INF02511: Knowledge Engineering

Reasoning about Knowledge
(a very short introduction)

Iyad Rahwan

Overview

- The partition model of knowledge
- Introduction to modal logic
- The *S5* axioms
- Common knowledge
- Applications to robotics
- Knowledge and belief

The Muddy Children Puzzle

- *n* children meet their father after playing in the mud. The father notices that *k* of the children have mud on their foreheads.
- Each child sees everybody else's foreheads, but not his own.
- The father says: "*At least one of you has mud on his forehead*."
- The father then says: "*Do any of you know that you have mud on your forehead? If you do, raise your hand now.*"
- No one raises his hand.
- The father repeats the question, and again no one moves.
- After exactly *k* repetitions, all children with muddy foreheads raise their hands simultaneously.

Muddy Children (cont.)

- Suppose $k = 1$
- The muddy child knows the others are clean
- When the father says at least one is muddy, he concludes that it's him

Muddy Children (cont.)

- Suppose $k = 2$
- Suppose you are muddy
- After the first announcement, you see another muddy child, so you think perhaps he’s the only muddy one.
- But you note that this child did not raise his hand, and you realise you are also muddy.
- So you raise your hand in the next round, and so does the other muddy child

The Partition Model of Knowledge

- An n-agent a partition model over language Σ is $A=(W, \pi, I_1, \ldots, I_n)$ where
  - W is a set of possible worlds
  - $\pi : \Sigma \rightarrow 2^W$ is an interpretation function that determines which sentences are true in which worlds
  - Each I_i is a partition of W for agent i
    - Remember: a partition chops a set into disjoint sets
    - $I_i(w)$ includes all the worlds in the partition of world w

Partition Model (cont.)

- What?
  - Each I_i is a partition of W for agent i
    - Remember: a partition chops a set into disjoint sets
    - $I_i(w)$ includes all the worlds in the partition of world w
- Intuition:
  - if the actual world is w, then $I_i(w)$ is the set of worlds that agent i cannot distinguish from w
  - i.e. all worlds in $I_i(w)$ all possible as far as i knows

Partition Model (cont.)

- Suppose there are two propositions *p* and *q*
- There are 4 possible worlds:
  - w_1: $p \wedge q$
  - w_2: $p \wedge \neg q$
  - w_3: $\neg p \wedge q$
  - w_4: $\neg p \wedge \neg q$
- Suppose the real world is w_1, and that in w_1 agent *i* cannot distinguish between w_1 and w_2
- We say that $I_i(w_1)=\{w_1, w_2\}$

The Knowledge Operator

- Let $K_i\varphi$ mean that "agent i knows that φ"
- Let $A=(W, \pi, I_1, \ldots, I_n)$ be a partition model over language Σ and let $w \in W$
- We define logical entailment |= as follows:
  - For $\varphi \in \Sigma$ we say $(A,w \models \varphi)$ if and only if $w \in \pi(\varphi)$
  - We say $A,w \models K_i\varphi$ if and only if $\forall w'$,

    if $w' \in I_i(w)$, then $A,w \models \varphi$

The Knowledge Operator (cont.)

- What?
  - We say $A,w \models K_i\varphi$ if and only if $\forall w'$,

    if $w' \in I_i(w)$, then $A,w \models \varphi$

- <u>Intuition:</u> in partition model A, if the actual world is w, agent i knows φ if and only if φ is true in all worlds he cannot distinguish from w

Muddy Children Revisited

- *n* children meet their father after playing in the mud. The father notices that *k* of the children have mud on their foreheads.
- Each child sees everybody else's foreheads, but not his own.

Muddy Children Revisited (cont.)

- Suppose $n = k = 2$ (two children, both muddy)
- Possible worlds:
  - w_1: muddy1 $\wedge$ muddy2 (actual world)
  - w_2: muddy1 $\wedge \neg$ muddy2
  - w_3: $\neg$ muddy1 $\wedge$ muddy2
  - w_4: $\neg$ muddy1 $\wedge \neg$ muddy2
- At the start, no one sees or hears anything, so all worlds are possible for each child
- After seeing each other, each child can tell apart worlds in which the other child's state is different

Muddy Children Revisited (cont.)

<u>Note:</u> in w_1 we have:

K_1 muddy2

K_2 muddy1

$K_1 \neg K_2$ muddy2

…

But we don't have:

K_1 muddy1

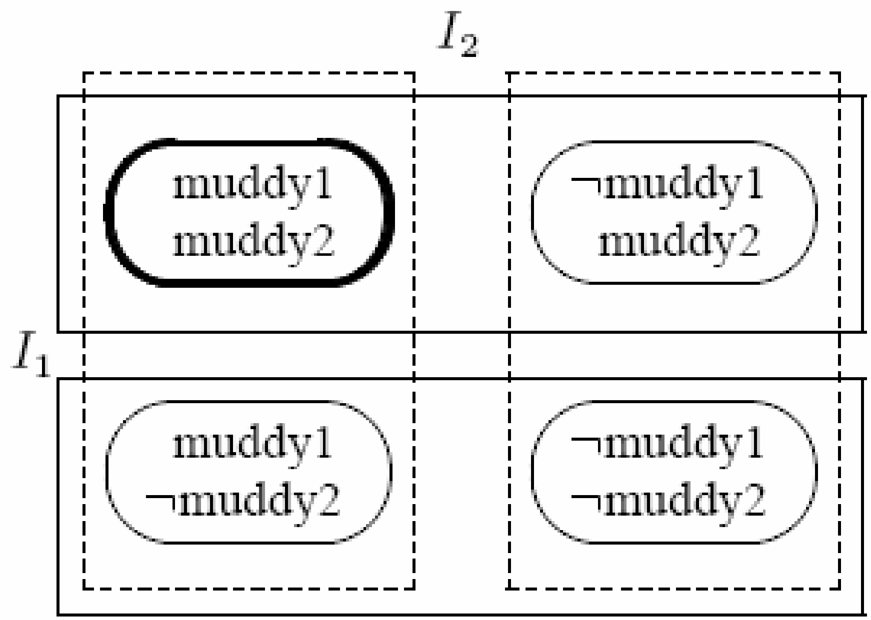


Figure 13.1: Partition model after the children see each other.

Bold oval = actual world

Solid boxes = equivalence classes in I_1

Dotted boxes = equivalence classes in I_2

Muddy Children Revisited (cont.)

- The father says: "*At least one of you has mud on his forehead*."
  - This eliminates the world:

    w_4: $\neg$ muddy1 $\wedge \neg$ muddy2

Muddy Children Revisited (cont.)

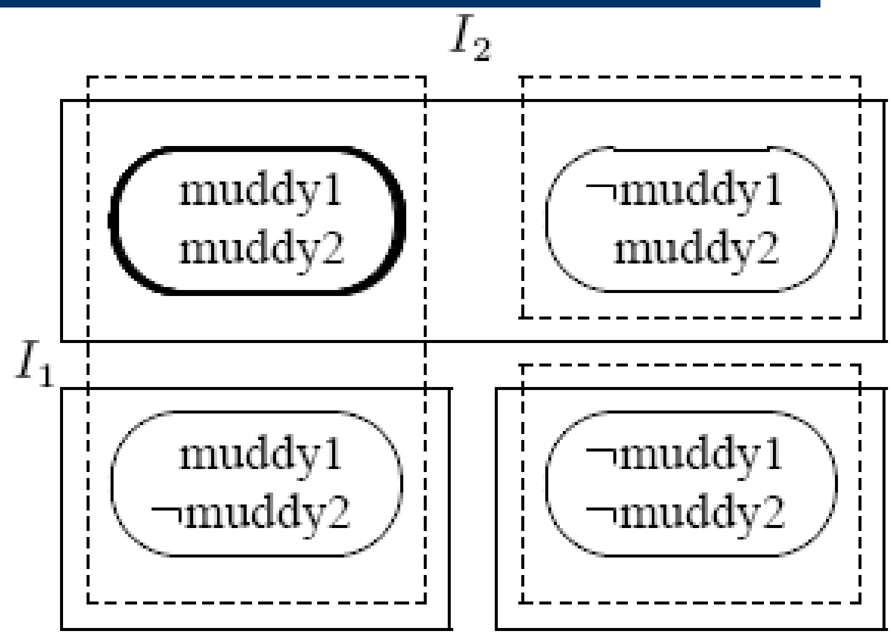


Figure 13.2: Partition model after the father's announcement.

Bold oval = actual world

Solid boxes = equivalence classes in I_1

Dotted boxes = equivalence classes in I_2

Muddy Children Revisited (cont.)

- The father then says: "*Do any of you know that you have mud on your forehead? If you do, raise your hand now.*"
  - Here, no one raises his hand.
  - But by observing that the other did not raise his hand (i.e. does not know whether he's muddy), each child concludes the true world state.
  - So, at the second announcement, they both raise their hands.

Muddy Children Revisited (cont.)

Note: in w_1 we have:

K_1 muddy1

K_2 muddy2

$K_1 K_2$ muddy2

…

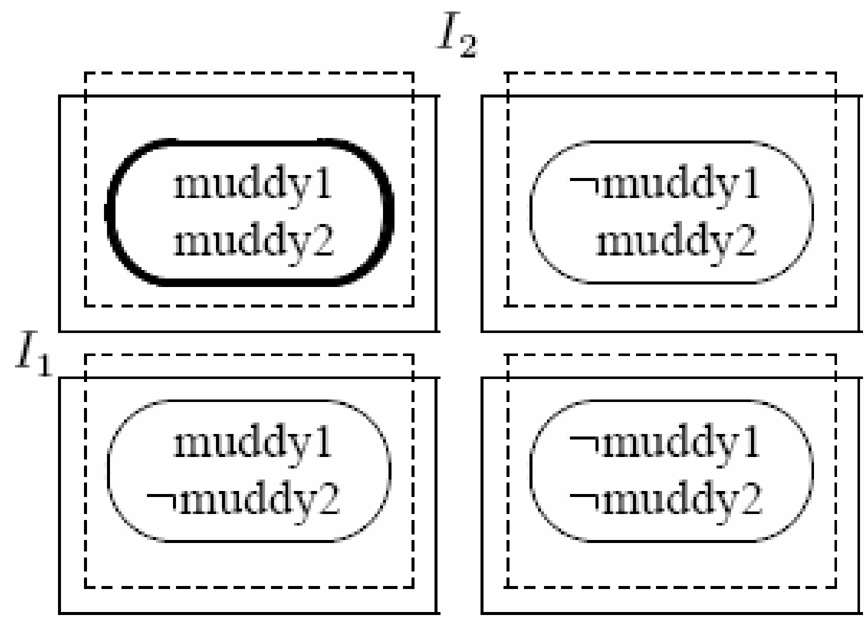


Figure 13.3: Final partition model.

Bold oval = actual world

Solid boxes = equivalence classes in I_1

Dotted boxes = equivalence classes in I_2

Modal Logic

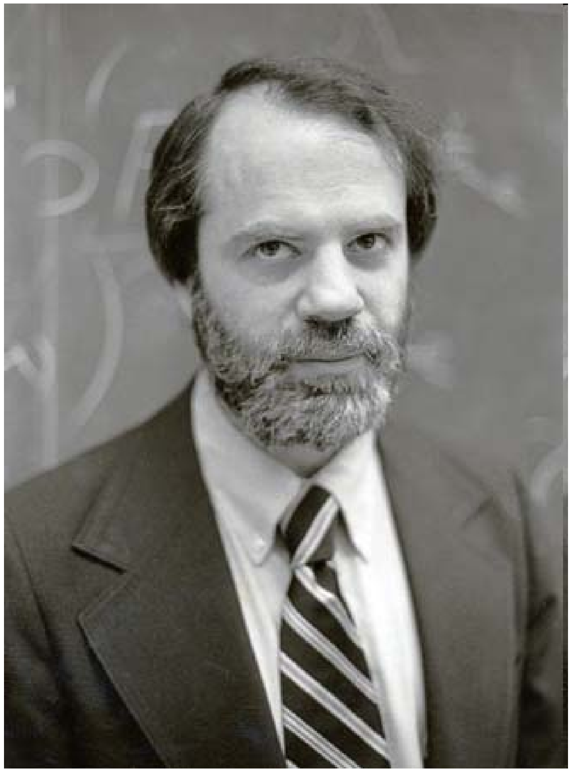
- Can be built on top of any language
- Two modal operators:
  - $\Box\varphi$ reads "φ is necessarily true"
  - $\Diamond\varphi$ reads "φ is possibly true"
- Equivalence:
  - $\Diamond\varphi \equiv \neg\Box\neg\varphi$
  - $\Box\varphi \equiv \neg\Diamond\neg\varphi$
- So we can use only one of the two operators

Modal Logic: Syntax

- Let P be a set of propositional symbols
- We define modal language $\mathcal{L}$ as follows:
- If $p \in P$ and φ, $\psi \in \mathcal{L}$ then:
  - $p \in \mathcal{L}$
  - $\neg\varphi \in \mathcal{L}$
  - $\varphi \wedge \psi \in \mathcal{L}$
  - $\Box\varphi \in \mathcal{L}$
- Remember that $\Diamond\varphi \equiv \neg\Box\neg\varphi$, and $\varphi \vee \psi \equiv \neg(\neg\varphi \wedge \neg\psi)$ and $\varphi \rightarrow \psi \equiv \neg\varphi \vee \psi$

Modal Logic: Semantics

- Semantics is given in terms of Kripke Structures (also known as possible worlds structures)
- Due to American logician Saul Kripke, City University of NY
- A Kripke Structure is (W, R)
  - W is a set of possible worlds
  - $R : W \times W$ is an binary accessibility relation over W



Modal Logic: Semantics (cont.)

- A Kripke model is a pair M,w where
  - $M = (W, R)$ is a Kripke structure and
  - $w \in W$ is a world
- The entailment relation is defined as follows:
  - $M,w \models \varphi$ if φ is true in w
  - $M,w \models \varphi \wedge \psi$ if $M,w \models \varphi$ and $M,w \models \psi$
  - $M,w \models \neg\varphi$ if and only if we do not have $M,w \models \varphi$
  - $M,w \models \Box\varphi$ if and only if $\forall w' \in W$ such that $R(w,w')$ we have $M,w' \models \varphi$

Modal Logic: Semantics (cont.)

- As in classical logic:
  - Any formula φ is valid (written $\models \varphi$) if and only if φ is true in all Kripke models

    E.g. $\Box\varphi \vee \neg\Box\varphi$ is valid
  - Any formula φ is satisfiable if and only if φ is true in some Kripke models
- We write $M, \models \varphi$ if φ is true in all worlds of M

Modal Logic: Axiomatics

- Is there a set of minimal axioms that allows us to derive precisely all the valid sentences?
- Some well-known axioms:
  - **Axiom(Classical)** All propositional tautologies are valid
  - **Axiom (K)** $(\Box\varphi \wedge \Box(\varphi \to \psi)) \to \Box\psi$ is valid
  - **Rule (Modus Ponens)** if φ and $\varphi \to \psi$ are valid, infer that ψ is valid
  - **Rule (Necessitation)** if φ is valid, infer that $\Box\varphi$ is valid

Modal Logic: Axiomatics

- Refresher: remember that
  - A set of inference rules (i.e. an inference procedure) is sound if everything it concludes is true
  - A set of inference rules (i.e. an inference procedure) is complete if it can find all true sentences
- Theorem: System **K** is sound and complete for the class of all Kripke models.

Multiple Modal Operators

- We can define a modal logic with *n* modal operators $\Box_1, \ldots, \Box_n$ as follows:
  - We would have a single set of worlds W
  - n accessibility relations $R_1, \ldots, R_n$
  - Semantics of each $\Box_i$ is defined in terms of R_i

Axiomatic theory of the partition model

- Objective: Come up with a sound and complete axiom system for the partition model of knowledge.
- Note: This corresponds to a more restricted set of models than the set of all Kripke models.
- In other words, we will need more axioms.

Axiomatic theory of the partition model

- The modal operator $\Box_i$ becomes K_i
- Worlds accessible from w according to R_i are those indistinguishable to agent i from world w
- K_i means "agent i knows that"
- Start with the simple axioms:
  - **(Classical)** All propositional tautologies are valid
  - **(Modus Ponens)** if φ and $\varphi \rightarrow \psi$ are valid, infer that ψ is valid

Axiomatic theory of the partition model (More Axioms)

- **(K)** From $(K_i\varphi \wedge K_i(\varphi \rightarrow \psi))$ infer $K_i\psi$
  - Means that the agent knows all the consequences of his knowledge
  - This is also known as logical omniscience
- **(Necessitation)** From φ, infer that $K_i\varphi$
  - Means that the agent knows all propositional tautologies

Axiomatic theory of the partition model (More Axioms)

- **Axiom (D)** $\neg K_i(\varphi \wedge \neg\varphi)$
  - This is called the axiom of consistency
- **Axiom (T)** $(K_i\,\varphi) \rightarrow \varphi$
  - This is called the veridity axiom
  - Means that if an agent cannot know something that is not true.
  - Corresponds to assuming that R_i is reflexive

Axiomatic theory of the partition model (More Axioms)

- **Axiom (4)** $K_i \varphi \rightarrow K_i K_i \varphi$
  - Called the positive introspection axiom
  - Corresponds to assuming that R_i is transitive
- **Axiom (5)** $\neg K_i \varphi \rightarrow K_i \neg K_i \varphi$
  - Called the negative introspection axiom
  - Corresponds to assuming that R_i is Euclidian
- Refresher: Binary relation R over domain Y is Euclidian if and only if $\forall y, y', y'' \in Y$, if $(y,y') \in R$ and $(y,y'') \in R$ then $(y',y'') \in R$

Axiomatic theory of the partition model (Overview of Axioms)

Name	Axiom	Accessibility Relation
Axiom K	$(K_i(\varphi) \wedge K_i(\varphi \rightarrow \psi)) \rightarrow K_i(\psi)$	NA
Axiom D	$\neg K_i(p \wedge \neg p)$	Serial
Axiom T	$K_i\varphi \rightarrow \varphi$	Reflexive
Axiom 4	$K_i\varphi \rightarrow K_iK_i\varphi$	Transitive
Axiom 5	$\neg K_i\varphi \rightarrow K_i\neg K_i\varphi$	Euclidean

Table 13.1: Axioms and corresponding constraints on the accessibility relation.

Proposition: a binary relation is an equivalence relation if and only if it is reflexive, transitive and Euclidean

Axiomatic theory of the partition model (back to the partition model)

- System **KT45** exactly captures the properties of knowledge defined in the partition model
- System **KT45** is also known as **S5**
- **S5** is sound and complete for the class of all partition models

The Coordinated Attack Problem
(aka, Two Generals' or Warring Generals Problem)

- Two generals standing on opposite hilltops, trying to coordinate an attack on a third general in a valley between them.
- Communication is via messengers who must travel across enemy lines (possibly get caught).
- If a general attacks on his own, he loses.
- If both attack simultaneously, they win.
- What protocol can ensure simultaneous attack?

The Coordinated Attack Problem

The Coordinated Attack Problem (A Naive Protocols)

- Let us call the generals:
  - *S* (sender)
  - *R* (receiver)
- Protocol for general *S*:
  - Send an “attack” message to *R*
  - Keeps sending until acknowledgement is received
- Protocol for general *R*:
  - Do nothing until he receives a message “attack” from *S*
  - If you receive a message, send an acknowledgement to *S*

The Coordinated Attack Problem (States)

- State of general *S*:
  - A pair (msg_S, ack_S) where msg $\in$ {0,1}, ack $\in$ {0,1}
  - msg_S = 1 means a message "attack" was sent
  - ack_S = 1 means an acknowledgement was received
- State of general *R*:
  - A pair (msg_R, ack_R) where msg $\in$ {0,1}, ack $\in$ {0,1}
  - msg_R = 1 means a message "attack" was received
  - ack_R = 1 means an acknowledgement was sent
- Global state: $<(msg_S, ack_S),(msg_R, ack_R)>$
- 4 possible local states per general &16 global states

The Coordinated Attack Problem (Possible Worlds)

- Initial global state: <(0,0),(0,0)>
- State changes as a result of:
  - Protocol events
  - Nondeterministic effects of nature
- Change in states captured in a history
- Example:
  - *S* sends a message to *R*, *R* receives it and sends an acknowledges, which is then received by *S*
  - <(0,0),(0,0)>, <(1,0),(1,0)>, <(1,1),(1,1)>
- In our model: **possible world = possible history**

The Coordinated Attack Problem (Indistinguishable Worlds)

- Defining the accessibility relation R_i:
  - Two histories are indistinguishable **to agent *i*** if their final global states have identical ***local states*** for agent *i*
- Example: world

  <(0,0),(0,0)>, <(1,0),(1,0)>, <(1,0),(1,1)>

  is indistinguishable to general *S* from this world:

  <(0,0),(0,0)>, <(1,0),(0,0)>, <(1,0),(0,0)>

  - <u>In words:</u> *S* sends a message to *R*, but does not get an acknowledgement. This could be because *R* never received the message, or because he did but his acknowledgement did not make reach *S*

The Coordinated Attack Problem (What do generals know?)

- Suppose the actual world is:
  - <(0,0),(0,0)>, <(1,0),(1,0)>, <(1,1),(1,1)>
- In this world, the following hold:
  - $K_S$attack
  - $K_R$attack
  - $K_S K_R$attack
- Unfortunately, this ***also*** holds:
  - $\neg K_R K_S K_R$attack
- *R* does not known that *S* knows that *R* knows that *S* intends to attack. Why? Because, from *R*'s perspective, the message could have been lost

The Coordinated Attack Problem (What do generals know?)

- Possible solution:
  - S acknowledges R's acknowledgement
- Then we have:
  - $K_R K_S K_R$attack
- Unfortunately, we ***also*** have:
  - $\neg K_S K_R K_S K_R$attack
- Is there a way out of this?

The "Everyone Knows" Operator

- $E_G\varphi$ denotes that everyone in group G knows φ
- **Semantics of "everyone knows":**

  Let:
  - M be a Kripke structure
  - w be a possible world in M
  - G be a group of agents
  - φ be a sentence of modal logic

  $M,w \models E_G\varphi$ if and only if $\forall i \in G$ we have $M,w \models K_i\varphi$

The "Common Knowledge" Operator

- When we say something is common knowledge, we mean that any fool knows it!
- If any fool knows φ, we can assume that everyone knows it, and everyone knows that everyone knows that everyone knows it, and so on (infinitely).

The “Common Knowledge” Operator (formal definition)

- $C_G\varphi$ denotes that φ is common knowledge among G
- **Semantics of “common knowledge”:**

  Let:

  - M be a Kripke structure
  - w be a possible world in M
  - G be a group of agents
  - φ be a sentence of modal logic

  $M,w \models C_G\varphi$ if and only if $M,w \models E_G(\varphi \wedge C_i\varphi)$

Notice the recursion in the definition.

The "Common Knowledge" Operator (Axiomatization)

- All we need is **S5** plus the following:
- **Axiom (A3)** $E_G\varphi \leftrightarrow (K_1\varphi \wedge \ldots \wedge K_n\varphi)$
  - given $G=\{1,\ldots,n\}$
- **Axiom (A4)** $C_G\varphi \rightarrow E_G(\varphi \wedge C_i\varphi)$
- **Rule (R3)** From $\varphi \rightarrow E_G(\psi \wedge \varphi)$

  infer $\varphi \rightarrow C_G\psi$
  - This is called the induction rule.

Back to Coordinated Attack

- Whenever any communication protocol guarantees a coordinated attack in a particular history, in that history we must have common knowledge between the two generals that an attack is about to happen.
- No finite exchange of acknowledgements will ever lead to such common knowledge.
- There is no communication protocol that solves the Coordinated Attack problem.

Reading

- *Logics for Knowledge and Belief. Chapter 13 of* ***Multiagent Systems: Algorithmic, Game-Theoretic, and Logical Foundations.*** Y. Shoham, K. Leyton-Brown. Cambridge University Press, 2009.