Multi-Agent Systems

Albert-Ludwigs-Universität Freiburg

Bernhard Nebel, Felix Lindner, and Thorsten Engesser

Winter Term 2018/19

Recap and Overview

- Last time
  - The language of modal logics
  - Kripkean semantics for modal logics
  - Problems you are able to deal with:
    - Model Checking: Checking truth of formulas in a possible world in a Kripke model.
    - Theorem Proving: Checking validity of formulas w.r.t. a class of Kripke models.
- Today
  - The logic of knowledge (and belief)
    - Logical properties of knowledge (and belief)
    - Knowledge of groups of agents
    - Dynamics of knowledge and puzzles

Motivation: Theory of Mind

- Sophisticated modes of social behavior require the ability to "put oneself in the position of someone else"
- Varieties of knowledge
  - Knowledge about others' knowledge:
    - First order:"John knows that the sun is shining"
    - Second order:"John knows that Mary knows that the sun is shining"
    - Third order:"John knows that Mary knows that Peter knows that the sun is shining"
    - ...
  - Knowledge about one's own knowledge
    - Positive introspection: "I know that I know that the sun is shining"
    - Negative introspection: "I know that I don't know that the sun is shining"

Motivation: False-Belief Task

Video: False Belief Task

Knowledge and Belief

- Belief is the attitute of assent towards the truth of particular propositions.
- Knowledge is true justified belief (Plato).
  - Justification: Evidence, or support, for your belief. I.e., if you just claim some truth without evidence, this does not count as knowledge.
- This definition is challenged by philosophical arguments (viz., by so-called Gettier cases).
- Standard epistemic logic is much more pragmatic, though: $\approx$ knowledge is true belief.

Operators for Knowledge and Belief

- $K_i\varphi$: Agent i knows φ.
- $B_i\varphi$: Agent i believes φ.
- Some authors also write $K(i)(p)$ or iKp resp. $B(i)(p)$ or iBp..
- One distinguishes the logic of knowledge (epistemic logic) from the logic of belief (doxastic logic). We will discuss empistemic logic and refer to doxastic logic when the distinction is interesting to us.

Epistemic Alternatives

Def. Epistemically Accessible, Epistemic Alternative

A particular world w' is epistemically accessible to an agent i in world w iff the set of all propositions p that agent i knows in w are compatible with all true propositions in w'. All such worlds w' are considered epistemic alternatives.

- We also say that the epistemic alternatives are epistemically indistinguishable to the agent.

Knowing and Believing: Semantics

Agent i knows φ in world w iff. φ is true in all worlds w' epistemically accessible from w for i:

- $M,w \models K_i\varphi$ iff. for all w' s.th. $(w,w') \in R(K_i), M,w' \models \varphi$

Agent i believes φ in world w iff. φ is true in all worlds w' doxastically accessible from w for i:

- $M,w \models B_i\varphi$ iff. for all w' s.th. $(w,w') \in R(B_i), M,w' \models \varphi$

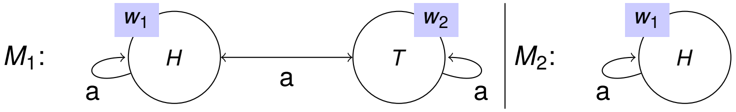
Additionally:

- $M,w \models \hat{K}_i\varphi$ iff. $M,w \models \neg K_i\neg\varphi$.
- $M,w \models \hat{B}_i\varphi$ iff. $M,w \models \neg B_i\neg\varphi$.

Example: Coin Flipping

Quote [Hintikka & Halonen 98]

When you know that S, you can legitimately omit from consideration all possibilities under which it is not the case that S. In other words you can restrict your attention to the situations in which it is true that S.



- $M_1, w_1 \models \neg K_a H \wedge \neg K_a T$
- $M_2, w_1 \models K_a H$

Recap: Axioms

- Some sessions ago we agreed on classifying knowledge as **S5** modality, viz.,:
  - K-axiom: $K_i(\varphi \rightarrow \psi) \rightarrow (K_i\varphi \rightarrow K_i\psi)$
  - T-axiom: $K_i\varphi \rightarrow \varphi$
  - 4-axiom: $K_i\varphi \rightarrow K_iK_i\varphi$
  - 5-axiom: $\neg K_i\varphi \rightarrow K_i\neg K_i\varphi$
- T was considered inappropriate for belief. Instead, the weaker axiom D was considered appropriate:
  - D-axiom: $B_i\varphi \rightarrow \neg B_i\neg\varphi$

Axiom K

$K_i(\varphi \rightarrow \psi) \rightarrow (K_i\varphi \rightarrow K_i\psi)$

- Objection: Logical Omniscience
  - The agent knows all implications of its knowledge.
  - The agent knows all tautologies.
- Is there a way out?
  - Not if we stick to normal modal logics. Approaches exist based on generalizations of Kripke models (impossible worlds). $\Rightarrow$We will not go into detail. If you are interested, see Fagin et al., 1995 [2].

Axiom N

- Necessity: If $\models \varphi$ then $\models K_i\varphi$

Axiom T

$K_i\varphi \rightarrow \varphi$

- Uncontroversial for epistemic logic. Indeed, **KT** is considered weakest system for knowledge.
- $K_aK_b\varphi \models_{\mathbf{KT}} K_a\varphi$, called transmissibility of knowledge. You know everything you know that others know.
  - $K_b\varphi \rightarrow \varphi$ (Axiom T)
  - $K_a(K_b\varphi \rightarrow \varphi)$ (Axiom N)
  - $K_a(K_b\varphi \rightarrow \varphi) \rightarrow (K_aK_b\varphi \rightarrow K_a\varphi)$ (K-Axiom)
  - $K_aK_b\varphi \rightarrow K_a\varphi$ (MP) $\square$
- Do you also believe everything you believe that others believe? $\Rightarrow$Doxastic: Uncontroversial that it does not hold. Weaker axiom D instead.

Axiom D

$K_i\varphi \rightarrow \neg K\neg\varphi$

- Epistemic: Axiom T already entails axiom D.
- Doxastic: Considered as a substitute for T. Psychological objection: People do hold contradictory beliefs.

Axiom 4

$K_i\varphi \rightarrow K_iK_i\varphi$

- Doxastic: Uncontroversial.
- Epistemic:
  - Objection I: So-called *KK*-regress: Agents are required to have infinitely nested knowledge.
  - Objection II: A corollary of the axiom is $\neg K_iK_i\varphi \rightarrow \neg K_i\varphi$ (counterposition). But it seems possible to have knowledge that comes into ones mind only after one gets some hints.

However, as Hintikka notes, "Knowing to know differs only in words from knowing." In system **KT4** we have $K_iK_i\varphi \rightarrow K_i\varphi$, $K_i\varphi \rightarrow K_iK_i\varphi$, hence $K_iK_i\varphi \leftrightarrow K_i\varphi$.

Axiom 5

$\neg K_i\varphi \rightarrow K_i\neg K_i\varphi$

- Objection: A corollary is $K_i\varphi \vee K_i\neg K_i\varphi$ ($\equiv_{Def.}$ awareness). No room for ignorance! Consider John living in 17th century: According to axiom 5, John either knew Einstein's theory of relativity, or he knew that he does not know Einstein's theory of relativity. In fact, it is more appropriate to claim that John was not aware.

Axiom B

From reflexivity (T) and Euclideanness (5) follows symmetry (B): $\neg\varphi \rightarrow K_i\neg K_i\varphi$ (Proof: $\neg\varphi \Rightarrow_{\mathbf{T}} \neg K_i\varphi \Rightarrow_{\mathbf{5}} K_i\neg K_i\varphi \quad \square$)

- Objection: What is actually true must be known to be possibly true.
- Objection: A corollary of axiom B is $\neg K_i\neg K\varphi \rightarrow \varphi$, which is the same as $\hat{K}K\varphi \rightarrow \varphi$. In words: Only true things are considered possible to be known resp. believed.

As this seems too strong to most epistemologists, 5 (and thus B) is often rejected. $\Rightarrow$Alternative Axioms proposed: 4.2, 4.3, 4.4
However, in many computer science applications, B is considered appropriate: If the agent cannot find φ in its database, it can conclude that it knows that it does not know φ (closed-world assumption).

Axiom 4.4

$\varphi \rightarrow (\neg K_i\neg K_i\varphi \rightarrow K_i\varphi)$

- A corollary is $\neg\varphi \vee K_i\neg K_i\varphi \vee K_i\varphi$. Thus, the agent is not required to know something about things that do not hold.
- In **KT4.4**: If φ is not the case, then $\neg K_i\varphi$ (by T), but not $K_i\neg K_i\varphi$. If φ is the case, then either it knows φ or it knows that it does not know φ.

Axiom 4.3

$K_i(K_i\varphi \rightarrow K_i\psi) \vee K_i(K_i\psi \rightarrow K_i\varphi)$

- Variant: $K_i(\neg K_i\varphi \vee K_i\psi) \vee K_i(\neg K_i\psi \vee K_i\varphi)$
- $K_i(\neg(K_i\varphi \wedge \neg K_i\psi)) \vee K_i(\neg(K_i\psi \wedge \neg K_i\varphi))$ (De Morgan)
- The agent has introspection regarding the fact that what it knows and does not know is consistent. E.g., it knows that it cannot at the same time know that φ and not know that φ.

Axiom 4.2

$\neg K_i\neg K_i\varphi \rightarrow K_i\neg K_i\neg\varphi$

- Corollary: $K_i\hat{K}_i\neg\varphi \vee K_i\hat{K}\varphi$. Agents are only required to know whether they consider some proposition or its negation possible.

Mixing Knowledge and Belief

Building a logic with both Knowledge and Belief modalities interacting is non-trivial. Consider interaction axioms:

- Entailment property: $K_i\varphi \rightarrow B_i\varphi$
- Positive certainty property: $B_i\varphi \rightarrow B_iK_i\varphi$
- And let B be a **KD45** modality and K a **S5** modality. Then $\neg B_i\varphi \rightarrow B_i\neg K_i\varphi$ holds (seems reasonable):
  - $\neg B_i\varphi \Rightarrow_{\textbf{Entailment}} \neg K_i\varphi \Rightarrow_{\textbf{5}} K_i\neg K_i\varphi \Rightarrow_{\textbf{Entailment}} B_i\neg K_i\varphi$ $\square$
- Objection
  - Let p be a proposition the agent believes, but in fact p is false: $B_ip \wedge \neg p$
  - $\Rightarrow_{\textbf{Positive Certainty}} B_iK_ip$
  - $\Rightarrow_{\textbf{T}} \neg K_ip \Rightarrow_{\textbf{5}} K_i\neg K_ip \Rightarrow_{\textbf{Entailment}} B_i\neg K_ip$ ↯

Group Knowledge

- Extend language by three modal operators:
  - $E_G\varphi$: Everyone in the group G knows φ.
  - $C_G\varphi$: It is common knowledge among the agents in G that φ is the case.
  - $D_G\varphi$: It is distributed knowledge among the agents in G that φ is the case.
- Example
  - $K_3C_{\{1,2\}}p$: Agent 3 knows that it is common knowledge among agents 1 and 2 that p is the case.

Everyone Knows

- $M,w \models E_G\varphi$ iff. $M,w \models K_i\varphi$ for all $i \in G$.
- Write $E^0_G\varphi$ as an abbreviation of φ, $E^1_G\varphi$ as an abbreviation for $K_1\varphi \wedge K_1\varphi \wedge ...$, and let $E^{k+1}_G\varphi$ be an abbreviation for $E_GE^k_G\varphi$.

Def. G-reachable in k steps

A world w' is G-reachable from world w in $k \geq 1$ steps iff. there exists worlds $u_0u_1 \ldots u_k$ such that $u_0 = w$ and $u_k = w'$ and for all j with $0 \leq j \leq k-1$ there exists $i \in G$ s.th. $(u_j, u_{j+1}) \in R(K_i)$.

Lemma

$M,w \models E^k_G\varphi$ iff. $M,w' \models \varphi$ for all w' that are G-reachable from w in k steps.

Common Knowledge

- $M,w \models C_G\varphi$ iff. $M,w \models E^k_G\varphi$ for $k = 1,2,\ldots$.

Def. G-reachable

A world w' is G-reachable from world w iff. w' is G-reachable from w in k steps for some $k \geq 1$.

Lemma

$M,w \models C_G\varphi$ iff. $M,w' \models \varphi$ for all w' that are G-reachable from w.

Distributed Knowledge

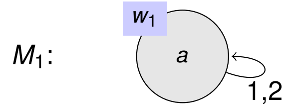
Distributed Knowledge

A group of agents G has distributed knowledge of φ iff the combined knowledge of the members of G implies φ. Idea: Eliminate all worlds that some agent in G considers impossible. Technically: Intersect the sets of worlds each agent in G considers possible. Hence:

- $M, w \models D_G\varphi$ iff. $M, w' \models \varphi$ for all w' s.th. $(w, w') \in \cap_{i \in G} K_i$

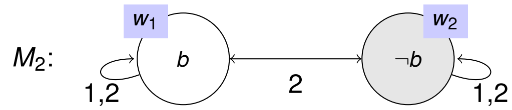
Theory of Mind Example Revisited

- First, it is common knowledge that the ball is in the box



$$M_1, w_1 \models E_{\{1,2\}}b \wedge C_{\{1,2\}}b \wedge D_{\{1,2\}}b$$

- Afterwards, the ball is not in the box. Agent 1 knows, Agent 2 does not.



$$M_2, w_2 \models \neg E_{\{1,2\}}\neg b \wedge \neg C_{\{1,2\}}\neg b \wedge D_{\{1,2\}}\neg b$$

Summary and Outlook

- Today
  - Critical discussion of the formal properties of knowledge and belief as modeled in modal logics. **S5** seems to be quasi-standard model for knowledge in computer science, there may be good reasons to make other choices for specific modeling purpose.
  - Foundations of group knowledge: Everyone knows, Common knowledge, and distributed knowledge.
- Next time: Dynamics and Puzzles

Literature

- J. Garson, 2004, https://plato.stanford.edu/entries/logic-modal/
- R. Fagin, J. Y. Halpern, Y. Moses, M. Y. Vardi, Reasoning about knowledge, The MIT Press, 1995.
- J. Symons, Logic and formal semantics for epistemology, in The Routledge Companion to Epistemology (Routledge Philosophy Companions) Duncan Pritchard and Sven Bernecker (eds.), New York: Routledge. 571–586, 2011.