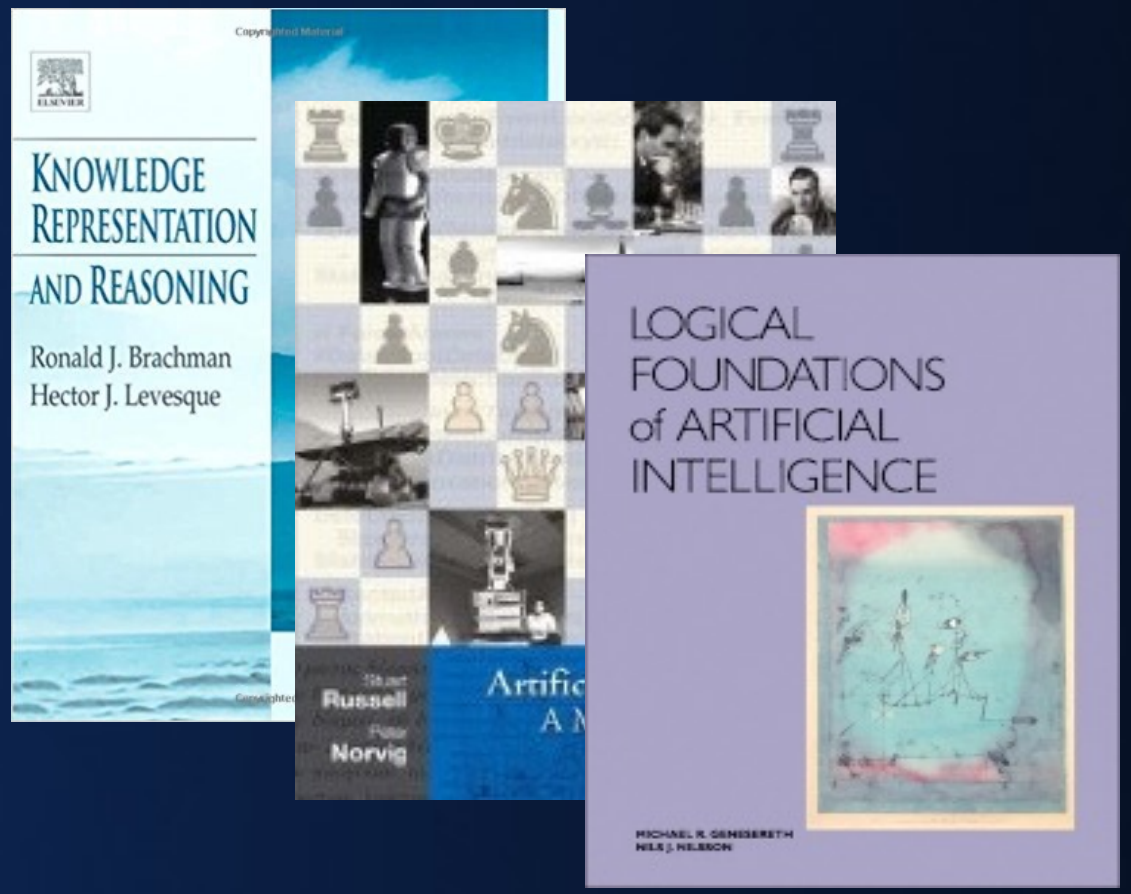


AI Fundamentals: Knowledge Representation and Reasoning

Maria Simi

Knowledge and beliefs

LESSON 4:REASONING ABOUT KNOWLEDGE AND BELIEFS

Multiple agents and their “attitudes”

Human intelligence is intrinsically social: humans need to negotiate and coordinate with other agents.

In multi-agent scenarios, we need methods for one agent to model **mental states** of other agents: high level representations of other agent’s belief, intentions and goals may be relevant for acting.

By *mental states* we mean the relation of an agent to a proposition.

Propositional attitudes that an agent can have include *Believes, Knows, Wants, Intends, Desires*, *Informs* … so called because the argument is a proposition.

Propositional attitudes do not behave as regular predicates.

Problem: referential transparency

Suppose we try to assert that "*Lois knows that Superman can fly*":

$Knows(Lois, CanFly(Superman))$

1. What is '*CanFly*'? *A predicate? A term?*
2. But since $Superman = Clark$, then we are able to reason as follows:

   $(Superman = Clark) \wedge Knows(Lois, CanFly(Superman)) \models$

   $Knows(Lois, CanFly(Clark))$ by the substitution of equal terms

This property is called **referential transparency**: what matters is the object that the term names, not the form of the term. Important property for reasoning in classical logic.

Propositional attitudes like *Believes* and *knows*, require **referential opacity** —the terms used do matter, because an agent may not be aware of which terms are **co-referential**.

Three approaches

1. **Reification**. We remain within FOL, as we did for the *situation calculus*, using terms to represent propositions [Mac Carthy]. Example: $Bel(a, On(b, \mathrm{c}))$. Referential transparency problem.
2. **Meta-linguistic representation**. We remain within FOL and represent propositions as strings. Example: $Bel(a, \text{“}On(b, \mathrm{c})\text{”})$.

In 1 and 2 problems are connecting the reified version of the proposition (a function or a string) and the proposition itself.

3. **Modal logics**. Propositional attitudes are represented as **modal operators** in specialized modal logics, with alternative semantics. Modal operators are an extension of classical logical operators.

   Example: $\mathbf{B}(a, On(b, \mathrm{c}))$ or $\mathbf{B}_A(On(b, \mathrm{c}))$, $\mathbf{K}_A(On(b, \mathrm{c}))$.

   to say that agent a believes/knows that block B is on C.

   Classical logic has only one modality (the *modality of truth*): P is the same as saying “P is true”

Modal logics

Strictly speaking modal logic is about **necessity** and **possibility**. However, the term is used more broadly to cover logics with different modelling goals.

- $\Box A$ It is necessary that A …
- $\Diamond A$ It is possible that A …

The simplest logic is called **K** (after Saul Kripke). **K** results from adding the following to the principles of propositional logic.

Necessitation Rule: If A is a theorem of **K**, then so is $\Box A$.

Distribution Axiom: $\Box(A \Rightarrow B) \Rightarrow (\Box A \Rightarrow \Box B)$

Note:

- $\Box$ some sort of universal quantification over interpretations
- $\Diamond$ some sort of existential quantification over interpretations

Other stronger logics

Logic **T** adds axiom

(*M*) $\Box A \Rightarrow A$

may be relevant for some modal operators and not for others

Example:

$Knows\ A \Rightarrow A$ seems plausible

$Bel\ A \Rightarrow A$ is not

Logic **S4** adds:

$\Box A \Rightarrow \Box\Box A$

Logic **S5** adds:

$\Diamond A \Rightarrow \Box\Diamond A$

Possible world semantics

Semantics for modal logics is defined wrt

1. a set W of possible worlds
2. an **accessibility relation** R between worlds

A formula A is now given an interpretation wrt a **possible world** w; we write $\mathcal{I}(A, w)$.

$(\neg)\ \mathcal{I}(\neg A, w) = T \quad iff \quad \mathcal{I}(A, w) = F$

$(\Rightarrow)\ \mathcal{I}(A \Rightarrow B, w) = T \quad iff \quad \mathcal{I}(A, w) = F$ or $\mathcal{I}(B, w) = T$

Classical

...

$(\Box)\ \mathcal{I}(\Box A, w) = T$ *iff* for **every** world w' in W such that wRw', $\mathcal{I}(A, w') = T$

$(\Diamond)\ \mathcal{I}(\Diamond A, w) = T$ *iff* for **some** world w' in W such that wRw', $\mathcal{I}(A, w') = T$

Different modal logics are defined according to the properties of the accessibility relation R (and corresponding axioms).

Modal logics and referential transparency

Modal logics address the problem of *referential transparency*, since the truth of a complex formula does not depend on the truth of the components in the same world/interpretation. Modal operators are not **compositional.**

Under possible worlds semantics it may be:

- ✓ $(Superman = Clark)$ is true in a world w
- ✓ $Knows(Lois, CanFly(Superman))$, i.e. $CanFly(Superman)$ in all the worlds accessible to Lois from w
- ✓ but not necessarily $Knows(Lois, CanFly(Clark))$, i.e. $CanFly(Clark)$ in all the accessible worlds

Modal logics for knowledge are easier than those of beliefs. We start with these.

Syntax of modal logic for knowledge

wff is an abbreviation for *Well Formed Formula;* $\mathbf{K}$ is the modal operator for *knowledge*

1. All the *wff* of ordinary FOL are also *wff* of the modal language
2. If Φ is a closed *wff* of the modal language and a is an agent, then $\mathbf{K}(a, \Phi)$ is a formula of the modal language.
3. If Φ and Ψ are *wff* so are the formulas that can be constructed from them with the usual logic connectives.

Examples:

$\mathbf{K}(A_1, \mathbf{K}(A_2, On(B, C)))$	A_1 knows that A_2 knows that B is on C.
$\mathbf{K}(A_1, On(B, C)) \vee \mathbf{K}(A_1, On(B, D))$	A_1 knows that B is on C or it knows that B is on D.
$\mathbf{K}(A_1, On(B, C) \vee On(B, D))$	A_1 knows that B is on C or that B is on D.
$\mathbf{K}(A_1, On(B, C)) \vee \mathbf{K}(A_1, \neg On(B, C))$	A_1 knows whether B is on C.
$\neg\mathbf{K}(A_1, On(B, C))$	A_1 does not know that B is on C.

Properties of knowledge

Desirable properties of knowledge:

- One agent can hold false beliefs but **cannot hold false knowledge;** if an agent knows something than this must be true. *Knowledge is justified true belief*.
- An agent **does not know all the truths**: something may be true without the agent knowing it.
- If two formulas Φ and Ψ are equivalent not necessarily $\mathbf{K}(A, \Phi)$ implies $\mathbf{K}(A, \Psi)$

The semantics of modal logic is given in terms of **possible worlds** and specific **accessibility relations** among them, one for each agent.

An agent knows a proposition just when that proposition is true in all the worlds accessible from the agent's world (those that the agent considers possible).

Possible world semantics of knowledge

Possible worlds roughly correspond to contexts within interpretations.

An **accessibility relation** (k for knowledge) is defined for agents and connects possible worlds:

if $k(a, w_i, w_j)$ is satisfied, then world w_j is **accessible** from world w_i for agent a.

Semantics:

1. Regular wffs (with no modal operators) are not simply true or false but they are **true or false wrt a possible world**.

   $\mathcal{I}(w_1, \Phi)$ may be different from $\mathcal{I}(w_2, \Phi)$
2. A modal formula $\boldsymbol{K}(a, \Phi)$ is true in w iff Φ is true in **all** the worlds accessible from w for agent a.
3. The semantics of complex formulas is determined by regular truth recursive rules.

Possible worlds semantics: visualization

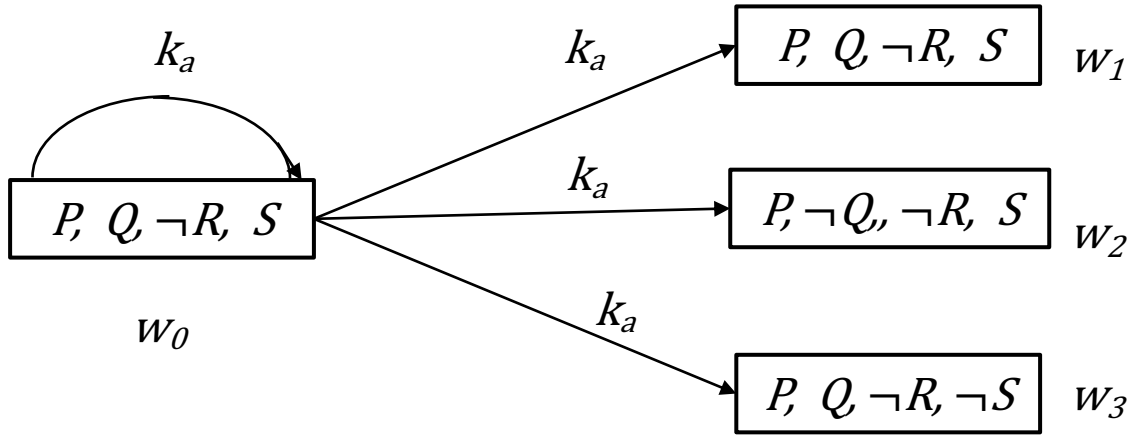
$K(a, \Phi)$ means that agent a knows the proposition denoted by Φ.

"**Not knowing Φ**" in w_0 (a specific world) is modelled by allowing worlds, accessible from w_0, in which Φ is *true* and some worlds in which Φ is *false*

Example: in the scenario represented below, where arrows represent a's accessibility rel,

$K(a, P)$ and $K(a, \neg R)$ in w_0 since P and $\neg R$ are true in worlds w_0, w_1, w_2 and w_3

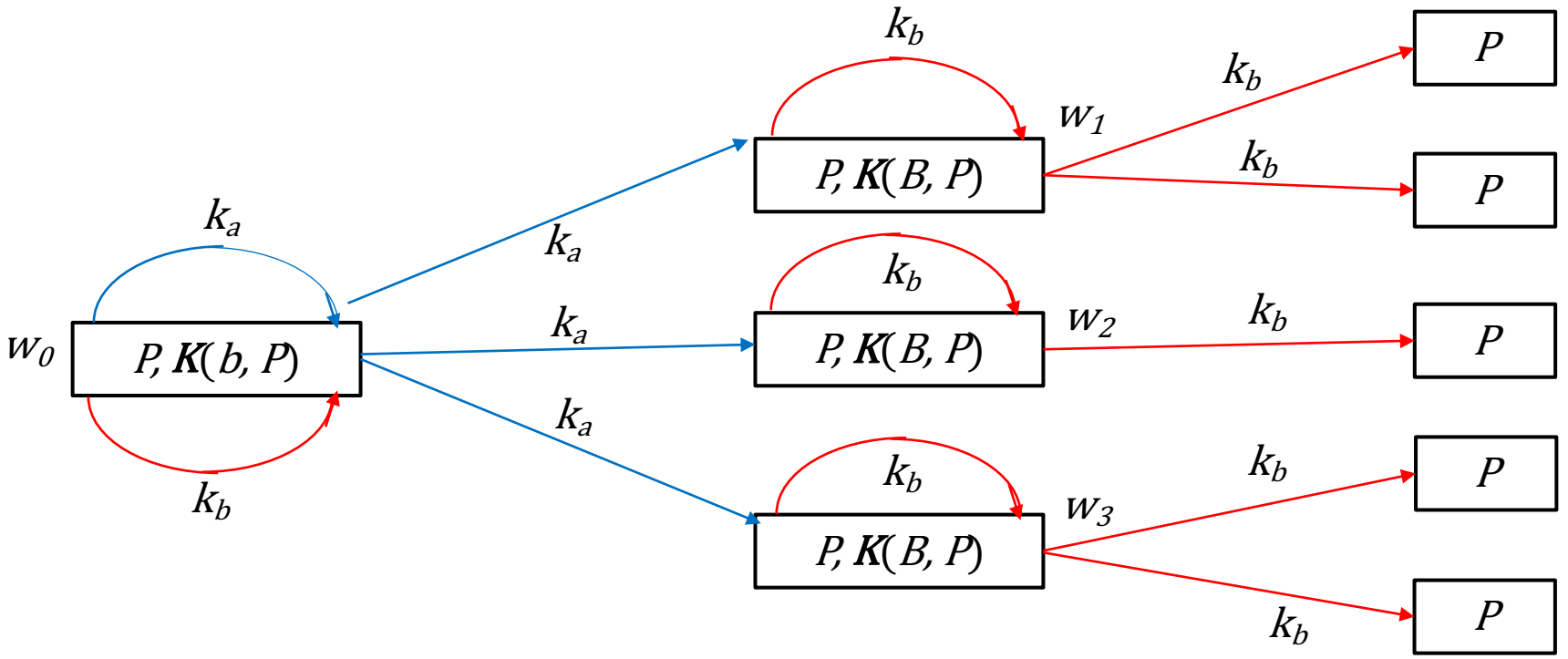
but $K(a, Q)$ is false in w_0



Nested knowledge statements

The accessibility relation also accounts for nested knowledge statements, also involving different agents.

$\boldsymbol{K}(\boldsymbol{a}, \boldsymbol{K}(b, P))$ holds in w_0 since $\boldsymbol{K}(b, P)$ holds in w_0, w_1, w_2 and w_3 accessible to a



Properties and axioms for knowledge - 1

Many of the properties that we desire for knowledge can be achieved by imposing constraints to the accessibility relation.

1. Agents should be able to reason with the knowledge they have

   $$\mathbf{K}(a, \alpha \Rightarrow \beta) \Rightarrow (\mathbf{K}(a, \alpha) \Rightarrow \mathbf{K}(a, \beta)) \qquad (\textit{Distribution axiom})$$

   This is implicit in possible world semantics.

2. Agents cannot have false knowledge (different for beliefs):

   $$\mathbf{K}(a, \alpha) \Rightarrow \alpha \qquad (\textit{Knowledge axiom})$$

   The knowledge axiom is satisfied if the accessibility relation is **reflexive**, i.e. $k(a, w, w)$ for every a and every w. An implication is that: $\neg\mathbf{K}(a, \textbf{\textit{false}})$.

   Moreover, reflexivity implies that there is at least a world accessible from w, i.e. the relation is also **serial.**

Knowledge axioms - 2

3. It is also reasonable to assume that if an agent knows something, then it knows that it knows

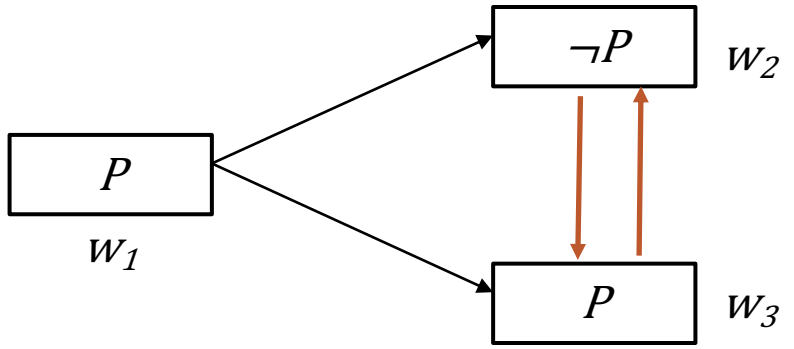
$$\mathbf{K}(a, \alpha) \Rightarrow \mathbf{K}(a, \mathbf{K}(a, \alpha)) \qquad (\textit{Positive introspection})$$

The accessibility relation must be **transitive**, i.e. $k(a, w_1, w_2)$ and $k(a, w_2, w_3)$ implies $k(a, w_1, w_3)$

4. In some axiomatization we also assume that if an agent doesn't know something, then it knows that it doesn't know it.

$$\neg\, \mathbf{K}(a, \alpha) \Rightarrow \mathbf{K}(a, \neg\mathbf{K}(\mathrm{a}, \alpha)) \qquad (\textit{Negative introspection})$$

The accessibility relation must be **Euclidean**, i.e. $k(a, w_1, w_2)$ and $k(a, w_1, w_3)$ implies $k(a, w_2, w_3)$



Knowledge axioms - 3

5. It is also intrinsic in possible world semantics that an agent knows all the logical theorems, including the ones characterizing knowledge.

   From $\vdash \alpha$ infer $\mathbf{K}(a, \alpha)$ (*Epistemic necessitation rule*)

6. From 1 and 5, in the propositional case we also get the rule:

   From $\alpha \vdash \beta$ and from $\mathbf{K}(a, \alpha)$ infer $\mathbf{K}(a, \beta)$ (*Logical omniscience*)

   From $\vdash \alpha \Rightarrow \beta$ infer $\mathbf{K}(a, \alpha) \Rightarrow \mathbf{K}(a, \beta)$ (*Logical omniscience*)

   Logical omniscience is considered problematic: we are assuming unbounded reasoning capabilities. As a corollary of logical omniscence:

   $\mathbf{K}(a, \alpha \wedge \beta) \equiv \mathbf{K}(a, \alpha) \wedge \mathbf{K}(a, \beta)$ (**K** *distribution over and*)

   It is not the case however that $\mathbf{K}(a, \alpha \vee \beta) \equiv \mathbf{K}(a, \alpha) \vee \mathbf{K}(a, \beta)$

Modal logics of knowledge

Modal epistemic logics are obtained with various combinations of axioms 1-4 plus inference rule 5:

- System K: axiom 1
- System T: axioms 1-2
- Logic S4: axioms 1-3
- Logic S5: axioms 1-4 (perfect reasoner)

Not any combination is possible since the properties of accessibility relations are interdependent. For example:

- Reflexive implies serial.
- If a relation is reflexive and Euclidian it is also transitive: axiom 2 and 4 imply 3.
- …

Properties and axioms for beliefs

Since an agent can hold wrong beliefs the knowledge axiom is not appropriate.

We include as axiom the following instead:

$\neg \mathbf{B}(a, \textbf{\textit{False}})$ (*lack of contradictions*)

The *distribution axiom* and the *necessitation rule* are controversial, since an agent cannot realistically believe all the logical consequences of its beliefs but only those that he is able to derive (limited/bounded rationality).

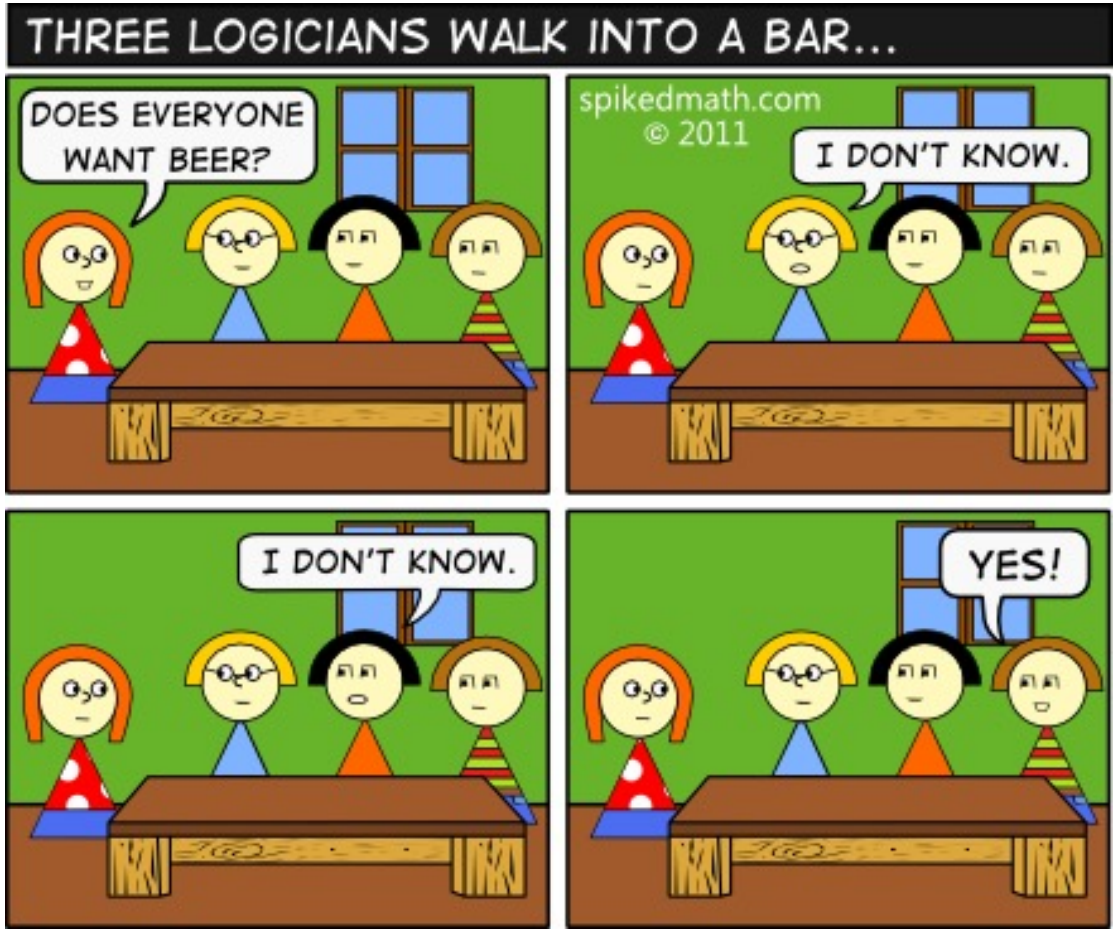
$\mathbf{B}(a, \alpha) \Rightarrow \mathbf{B}(a, \mathbf{B}(a, \alpha))$ (*Positive introspection*)

$\mathbf{B}(a, \alpha) \Rightarrow \mathbf{K}(a, \mathbf{B}(a, \alpha))$ also reasonable

Negative introspection is problematic. While the following special case of the knowledge axioms is safe:

$\mathbf{B}(a, \mathbf{B}(a, \alpha)) \Rightarrow \mathbf{B}(\mathrm{a}, \alpha)$ … and so on

The wise-men puzzle



Three wise man who are told by their king that at least one of them has a white spot in his forehead; actually all three have white spots.

Each wise-man can see the other's foreheads but not his own.

The first wise man says "I don't know whether I have a white spot".

The second wise man says "I don't know whether I have a white spot".

The third wise man can then conclude that he has a white spot.

The proof for **two** wise men

The two wise men are called A and B. The following facts are given, after B speaks:

1.	$\mathbf{K}_A(\neg White(A) \Rightarrow \mathbf{K}_B(\neg White(A))$	B can see A's forehead, and A knows it.
2.	$\mathbf{K}_A(\mathbf{K}_B(\neg White(A) \Rightarrow White(B)))$	At least one is white
3.	$\mathbf{K}_A(\neg\mathbf{K}_B(White(B)))$	B does not know the color on his forehead
4.	$\mathbf{K}_A(\mathbf{K}_B(\neg White(A)) \Rightarrow \mathbf{K}_B(White(B)))$	2 and Distribution axiom, in A's mind
5.	$\mathbf{K}_A(\neg White(A) \Rightarrow \mathbf{K}_B(White(B)))$	from 1 and 4, by transitivity
6.	$\mathbf{K}_A(\neg\mathbf{K}_B(White(B)) \Rightarrow White(A))$	5, contrapositive
7.	$\mathbf{K}_A(White(A))$	3, 6 Modus Ponens

Autoepistemic logic

… AT THE INTERSECTION BETWEEN MODAL LOGICS FOR BELIEFS AND NONMONOTONIC LOGIC

Autoepistemic logic for nonmonotonic reasoning

$\frac{\alpha : \beta}{\gamma}$ in default logic is read as "if α and *it is consistent to believe* β then γ"

A different approach to nonmonotonic reasoning is to model "*it is consistent to believe* β" as the lack of belief in $\neg\beta$ with a suitable logic for belief.

Autoepistemic logic uses a belief operator $\mathbf{B}$. $\mathbf{B}\alpha$ stands for "α is believed to be true".

The $\mathbf{B}$ operator could be used to represent defaults, for example, as follows:

$$\forall x \; Bird(x) \wedge \neg\mathbf{B}\neg Flies(x) \Rightarrow Flies(x)$$

Any bird not believed to be unable to flight, does fly.

Note that: $\mathbf{B}\neg Flies(x)$ is different from $\neg Flies(x)$

Autoepistemic logic

Given a KB that contains sentences with the $\mathbf{B}$ "autoepistemic" operator, what is a reasonable set of beliefs E to hold?

Minimal properties for a **set of beliefs** E to be considered **stable**:

1. Closure under entailment: if $E \models \alpha$, then $\alpha \in E$
2. Positive introspection: if $\alpha \in E$, then $\mathbf{B}\alpha \in E$
3. Negative introspection: if $\alpha \notin E$, then $\neg\mathbf{B}\alpha \in E$

This leads to the following definition of **stable expansion of a KB** (a **minimal** set satisfying 1-3)

Stable expansion of the KB [Moore]: A set E is a stable expansion of KB if and only if for every sentence π, it is the case that:

$$\pi \in E \textit{ iff } KB \cup \{\mathbf{B}\alpha \mid \alpha \in E\} \cup \{\neg\mathrm{B}\alpha \mid \alpha \notin E\} \models \pi$$

---------- Δ ---------- *assumed beliefs*

The implicit beliefs E are those sentences that are entailed by KB plus the assumptions deriving from positive and negative introspection.

Default reasoning with stable expansions

Example:

1. $Bird(Chilly)$, $Bird(Tweety)$, $Tweety \neq Chilly$, $\neg Flies(Chilly)$
2. $\forall x\, Bird(x) \wedge \neg \mathbf{B} \neg Flies(x) \Rightarrow Flies(x)$

Can we infer $Flies(Tweety)$?

$\mathrm{KB} \nvDash \neg Flies(Tweety)$ then $\neg Flies(Tweety) \notin E$

$\neg \mathbf{B} \neg Flies(Tweety)$ must be assumed by negative introspection

$Flies(Tweety)$ by $(1, 2)$

Can we infer $Flies(Chilly)$?

$\mathrm{KB} \vDash \neg Flies(Chilly)$

$\mathbf{B} \neg Flies\,(Chilly)$

Enumerating and checking expansions

Example (**one stable expansion**):

$Bird(Chilly)$, $Bird(Tweety)$, $(Tweety \neq Chilly)$, $\neg Flies(Chilly)$, objective subset

$Bird(Tweety) \wedge \neg \mathbf{B} \neg Flies(Tweety) \Rightarrow Flies(Tweety)$

$Bird(Chilly) \wedge \neg \mathbf{B} \neg Flies(Chilly) \Rightarrow Flies(Chilly)$

} ground instances of the general rule $\forall x\, Bird(x) \wedge \neg \mathrm{B} \neg Flies(x) \Rightarrow Flies(x)$

Subjective subset: $\mathrm{B} \neg Flies(Tweety)$, $\mathrm{B} \neg Flies(Chilly)$

Four cases for the **subjective subset**: *check*

1. $\mathbf{B} \neg Flies(Tweety)$ *true* and $\mathbf{B} \neg Flies(Chilly)$ *true* ➢ $\mathrm{KB} \models \neg Flies(Tweety)$? No. *FAIL*
2. $\mathbf{B} \neg Flies(Tweety)$ *true* and $\mathbf{B} \neg Flies(Chilly)$ *false* ➢ $\mathrm{KB} \not\models \neg Flies(Tweety)$; $\mathrm{KB} \models \neg Flies(Chilly)$ FAIL
3. $\mathbf{B} \neg Flies(Tweety)$ *false* and $\mathbf{B} \neg Flies(Chilly)$ *true* ➢ $\mathrm{KB} \models Flies(Tweety)$; $\mathrm{KB} \models \neg Flies(Chilly)$ OK
4. $\mathbf{B} \neg Flies(Tweety)$ *false* and $\mathbf{B} \neg Flies(Chilly)$ *false* ➢ $\mathrm{KB} \models Flies(Tweety)$; $\mathrm{KB} \models \neg Flies(Chilly)$ FAIL

Case 3 is the unique stable espansion.

Stable expansions: other cases

The KB consisting of the sentence $(\neg \mathbf{B}p \Rightarrow p)$ has **no stable expansion**:

- ✓ If $\mathbf{B}p$ is false, then the expansion entails p;
- ✓ if $\mathbf{B}p$ is true, then the expansion does not include p .

The KB consisting of the sentences $(\neg \mathrm{B}p \Rightarrow q)$ and $(\neg \mathrm{B}q \Rightarrow p)$ has exactly **two stable expansions:**

- ✓ If $\mathbf{B}p$ is true and $\mathbf{B}q$ false, the KB entails p and does not entail q, and so this is the first stable expansion
- ✓ the other stable expansion is when $\mathbf{B}p$ is false and $\mathbf{B}q$ is true
- ✓ if $\mathbf{B}p$ is true and $\mathbf{B}q$ true, but the KB does not entail p nor q
- ✓ if $\mathbf{B}p$ is false and $\mathbf{B}q$ is false, then KB entails p and q.

Conclusions

- ✓ In multi-agent environments there is a need to represent and reason about other agent's propositional attitudes.
- ✓ We have reviewed modal logics, based on possible world semantics, and discussed the properties that are appropriate for knowledge and beliefs.
- ✓ Auto-epistemic logic can be regarded as another approach to nonmonotonic reasoning.
- Reason Maintenance systems were designed to support efficiently beliefs revision on behalf of a problem solver [skipped].
- An approach similar to the stable expansions of autoepistemic logic will be used in **Answer Set Programming**, a "modern" reasoning class of systems which are an extension of logic programming.

References

Stuart J. Russell and Peter Norvig. *Artificial Intelligence: A Modern Approach* (3rd edition). Pearson Education 2010 (Ch. 12)

Genesereth, M., and Nilsson, N., Logical Foundations of Artificial Intelligence, San Francisco: Morgan Kaufmann, 1987 (Ch. 9).

Ronald Brachman and Hector Levesque. *Knowledge Representation and Reasoning*. Morgan Kaufmann Publishers Inc., San Francisco, CA, USA. 2004 (Ch. 11.5)