\phi \land \psi) \rightarrow \phi$ (classical tautology)
2. $K_a((\phi \land \psi) \rightarrow \phi)$ (necessitation to 1.)
3. $K_a((\phi \land \psi) \rightarrow \phi) \rightarrow (K_a(\phi \land \psi) \rightarrow K_a\phi)$ (Axiom K)
4. $K_a(\phi \land \psi) \rightarrow K_a\phi$ (modus ponens to 2,3)
Why axiomatization

- the computation of knowledge is modeled;
- enables to explain why an agent knows something; (link with justification logic)
- axiomatization helps to understand the principle of the logics
- we do not have to design a specific epistemic state, as in model checking („open world“)
# Classes of epistemic states

<table>
<thead>
<tr>
<th>Properties</th>
<th>Related axioms</th>
</tr>
</thead>
<tbody>
<tr>
<td>K</td>
<td>any accessibility relation</td>
</tr>
<tr>
<td>T</td>
<td>Reflexive</td>
</tr>
<tr>
<td>D</td>
<td>Serial</td>
</tr>
<tr>
<td>4</td>
<td>Transitive</td>
</tr>
<tr>
<td>5</td>
<td>Euclidean</td>
</tr>
</tbody>
</table>

- **K**
  - Related axiom: \( K_a \phi \rightarrow \phi \)

- **T**
  - Reflexive
  - Related axiom: \( K_a \top \)

- **D**
  - Serial
  - Related axiom: \( \overline{K}_a \top \)

- **4**
  - Transitive
  - Related axiom: \( K_a \phi \rightarrow K_a K_a \phi \)

- **5**
  - Euclidean
  - Related axiom: \( \neg K_a \phi \rightarrow K_a \neg K_a \phi \)
Each row in the table is a completeness and correctness statement of calculi w.r.t. the given class of epistemic states.

**Definition**

A formula $\phi$ is KD45-valid iff it is true in all epistemic states $\mathcal{M}, w$ in which accessibility relations are serial, transitive, and Euclidean.

**Theorem**

A formula $\phi$ is KD45-valid iff it is provable in the axiomatization $K$ extended with the axioms $D, 4, 5$. 
(No it’s not coffee time but) Time to wake up

- Show that if we confine Kripke structures to those with reflexive relations, then $K_a \phi \rightarrow \phi$ is valid w.r.t. that class

- Show that there might be pointed models which are not reflexive but make $K_a \phi \rightarrow \phi$ true

(But at least you cannot find a frame F (i.e. a model without the evaluation V), such that
  - F is not reflexive and
  - for all models (F,V,w) based on it $K_a \phi \rightarrow \phi$ is made true)
Complexity of checking validity

- Without common knowledge:

<table>
<thead>
<tr>
<th></th>
<th>Single agent</th>
<th>Several agents</th>
</tr>
</thead>
<tbody>
<tr>
<td>K</td>
<td>PSPACE-complete</td>
<td>PSPACE-complete</td>
</tr>
<tr>
<td>KD45, S5</td>
<td>NP-complete</td>
<td>PSPACE-complete</td>
</tr>
</tbody>
</table>

- With common knowledge (and several agents): EXPTIME-complete

- In general and here: Model checking is more practical than theorem proving
LANGUAGE PROPERTIES
Expressivity

Definition

Two formulas $\phi, \psi$ are equivalent iff for all pointed models $\mathcal{M}, w$: $\mathcal{M}, w \models \phi$ iff $\mathcal{M}, w \models \psi$

Theorem

$\mathcal{L}_{ELCK}$ is strictly more expressive than $\mathcal{L}_{EL}$: no formula in $\mathcal{L}_{EL}$ is equivalent to $C_{\{a,b\}}p$

Proof sketch:

• By contradiction, suppose $\phi \in \mathcal{L}_{EL}$ equivalent to $C_{\{a,b\}}p$.
• Let $d$ be the modal depth of $\phi$, e.g., $d = 3$
• Consider two models (from Hinntikka’s world)
• $\phi$ has same value in both models but $C_{\{a,b\}}p$ can distinguish them
Expressivity

• Some operators are mere syntactic sugar such as operator $E_G \phi$, read as "every agent in G knows $\phi$"

• Define
  
  - $\mathcal{M}, w \models E_G \phi$ iff for all agents in $a \in G$: $\mathcal{M}, w \models K_a \phi$

\textbf{Theorem}

$\mathcal{L}_{EL}$ augmented with $E_G$ is equally expressive as $\mathcal{L}_{EL}$

Proof: $E_G \phi \equiv \bigwedge_{a \in G} K_a \phi$

• $E_G$ gives intuitive reading for common knowledge:
  $C_G \phi$ means $E_G^n \phi$ for all $n \in \mathbb{N}$
Bisimulation

- Modal logics and epistemic logics cannot distinguish between structures with same "transition" behaviour
- Captured by notion of bisimulation

### Definition

For two models $M = (W, (R_a)_{a \in AGT}, V)$ and $M' = (W', (R'_a)_{a \in AGT}, V')$ a set $\mathcal{R} \subseteq W \times W'$ is a bisimulation iff for all $w \in W, w' \in W'$ with $(w, w') \in \mathcal{R}$

- $V(w) = V'(w')$
- For all $a \in AGT$, for all $v \in W$: If $R_a(w, v)$ then there is $v' \in W'$ with $R'_a(w', v')$ and $(v, v') \in \mathcal{R}$
- For all $a \in AGT$, for all $v' \in W$: If $R'_a(w', v')$ then there is $v \in W$ with $R_a(w, v)$ and $(v, v') \in \mathcal{R}$
- $(M, w) \bisim (M', w')$ iff there is a bisimulation linking $w$ and $w'$
Bisimilarity preserves formulae

**Theorem**

Suppose that \((M, w) \sim (M', w')\). Then, for all formulas \(\phi \in \mathcal{L}_{ELCK}\) it holds that: \((M, w) \vDash \phi \iff (M', w') \vDash \phi\)

So: Though the common knowledge operator can see arbitrarily far (transitive closure of accessibility relations!; see example on two models before), it can only do in a accessibility-guarded way.

One big meta result regarding bisimulation in the so-called area of correspondence theory (but not directly relevant here)

**Theorem**

Modal logics are exactly those fragments of FOL whose formulae are invariant under bisimilarity

(see, e.g., Blackburn et al, 02)
**Definition**

Given a class \( X \) of models, \( L_1 \) is exponentially more succinct than \( L_2 \) on \( X \) iff the following conditions hold:

- for every formula \( \beta \in L_2 \) there is a formula \( \alpha \in L_1 \) such that \( \alpha \equiv_X \beta \) and \( |\alpha| \leq |\beta| \).
- there exist \( k_1, k_2 > 0 \), a sequence of formulas \( \alpha_1, \alpha_2, \ldots \in L_1 \), and a sequence of formulas \( \beta_1, \beta_2, \ldots \in L_2 \) such that, for all \( n \), we have:
  - \( |\alpha_n| \leq k_1 n \)
  - \( |\beta_n| \geq 2^{k_2 n} \)
  - \( \beta_n \) is the shortest formula in \( L_2 \) that is equivalent to \( \alpha_n \) on \( X \)
Theorem

\( \mathcal{L}_{EL} \) augmented with \( E_G \)’s is exponentially more succinct than \( \mathcal{L}_{EL} \)

- \( E_{\{a,b\}} E_{\{a,b\}} E_{\{a,b\}} \phi \equiv K_a K_a K_a \phi \land K_a K_a K_b \phi \land K_a K_b K_a \phi \land K_a K_b K_b \phi \land K_a K_b K_a \phi \land K_b K_b K_a \phi \land K_b K_b K_b \phi \)
- \( E_{\{a\}} \ldots E_{\{a\}} \equiv \ldots \)

- Proof is involved
  (French, van der Hoek, Illiev, Kooi 2013)
With epistemic logics there is a firm formal foundation for dealing with knowledge in multi-agents
Uhhh, a lecture with a hopefully useful

APPENDIX
References


Book references

- Jaakko Hintikka. Knowledge and Belief: An Introduction to the Logic of the Two Notions (1962)
- J-J Ch. Meyer, van der Hoek, Epistemic logic in AI and computer science, 1995
- van Ditmarsch, van der Hoek, Kooi, Dynamic epistemic logic, 2007
- van Ditmarsch, Joseph Y. Halpern, van der Hoek, Kooi, Handbook of epistemic logic, 2015
Color Convention in this course

- Formulae, when occurring inline
- Newly introduced terminology and definitions
- Important results (observations, theorems) as well as emphasizing some aspects
- Examples are given with standard orange with possibly light orange frame
- Comments and notes
- Algorithms