Modal Logic

Epistemic and Doxastic Logic

Eric Pacuit

University of Maryland, College Park

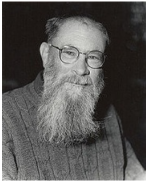
pacuit.org
epacuit@umd.edu

October 19, 2015

Literature

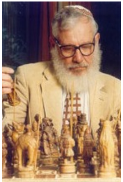
1. W. Holliday, Epistemic Logic and Epistemology, *Handbook of Formal Philosophy*, Springer, forthcoming
2. E. Pacuit, Dynamic Epistemic Logic I: Modeling Knowledge and Belief, *Philosophy Compass*, 2013
3. E. Pacuit, Dynamic Epistemic Logic II: Logics of Information Change, *Philosophy Compass*, 2013
4. R. Sorensen, Epistemic Paradoxes, Stanford Encyclopedia of Philosophy, 2011

Foundations of Epistemic Logic



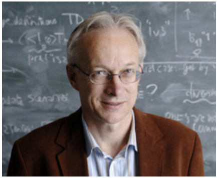
David Lewis

Jakko Hintikka



Robert Aumann

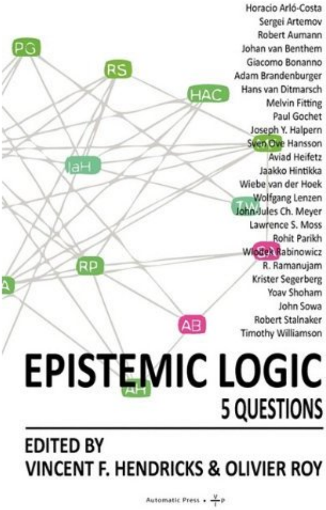
Larry Moss



Johan van Benthem

Alexandru Baltag

Foundations of Epistemic Logic



Ten Puzzles and Paradoxes

1. Surprise Exam
2. The Knower
3. Logical Omniscience/Knowledge Closure
4. Lottery Paradox & Preface Paradox
5. Margin of Error Paradox
6. Fitch's Paradox
7. Aumann's Agreeing to Disagree Theorem
8. Brandenburger-Keisler Paradox
9. Absent-Minded Driver
10. Common Knowledge of Rationality and Backwards Induction

Three introductory examples

Epistemic Logic

Let K_aP informally mean "agent a knows that P (is true)".

Epistemic Logic

Let $K_a P$ informally mean "agent a knows that P (is true)".

$K_a(P \to Q)$: "Ann knows that P implies Q"

Epistemic Logic

Let K_aP informally mean "agent a knows that P (is true)".

$K_a(P \to Q)$: "Ann knows that P implies Q"

$K_aP \vee \neg K_aP$: "either Ann does or does not know P"

Epistemic Logic

Let $K_a P$ informally mean "agent a knows that P (is true)".

$K_a(P \rightarrow Q)$: "Ann knows that P implies Q"

$K_a P \vee \neg K_a P$: "either Ann does or does not know P"

$K_a P \vee K_a \neg P$: "Ann knows whether P is true"

Epistemic Logic

Let K_aP informally mean "agent a knows that P (is true)".

$K_a(P \to Q)$: "Ann knows that P implies Q"

$K_aP \vee \neg K_aP$: "either Ann does or does not know P"

$K_aP \vee K_a\neg P$: "Ann knows whether P is true"

$\neg K_a \neg P$: "P is an epistemic possibility for Ann"

Epistemic Logic

Let $K_a P$ informally mean "agent a knows that P (is true)".

$K_a(P \to Q)$: "Ann knows that P implies Q"

$K_a P \vee \neg K_a P$: "either Ann does or does not know P"

$K_a P \vee K_a \neg P$: "Ann knows whether P is true"

$\neg K_a \neg P$: "P is an epistemic possibility for Ann"

$K_a K_a P$: "Ann knows that she knows that P"

Example

Suppose there are three cards: 1, 2 and 3.

Ann is dealt one of the cards, one of the cards is placed face down on the table and the third card is put back in the deck.

Example

Suppose there are three cards: 1, 2 and 3.

Ann is dealt one of the cards, one of the cards is placed face down on the table and the third card is put back in the deck.

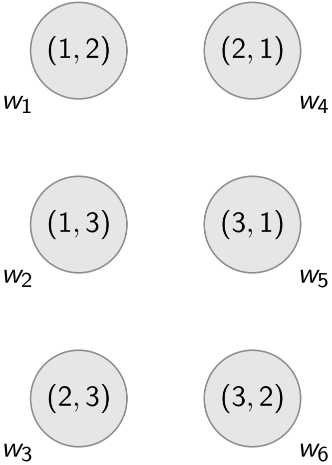
What are the relevant states?

Example

Suppose there are three cards: 1, 2 and 3.

Ann is dealt one of the cards, one of the cards is placed face down on the table and the third card is put back in the deck.

What are the relevant states?

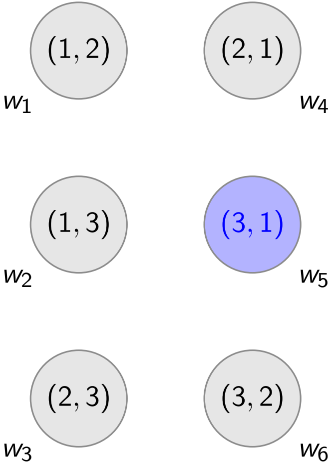


Example

Suppose there are three cards: 1, 2 and 3.

Ann is dealt one of the cards, one of the cards is placed face down on the table and the third card is put back in the deck.

Ann receives card 3 and card 1 is put on the table

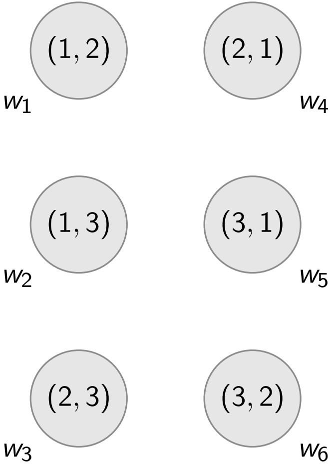


Example

Suppose there are three cards: 1, 2 and 3.

Ann is dealt one of the cards, one of the cards is placed face down on the table and the third card is put back in the deck.

What information does Ann have?

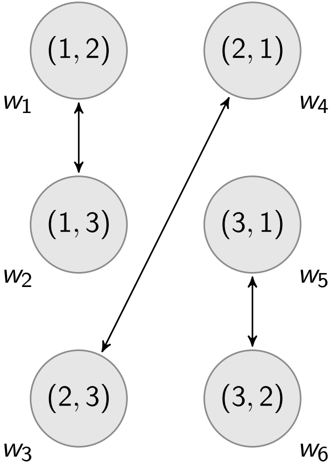


Example

Suppose there are three cards: 1, 2 and 3.

Ann is dealt one of the cards, one of the cards is placed face down on the table and the third card is put back in the deck.

What information does Ann have?

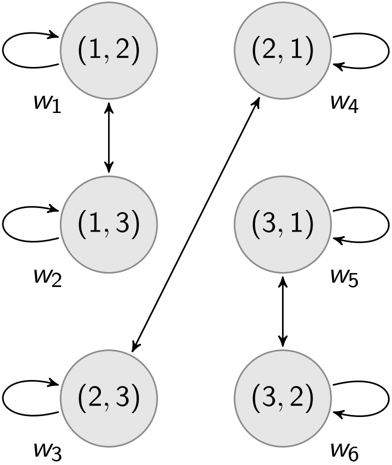


Example

Suppose there are three cards: 1, 2 and 3.

Ann is dealt one of the cards, one of the cards is placed face down on the table and the third card is put back in the deck.

What information does Ann have?



Example

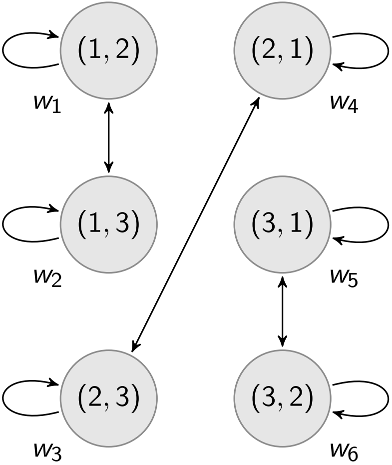
Suppose there are three cards: 1, 2 and 3.

Ann is dealt one of the cards, one of the cards is placed face down on the table and the third card is put back in the deck.

Suppose H_i is intended to mean "Ann has card i"

T_i is intended to mean "card i is on the table"

Eg., $V(H_1) = \{w_1, w_2\}$



Example

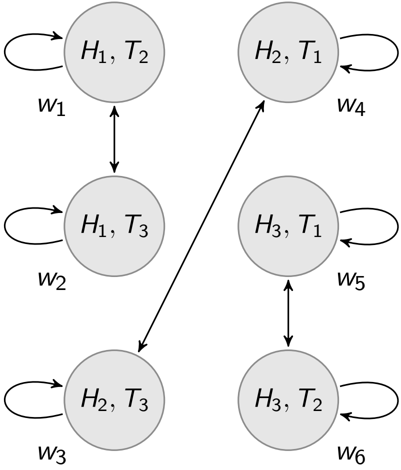
Suppose there are three cards: 1, 2 and 3.

Ann is dealt one of the cards, one of the cards is placed face down on the table and the third card is put back in the deck.

Suppose H_i is intended to mean "Ann has card i"

T_i is intended to mean "card i is on the table"

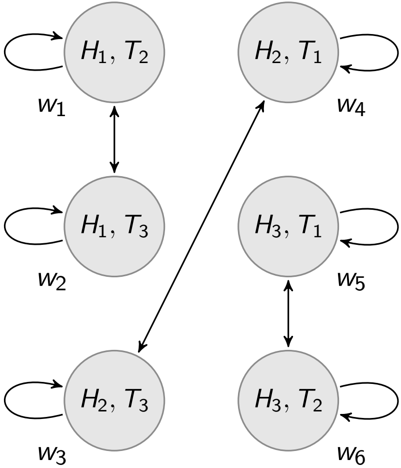
Eg., $V(H_1) = \{w_1, w_2\}$



Example

Suppose there are three cards: 1, 2 and 3.

Ann is dealt one of the cards, one of the cards is placed face down on the table and the third card is put back in the deck.

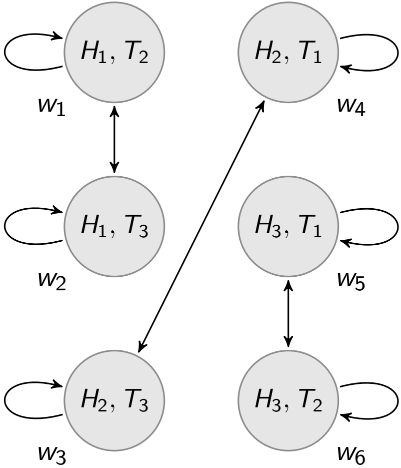


Example

Suppose there are three cards: 1, 2 and 3.

Ann is dealt one of the cards, one of the cards is placed face down on the table and the third card is put back in the deck.

Suppose that Ann receives card 1 and card 2 is on the table.

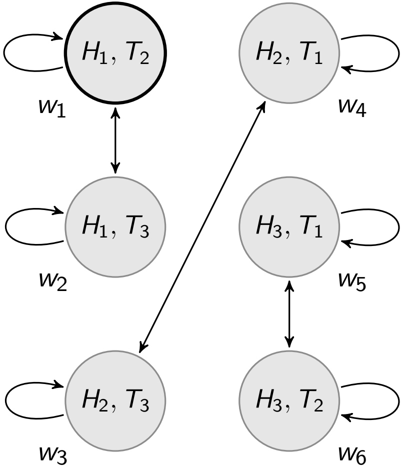


Example

Suppose there are three cards: 1, 2 and 3.

Ann is dealt one of the cards, one of the cards is placed face down on the table and the third card is put back in the deck.

Suppose that Ann receives card 1 and card 2 is on the table.

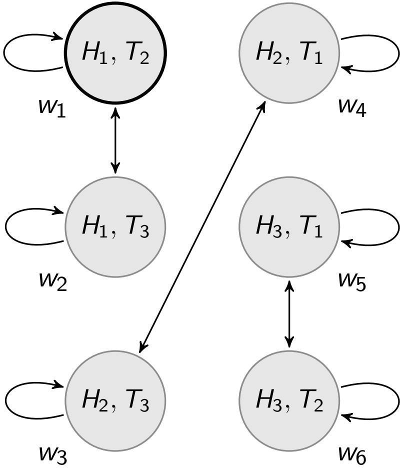


Example

Suppose there are three cards: 1, 2 and 3.

Ann is dealt one of the cards, one of the cards is placed face down on the table and the third card is put back in the deck.

$\mathcal{M}, w_1 \models K_a H_1$

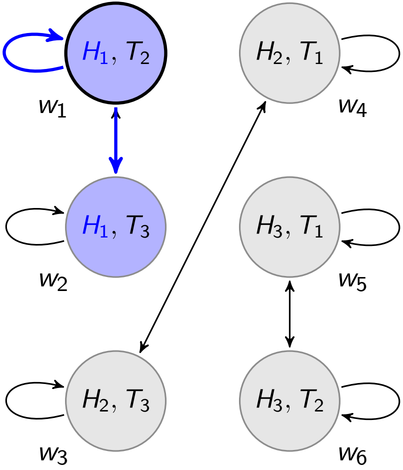


Example

Suppose there are three cards: 1, 2 and 3.

Ann is dealt one of the cards, one of the cards is placed face down on the table and the third card is put back in the deck.

$\mathcal{M}, w_1 \models K_a H_1$



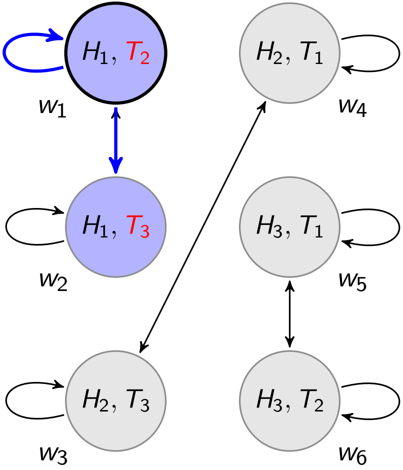
Example

Suppose there are three cards: 1, 2 and 3.

Ann is dealt one of the cards, one of the cards is placed face down on the table and the third card is put back in the deck.

$\mathcal{M}, w_1 \models K_a H_1$

$\mathcal{M}, w_1 \models K_a \neg T_1$

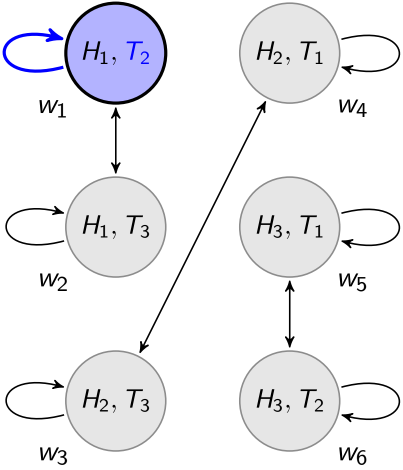


Example

Suppose there are three cards: 1, 2 and 3.

Ann is dealt one of the cards, one of the cards is placed face down on the table and the third card is put back in the deck.

$\mathcal{M}, w_1 \models \neg K_a \neg T_2$

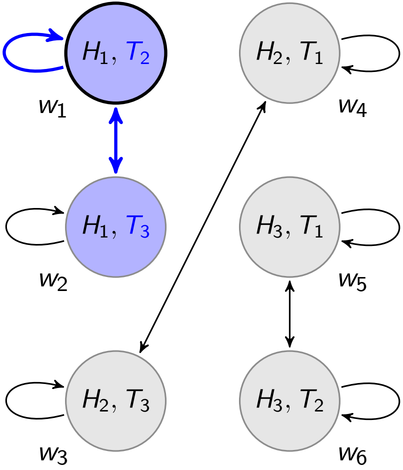


Example

Suppose there are three cards: 1, 2 and 3.

Ann is dealt one of the cards, one of the cards is placed face down on the table and the third card is put back in the deck.

$\mathcal{M}, w_1 \models K_a(T_2 \vee T_3)$



Multiagent Epistemic Logic

Many of the examples we are interested in involve more than one agent!

Multiagent Epistemic Logic

Many of the examples we are interested in involve more than one agent!

$K_a P$ means "Ann knows P"

$K_b P$ means "Bob knows P"

Multiagent Epistemic Logic

Many of the examples we are interested in involve more than one agent!

K_aP means "Ann knows P"

K_bP means "Bob knows P"

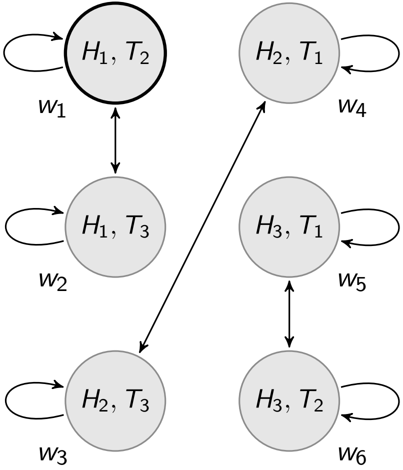
- $K_aK_b\varphi$: "Ann knows that Bob knows φ"
- $K_a(K_b\varphi \vee K_b\neg\varphi)$: "Ann knows that Bob knows whether φ
- $\neg K_bK_aK_b(\varphi)$: "Bob does not know that Ann knows that Bob knows that φ"

Example

Suppose there are three cards: 1, 2 and 3.

Ann is dealt one of the cards, one of the cards is placed face down on the table and the third card is put back in the deck.

Suppose that Ann receives card 1 and card 2 is on the table.

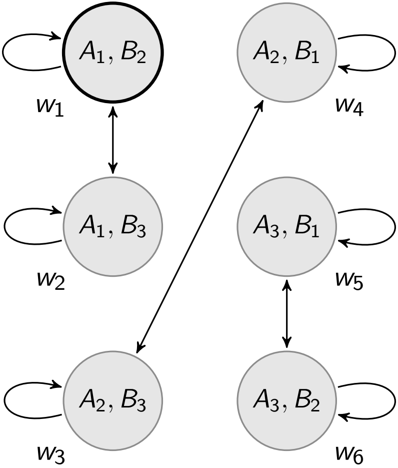


Example

Suppose there are three cards: 1, 2 and 3.

Ann is dealt one of the cards, Bob is given one of the cards and the third card is put back in the deck.

Suppose that Ann receives card 1 and Bob receives card 2.

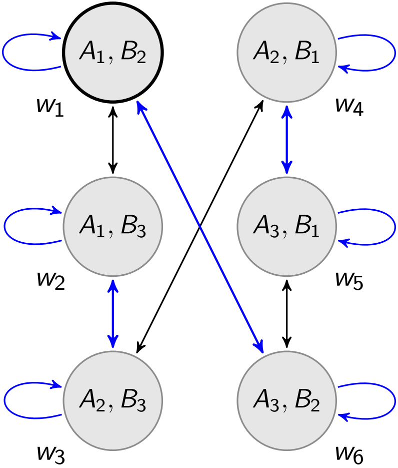


Example

Suppose there are three cards: 1, 2 and 3.

Ann is dealt one of the cards, Bob is given one of the cards and the third card is put back in the deck.

Suppose that Ann receives card 1 and Bob receives card 2.

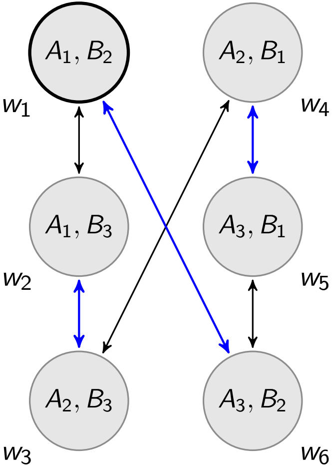


Example

Suppose there are three cards: 1, 2 and 3.

Ann is dealt one of the cards, Bob is given one of the cards and the third card is put back in the deck.

Suppose that Ann receives card 1 and Bob receives card 2.



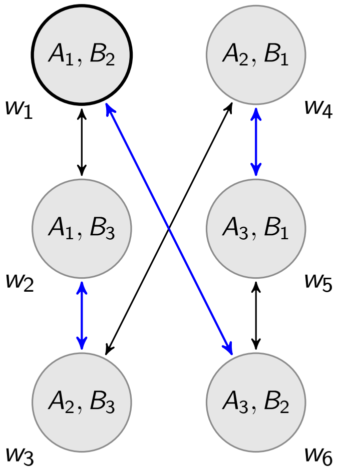
Example

Suppose there are three cards: 1, 2 and 3.

Ann is dealt one of the cards, Bob is given one of the cards and the third card is put back in the deck.

Suppose that Ann receives card 1 and Bob receives card 2.

$\mathcal{M}, w_1 \models K_b(K_a A_1 \vee K_a \neg A_1)$



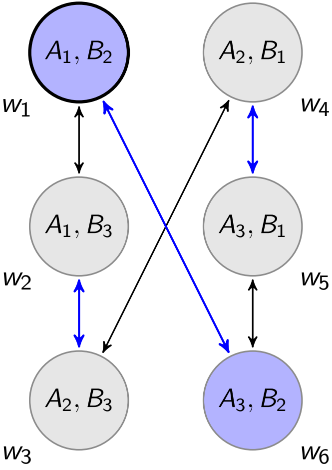
Example

Suppose there are three cards: 1, 2 and 3.

Ann is dealt one of the cards, Bob is given one of the cards and the third card is put back in the deck.

Suppose that Ann receives card 1 and Bob receives card 2.

$\mathcal{M}, w_1 \models K_b(K_a A_1 \vee K_a \neg A_1)$



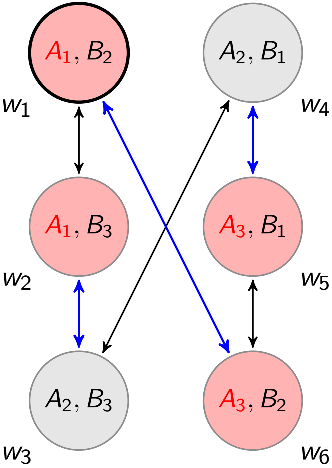
Example

Suppose there are three cards: 1, 2 and 3.

Ann is dealt one of the cards, Bob is given one of the cards and the third card is put back in the deck.

Suppose that Ann receives card 1 and Bob receives card 2.

$\mathcal{M}, w_1 \models K_b(K_a A_1 \vee K_a \neg A_1)$



College Park and Amsterdam

Let K_c stand for **agent** c **knows that** and K_a stand for **agent** a **knows that**. Suppose agent c, who lives in College Park, knows that agent a lives in Amsterdam. Let r stand for 'it's raining in Amsterdam'. Although c doesn't know whether it's raining in Amsterdam, c knows that a knows whether it's raining there:

College Park and Amsterdam

Let K_c stand for **agent** c **knows that** and K_a stand for **agent** a **knows that**. Suppose agent c, who lives in College Park, knows that agent a lives in Amsterdam. Let r stand for 'it's raining in Amsterdam'. Although c doesn't know whether it's raining in Amsterdam, c knows that a knows whether it's raining there:

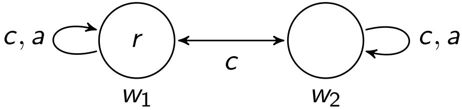
$$\neg(K_c r \vee K_c \neg r) \wedge K_c(K_a r \vee K_a \neg r).$$

College Park and Amsterdam

Let K_c stand for **agent** c **knows that** and K_a stand for **agent** a **knows that**. Suppose agent c, who lives in College Park, knows that agent a lives in Amsterdam. Let r stand for 'it's raining in Amsterdam'. Although c doesn't know whether it's raining in Amsterdam, c knows that a knows whether it's raining there:

$$\neg(K_c r \vee K_c \neg r) \wedge K_c(K_a r \vee K_a \neg r).$$

The following picture depicts a situation in which this is true, where an arrow represents *compatibility with one's knowledge*:

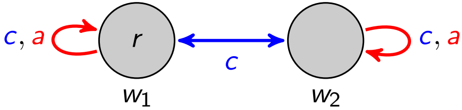


College Park and Amsterdam

Let K_c stand for **agent** c **knows that** and K_a stand for **agent** a **knows that**. Suppose agent c, who lives in College Park, knows that agent a lives in Amsterdam. Let r stand for 'it's raining in Amsterdam'. Although c doesn't know whether it's raining in Amsterdam, c knows that a knows whether it's raining there:

$$\neg(K_c r \vee K_c \neg r) \wedge K_c(K_a r \vee K_a \neg r).$$

The following picture depicts a situation in which this is true, where an arrow represents *compatibility with one's knowledge*:

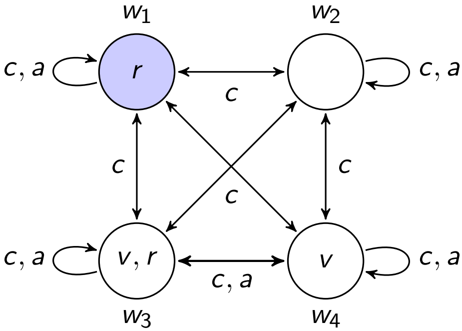


Now suppose that agent c doesn't know whether agent a has left Amsterdam for a vacation. (Let v stand for 'a has left Amsterdam on vacation'.) Agent c knows that if a is not on vacation, then a knows whether it's raining in Amsterdam; but if a is on vacation, then a won't bother to follow the weather.

$$K_c(\neg v \to (K_a r \vee K_a \neg r)) \wedge K_c(v \to \neg(K_a r \vee K_a \neg r)).$$

Now suppose that agent c doesn't know whether agent a has left Amsterdam for a vacation. (Let v stand for 'a has left Amsterdam on vacation'.) Agent c knows that if a is not on vacation, then a knows whether it's raining in Amsterdam; but if a is on vacation, then a won't bother to follow the weather.

$$K_c(\neg v \rightarrow (K_a r \vee K_a \neg r)) \wedge K_c(v \rightarrow \neg(K_a r \vee K_a \neg r)).$$



The Muddy Children Puzzle

Three children are outside playing. Two of them get mud on their forehead. They cannot see or feel the mud on their own foreheads, but can see who is dirty.

Three children are outside playing. Two of them get mud on their forehead. They cannot see or feel the mud on their own foreheads, but can see who is dirty.

Their mother enters the room and says "At least one of you have mud on your forehead".

Three children are outside playing. Two of them get mud on their forehead. They cannot see or feel the mud on their own foreheads, but can see who is dirty.

Their mother enters the room and says "At least one of you have mud on your forehead".

Then the children are repeatedly asked "do you know if you have mud on your forehead?"

Three children are outside playing. Two of them get mud on their forehead. They cannot see or feel the mud on their own foreheads, but can see who is dirty.

Their mother enters the room and says "At least one of you have mud on your forehead".

Then the children are repeatedly asked "do you know if you have mud on your forehead?"

What happens?

Three children are outside playing. Two of them get mud on their forehead. They cannot see or feel the mud on their own foreheads, but can see who is dirty.

Their mother enters the room and says "At least one of you have mud on your forehead".

Then the children are repeatedly asked "do you know if you have mud on your forehead?"

What happens?

Claim: After first question, the children answer "I don't know",

Three children are outside playing. Two of them get mud on their forehead. They cannot see or feel the mud on their own foreheads, but can see who is dirty.

Their mother enters the room and says "At least one of you have mud on your forehead".

Then the children are repeatedly asked "do you know if you have mud on your forehead?"

What happens?

Claim: After first question, the children answer "I don't know", after the second question the muddy children answer "I have mud on my forehead!" (but the clean child is still in the dark).

Three children are outside playing. Two of them get mud on their forehead. They cannot see or feel the mud on their own foreheads, but can see who is dirty.

Their mother enters the room and says "At least one of you have mud on your forehead".

Then the children are repeatedly asked "do you know if you have mud on your forehead?"

What happens?

Claim: After first question, the children answer "I don't know", after the second question the muddy children answer "I have mud on my forehead!" (but the clean child is still in the dark). Then the clean child says, "Oh, I must be clean."

Muddy Children

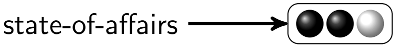
Assume:

- There are three children: Ann, Bob and Charles.
- (Only) Ann and Bob have mud on their forehead.

Muddy Children

Assume:

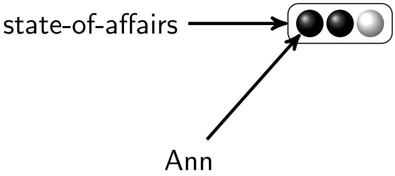
- There are three children: Ann, Bob and Charles.
- (Only) Ann and Bob have mud on their forehead.



Muddy Children

Assume:

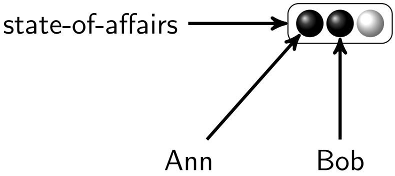
- There are three children: Ann, Bob and Charles.
- (Only) Ann and Bob have mud on their forehead.



Muddy Children

Assume:

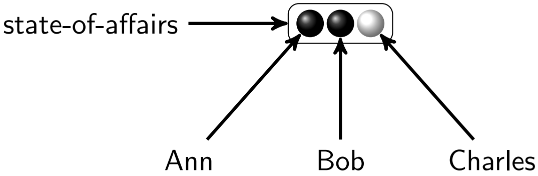
- There are three children: Ann, Bob and Charles.
- (Only) Ann and Bob have mud on their forehead.



Muddy Children

Assume:

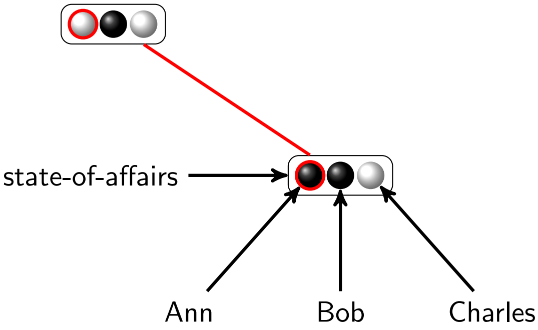
- There are three children: Ann, Bob and Charles.
- (Only) Ann and Bob have mud on their forehead.



Muddy Children

Assume:

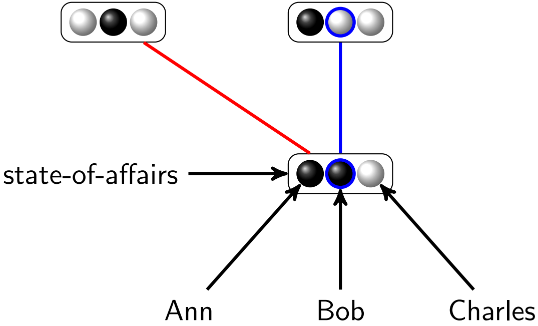
- There are three children: Ann, Bob and Charles.
- (Only) Ann and Bob have mud on their forehead.



Muddy Children

Assume:

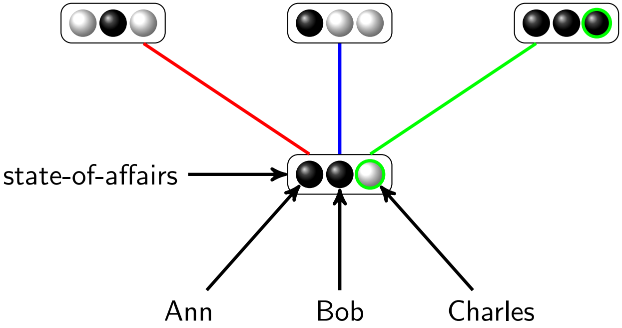
- There are three children: Ann, Bob and Charles.
- (Only) Ann and Bob have mud on their forehead.



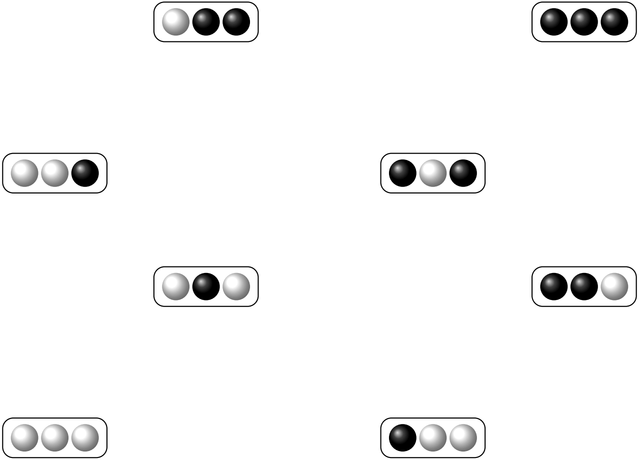
Muddy Children

Assume:

- There are three children: Ann, Bob and Charles.
- (Only) Ann and Bob have mud on their forehead.

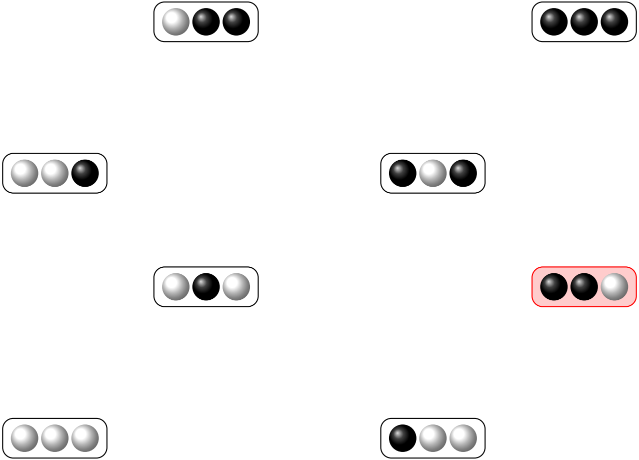


Muddy Children



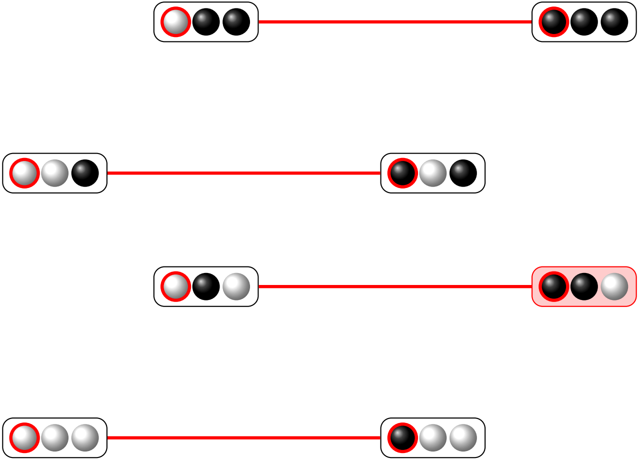
The 8 possible situations

Muddy Children



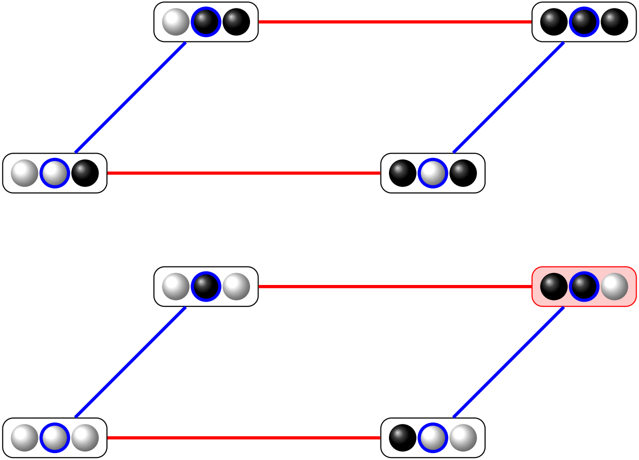
The actual situation

Muddy Children



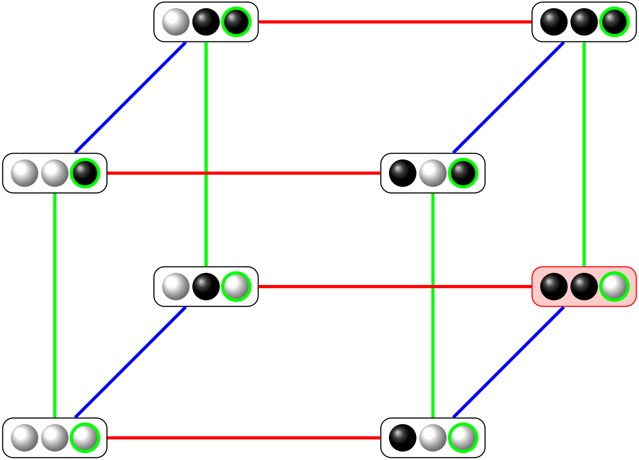
Ann's uncertainty

Muddy Children



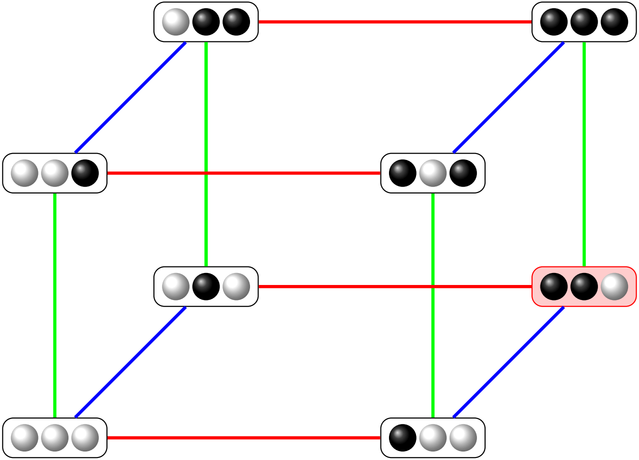
Bob's uncertainty

Muddy Children

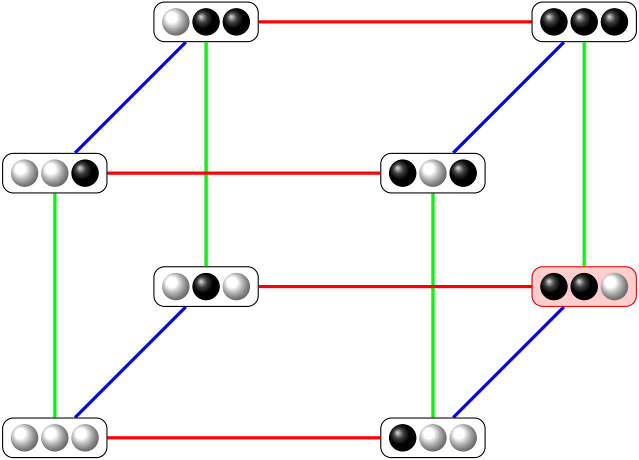


Charles' uncertainty

Muddy Children

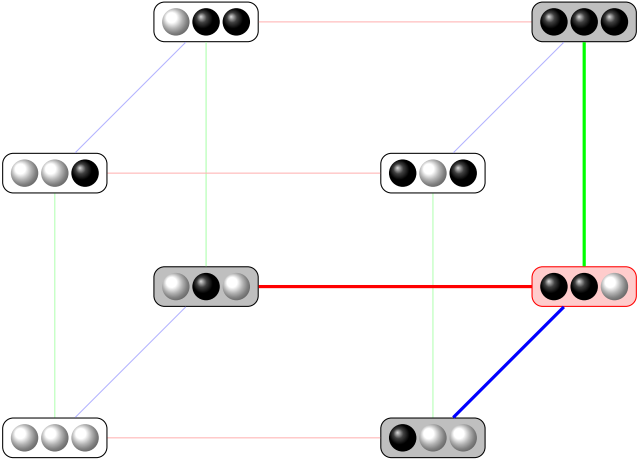


Muddy Children



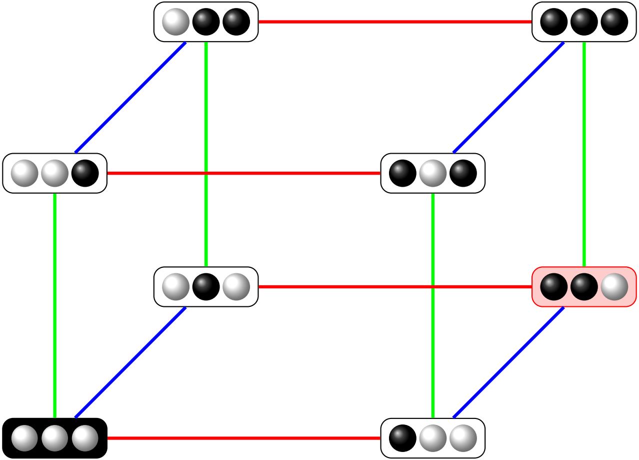
None of the children know if they are muddy

Muddy Children



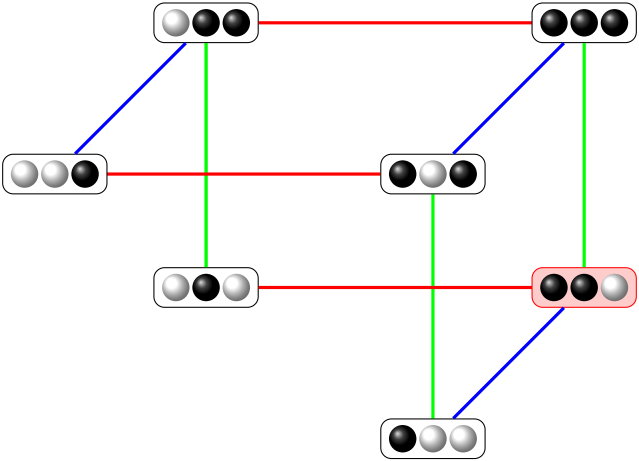
None of the children know if they are muddy

Muddy Children



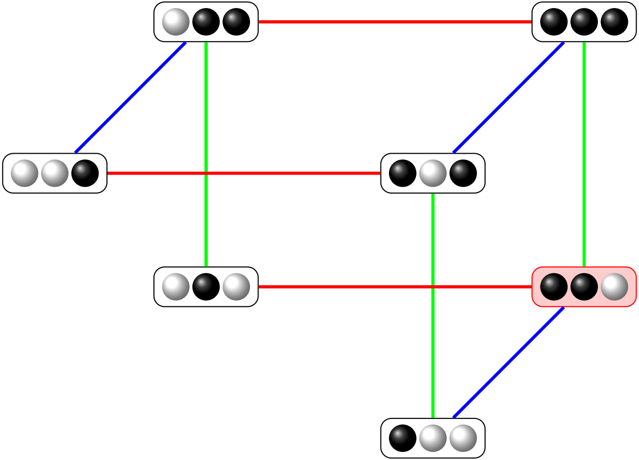
“At least one has mud on their forehead.”

Muddy Children



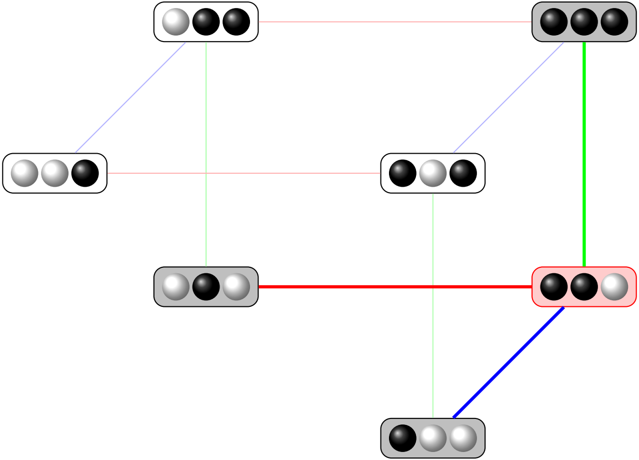
"At least one has mud on their forehead."

Muddy Children



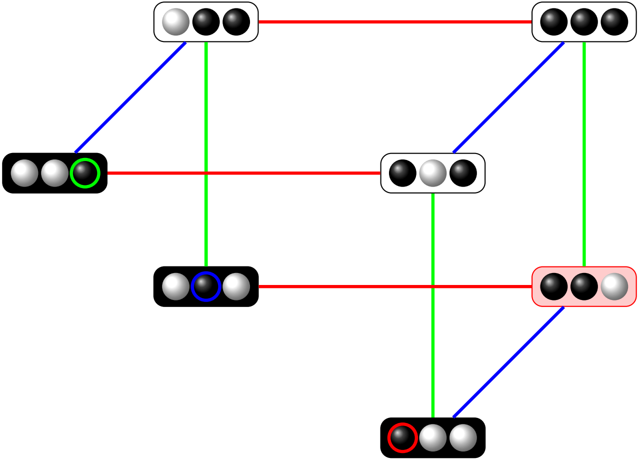
"Who has mud on their forehead?"

Muddy Children



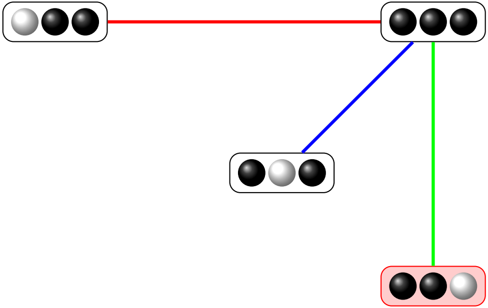
"Who has mud on their forehead?"

Muddy Children



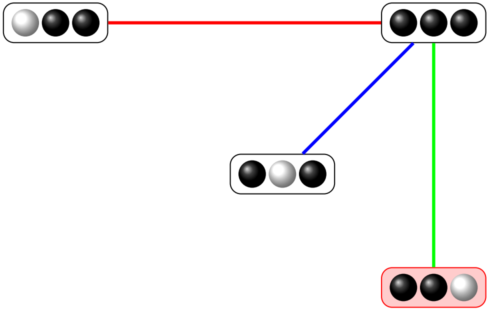
No one steps forward.

Muddy Children



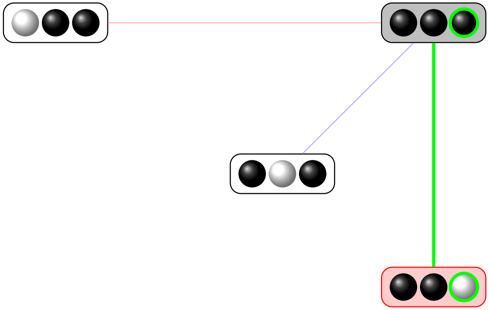
No one steps forward.

Muddy Children



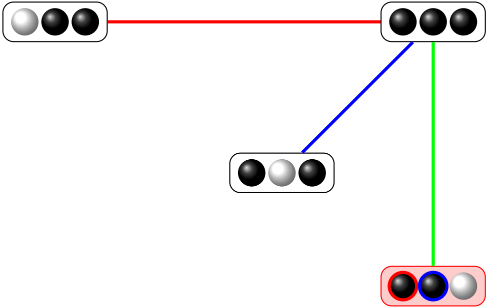
"Who has mud on their forehead?"

Muddy Children



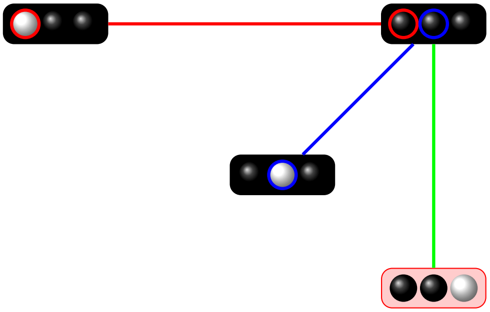
Charles does not know he is clean.

Muddy Children



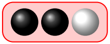
Ann and Bob step forward.

Muddy Children



Ann and Bob step forward.

Muddy Children



Now, Charles knows he is clean.

Epistemic Logic: The Language

φ is a formula of Epistemic Logic ($\mathcal{L}$) if it is of the form

$$\varphi \ := \ p \mid \neg\varphi \mid \varphi \wedge \psi \mid K_a\varphi$$

Epistemic Logic: The Language

φ is a formula of Epistemic Logic ($\mathcal{L}$) if it is of the form

$$\varphi \;:=\; p \mid \neg\varphi \mid \varphi \wedge \psi \mid K_a\varphi$$

- $p \in \mathsf{At}$ is an atomic fact.
  - "It is raining"
  - "The talk is at 2PM"
  - "The card on the table is a 7 of Hearts"

Epistemic Logic: The Language

φ is a formula of Epistemic Logic ($\mathcal{L}$) if it is of the form

$$\varphi \;:=\; p \mid \neg\varphi \mid \varphi \wedge \psi \mid K_a\varphi$$

- $p \in \mathrm{At}$ is an atomic fact.
- The usual propositional language ($\mathcal{L}_0$)

Epistemic Logic: The Language

φ is a formula of Epistemic Logic ($\mathcal{L}$) if it is of the form

$$\varphi \ := \ p \mid \neg\varphi \mid \varphi \wedge \psi \mid K_a\varphi$$

- $p \in \mathsf{At}$ is an atomic fact.
- The usual propositional language ($\mathcal{L}_0$)
- $K_a\varphi$ is intended to mean "Agent a knows that φ is true".

Epistemic Logic: The Language

φ is a formula of Epistemic Logic ($\mathcal{L}$) if it is of the form

$$\varphi \ := \ p \mid \neg\varphi \mid \varphi \wedge \psi \mid K_a\varphi$$

- $p \in \mathrm{At}$ is an atomic fact.
- The usual propositional language ($\mathcal{L}_0$)
- $K_a\varphi$ is intended to mean "Agent a knows that φ is true".
- The usual definitions for $\rightarrow, \vee, \leftrightarrow$ apply
- Define $L_a\varphi$ (or $\hat{K}_a$) as $\neg K_a \neg\varphi$

Epistemic Logic: The Language

φ is a formula of Epistemic Logic ($\mathcal{L}$) if it is of the form

$$\varphi := p \mid \neg\varphi \mid \varphi \wedge \psi \mid K_a\varphi$$

$K_a(p \rightarrow q)$: "Ann knows that p implies q"

$K_a p \vee \neg K_a p$:

$K_a p \vee K_a \neg p$:

$L_a\varphi$:

$K_a L_a \varphi$:

Epistemic Logic: The Language

φ is a formula of Epistemic Logic ($\mathcal{L}$) if it is of the form

$$\varphi \ := \ p \mid \neg\varphi \mid \varphi \wedge \psi \mid K_a\varphi$$

$K_a(p \rightarrow q)$: "Ann knows that p implies q"

$K_a p \vee \neg K_a p$: "either Ann does or does not know p"

$K_a p \vee K_a \neg p$: "Ann knows whether p is true"

$L_a\varphi$:

$K_a L_a\varphi$:

Epistemic Logic: The Language

φ is a formula of Epistemic Logic ($\mathcal{L}$) if it is of the form

$$\varphi \ := \ p \mid \neg\varphi \mid \varphi \wedge \psi \mid K_a\varphi$$

$K_a(p \rightarrow q)$: "Ann knows that p implies q"

$K_a p \vee \neg K_a p$: "either Ann does or does not know p"

$K_a p \vee K_a \neg p$: "Ann knows whether p is true"

$L_a\varphi$: "φ is an epistemic possibility"

$K_a L_a\varphi$: "Ann knows that she thinks φ is possible"

Epistemic Logic: Kripke Models

$$\mathcal{M} = \langle W, \{R_a\}_{a \in \mathcal{A}}, V \rangle$$

Epistemic Logic: Kripke Models

$$\mathcal{M} = \langle W, \{R_a\}_{a \in \mathcal{A}}, V \rangle$$

- $W \neq \emptyset$ is the set of all relevant situations (states of affairs, possible worlds)

Epistemic Logic: Kripke Models

$$\mathcal{M} = \langle W, \{R_a\}_{a \in \mathcal{A}}, V \rangle$$

- $W \neq \emptyset$ is the set of all relevant situations (states of affairs, possible worlds)
- $R_a \subseteq W \times W$ *represents* the agent a's knowledge

Epistemic Logic: Kripke Models

$$\mathcal{M} = \langle W, \{R_a\}_{a \in \mathcal{A}}, V \rangle$$

- $W \neq \emptyset$ is the set of all relevant situations (states of affairs, possible worlds)
- $R_a \subseteq W \times W$ *represents* the agent a's knowledge
- $V : \mathsf{At} \rightarrow \wp(W)$ is a valuation function assigning propositional variables to worlds

Epistemic Logic: Truth in a Model

Given $\varphi \in \mathcal{L}$, a Kripke model $\mathcal{M} = \langle W, \{R_a\}_{a \in \mathcal{A}}, V \rangle$ and $w \in W$

$\mathcal{M}, w \models \varphi$ means "in $\mathcal{M}$, if the actual state is w, then φ is true"

Epistemic Logic: Truth in a Model

Given $\varphi \in \mathcal{L}$, a Kripke model $\mathcal{M} = \langle W, \{R_a\}_{a \in \mathcal{A}}, V \rangle$ and $w \in W$

$\mathcal{M}, w \models \varphi$ is defined as follows:

- $\mathcal{M}, w \models p$ iff $w \in V(p)$ (with $p \in \mathrm{At}$)
- $\mathcal{M}, w \models \neg\varphi$ if $\mathcal{M}, w \not\models \varphi$
- $\mathcal{M}, w \models \varphi \wedge \psi$ if $\mathcal{M}, w \models \varphi$ and $\mathcal{M}, w \models \psi$
- $\mathcal{M}, w \models K_a\varphi$ if for each $v \in W$, if wR_av, then $\mathcal{M}, v \models \varphi$

Epistemic Logic: Truth in a Model

Given $\varphi \in \mathcal{L}$, a Kripke model $\mathcal{M} = \langle W, \{R_a\}_{a \in \mathcal{A}}, V \rangle$ and $w \in W$

$\mathcal{M}, w \models \varphi$ is defined as follows:

- ✓ $\mathcal{M}, w \models p$ iff $w \in V(p)$ (with $p \in \mathrm{At}$)
- $\mathcal{M}, w \models \neg\varphi$ if $\mathcal{M}, w \not\models \varphi$
- $\mathcal{M}, w \models \varphi \wedge \psi$ if $\mathcal{M}, w \models \varphi$ and $\mathcal{M}, w \models \psi$
- $\mathcal{M}, w \models K_a\varphi$ if for each $v \in W$, if wR_av, then $\mathcal{M}, v \models \varphi$

Epistemic Logic: Truth in a Model

Given $\varphi \in \mathcal{L}$, a Kripke model $\mathcal{M} = \langle W, \{R_a\}_{a\in\mathcal{A}}, V\rangle$ and $w \in W$

$\mathcal{M}, w \models \varphi$ is defined as follows:

- ✓ $\mathcal{M}, w \models p$ iff $w \in V(p)$ (with $p \in \mathrm{At}$)
- ✓ $\mathcal{M}, w \models \neg\varphi$ if $\mathcal{M}, w \not\models \varphi$
- ✓ $\mathcal{M}, w \models \varphi \wedge \psi$ if $\mathcal{M}, w \models \varphi$ and $\mathcal{M}, w \models \psi$
- $\mathcal{M}, w \models K_a\varphi$ if for each $v \in W$, if wR_av, then $\mathcal{M}, v \models \varphi$

Epistemic Logic: Truth in a Model

Given $\varphi \in \mathcal{L}$, a Kripke model $\mathcal{M} = \langle W, \{R_a\}_{a\in\mathcal{A}}, V\rangle$ and $w \in W$

$\mathcal{M}, w \models \varphi$ is defined as follows:

- ✓ $\mathcal{M}, w \models p$ iff $w \in V(p)$ (with $p \in \text{At}$)
- ✓ $\mathcal{M}, w \models \neg\varphi$ if $\mathcal{M}, w \not\models \varphi$
- ✓ $\mathcal{M}, w \models \varphi \wedge \psi$ if $\mathcal{M}, w \models \varphi$ and $\mathcal{M}, w \models \psi$
- ✓ $\mathcal{M}, w \models K_a\varphi$ if for each $v \in W$, if wR_av, then $\mathcal{M}, v \models \varphi$

Epistemic Logic: Truth in a Model

Given $\varphi \in \mathcal{L}$, a Kripke model $\mathcal{M} = \langle W, \{R_a\}_{a \in \mathcal{A}}, V \rangle$ and $w \in W$

$\mathcal{M}, w \models \varphi$ is defined as follows:

- ✓ $\mathcal{M}, w \models p$ iff $w \in V(p)$ (with $p \in \text{At}$)
- ✓ $\mathcal{M}, w \models \neg\varphi$ if $\mathcal{M}, w \not\models \varphi$
- ✓ $\mathcal{M}, w \models \varphi \wedge \psi$ if $\mathcal{M}, w \models \varphi$ and $\mathcal{M}, w \models \psi$
- ✓ $\mathcal{M}, w \models K_a\varphi$ if for each $v \in W$, if wR_av, then $\mathcal{M}, v \models \varphi$
- ✓ $\mathcal{M}, w \models L_a\varphi$ if there exists a $v \in W$ such that wR_av and $\mathcal{M}, v \models \varphi$

$K_a\varphi$: "Agent a is *informed* that φ", "Agent a *knows* that φ"

$K_a\varphi$: "Agent a is *informed* that φ", "Agent a *knows* that φ"

$\mathcal{M}, w \models K_a\varphi$ iff for all $v \in W$, if wR_av then $\mathcal{M}, v \models \varphi$

I.e., $R_a(w) = \{v \mid wR_av\} \subseteq [\![\varphi]\!]_{\mathcal{M}} = \{v \mid \mathcal{M}, v \models \varphi\}$:

$K_a\varphi$: "Agent a is *informed* that φ", "Agent a *knows* that φ"

$\mathcal{M}, w \models K_a\varphi$ iff for all $v \in W$, if wR_av then $\mathcal{M}, v \models \varphi$
I.e., $R_a(w) = \{v \mid wR_av\} \subseteq [\![\varphi]\!]_{\mathcal{M}} = \{v \mid \mathcal{M}, v \models \varphi\}$:

- wR_av if "everything a knows in state w is true in v

$K_a\varphi$: "Agent a is *informed* that φ", "Agent a *knows* that φ"

$\mathcal{M}, w \models K_a\varphi$ iff for all $v \in W$, if wR_av then $\mathcal{M}, v \models \varphi$
I.e., $R_a(w) = \{v \mid wR_av\} \subseteq [\![\varphi]\!]_{\mathcal{M}} = \{v \mid \mathcal{M}, v \models \varphi\}$:

- wR_av if "everything a knows in state w is true in v
- wR_av if "agent a has the same experiences and memories in both w and v"

$K_a\varphi$: "Agent a is *informed* that φ", "Agent a *knows* that φ"

$\mathcal{M}, w \models K_a\varphi$ iff for all $v \in W$, if wR_av then $\mathcal{M}, v \models \varphi$
I.e., $R_a(w) = \{v \mid wR_av\} \subseteq [\![\varphi]\!]_{\mathcal{M}} = \{v \mid \mathcal{M}, v \models \varphi\}$:

- wR_av if "everything a knows in state w is true in v
- wR_av if "agent a has the same experiences and memories in both w and v"
- wR_av if "agent a has cannot *rule-out* v, given her evidence and observations (at state w)"

$K_a\varphi$: "Agent a is *informed* that φ", "Agent a *knows* that φ"

$\mathcal{M}, w \models K_a\varphi$ iff for all $v \in W$, if wR_av then $\mathcal{M}, v \models \varphi$

I.e., $R_a(w) = \{v \mid wR_av\} \subseteq [\![\varphi]\!]_{\mathcal{M}} = \{v \mid \mathcal{M}, v \models \varphi\}$:

- wR_av if "everything a knows in state w is true in v
- wR_av if "agent a has the same experiences and memories in both w and v"
- wR_av if "agent a has cannot *rule-out* v, given her evidence and observations (at state w)"
- wR_av if "agent a is in the same *local state* in w and v"

$L_a\varphi$ iff there is a $v \in W$ such that $\mathcal{M}, v \models \varphi$

I.e., $R_a(w) = \{v \mid wR_av\} \cap [\![\varphi]\!]_{\mathcal{M}} = \{v \mid \mathcal{M}, v \models \varphi\} \neq \emptyset$

$L_a\varphi$ iff there is a $v \in W$ such that $\mathcal{M}, v \models \varphi$

I.e., $R_a(w) = \{v \mid wR_av\} \cap [\![\varphi]\!]_{\mathcal{M}} = \{v \mid \mathcal{M}, v \models \varphi\} \neq \emptyset$

- $L_a\varphi$: "Agent *a* thinks that φ might be true."
- $L_a\varphi$: "Agent *a* considers φ possible."

$L_a\varphi$ iff there is a $v \in W$ such that $\mathcal{M}, v \models \varphi$

I.e., $R_a(w) = \{v \mid wR_av\} \cap [\![\varphi]\!]_{\mathcal{M}} = \{v \mid \mathcal{M}, v \models \varphi\} \neq \emptyset$

- ~~$L_a\varphi$: "Agent a thinks that φ might be true."~~
- ~~$L_a\varphi$: "Agent a considers φ possible."~~
- $L_a\varphi$: "(*according to the model*), φ is consistent with what a knows ($\neg K_a \neg \varphi$)".

The Surprise Exam Paradox

A teacher announces to her student, a clever logician, that she will give him a **surprise exam** in a term of $n \geq 2$ days.

The Surprise Exam Paradox

A teacher announces to her student, a clever logician, that she will give him a **surprise exam** in a term of $n \geq 2$ days. He replies:

- you can't wait until day n to give the exam, because then I'd know on the morning of n that the exam must be that day;

The Surprise Exam Paradox

A teacher announces to her student, a clever logician, that she will give him a **surprise exam** in a term of $n \geq 2$ days. He replies:

- you can't wait until day n to give the exam, because then I'd know on the morning of n that the exam must be that day;
- you also can't wait until day $n-1$ to give the exam, because then I'd know on the morning of $n-1$ that it must be that day, having ruled out day n by the previous reasoning.

The Surprise Exam Paradox

A teacher announces to her student, a clever logician, that she will give him a **surprise exam** in a term of $n \geq 2$ days. He replies:

- you can't wait until day n to give the exam, because then I'd know on the morning of n that the exam must be that day;
- you also can't wait until day $n-1$ to give the exam, because then I'd know on the morning of $n-1$ that it must be that day, having ruled out day n by the previous reasoning.
- you also can't wait until day $n-2$ to give the exam, etc.

The Surprise Exam Paradox

A teacher announces to her student, a clever logician, that she will give him a **surprise exam** in a term of $n \geq 2$ days. He replies:

- you can't wait until day n to give the exam, because then I'd know on the morning of n that the exam must be that day;
- you also can't wait until day $n-1$ to give the exam, because then I'd know on the morning of $n-1$ that it must be that day, having ruled out day n by the previous reasoning.
- you also can't wait until day $n-2$ to give the exam, etc.

He concludes that the teacher cannot give him a surprise exam.

The Surprise Exam Paradox

A teacher announces to her student, a clever logician, that she will give him a **surprise exam** in a term of $n \geq 2$ days. He replies:

- you can't wait until day n to give the exam, because then I'd know on the morning of n that the exam must be that day;
- you also can't wait until day $n-1$ to give the exam, because then I'd know on the morning of $n-1$ that it must be that day, having ruled out day n by the previous reasoning.
- you also can't wait until day $n-2$ to give the exam, etc.

He concludes that the teacher cannot give him a surprise exam. But then he is surprised to receive an exam on, say, day $n-1$.

The Surprise Exam Paradox

A teacher announces to her student, a clever logician, that she will give him a **surprise exam** in a term of $n \geq 2$ days. He replies:

- you can't wait until day n to give the exam, because then I'd know on the morning of n that the exam must be that day;
- you also can't wait until day $n-1$ to give the exam, because then I'd know on the morning of $n-1$ that it must be that day, having ruled out day n by the previous reasoning.
- you also can't wait until day $n-2$ to give the exam, etc.

He concludes that the teacher cannot give him a surprise exam. But then he is surprised to receive an exam on, say, day $n-1$.

QUESTION: what went wrong in the student's reasoning?

We will follow in the tradition of those who have formalized the prediction paradox in static epistemic/doxastic logic:

R. Binkley. *The Surprise Examination in Modal Logic*. Journal of Philosophy, 1968.

C. Harrison. 1969.. *The Unanticipated Examination in View of Kripke's Semantics for Modal Logic*. Philosophical Logic..

J. McLelland and C. Chihara. *The Surprise Examination Paradox*. Journal of Philosophical Logic, 1975.

R. Sorensen. *Blindspots*. Oxford University Press, 1988.

Our brief discussion here is based on a more detailed analysis in:

W. Holliday. *Simplifying the Surprise Exam*. 2013 (email for manuscript).

Step 1: Choosing the Formalism (language)

To formalize the paradoxes, we use the epistemic language

$$\varphi ::= p_i \mid \neg\varphi \mid (\varphi \wedge \varphi) \mid K_i\varphi$$

where $i \in \mathbb{N}$.

Step 1: Choosing the Formalism (language)

To formalize the paradoxes, we use the epistemic language

$$\varphi ::= p_i \mid \neg\varphi \mid (\varphi \wedge \varphi) \mid K_i\varphi$$

where $i \in \mathbb{N}$. For the surprise exam paradox, we read

$K_i\varphi$ as "the student knows on the *morning* of day i that φ";

p_i as "there is an exam on the *afternoon* of day i".

Step 1: Choosing the Formalism (language)

To formalize the paradoxes, we use the epistemic language

$$\varphi ::= p_i \mid \neg\varphi \mid (\varphi \wedge \varphi) \mid K_i\varphi$$

where $i \in \mathbb{N}$. For the surprise exam paradox, we read

$K_i\varphi$ as "the student knows on the *morning* of day i that φ";

p_i as "there is an exam on the *afternoon* of day i".

For the designated student paradox, we read

$K_i\varphi$ as "the i-th student in line knows that φ";

p_i as "there is a gold star on the back of the i-th student".

Step 1: Choosing the Formalism (reasoning system)

To formalize the *reasoning* in the paradoxes, we will use the minimal "normal" modal proof system **K**, extending propositional logic with the following rule for each $i \in \mathbb{N}$ (Chellas 1980, §4.1):

$$\mathsf{RK}_i \ \frac{(\varphi_1 \wedge \cdots \wedge \varphi_m) \rightarrow \psi}{(K_i\varphi_1 \wedge \cdots \wedge K_i\varphi_m) \rightarrow K_i\psi},$$

which states that if the premise is a theorem, so is the conclusion.

Step 1: Choosing the Formalism (reasoning system)

To formalize the *reasoning* in the paradoxes, we will use the minimal "normal" modal proof system **K**, extending propositional logic with the following rule for each $i \in \mathbb{N}$ (Chellas 1980, §4.1):

$$\mathsf{RK}_i \ \frac{(\varphi_1 \wedge \cdots \wedge \varphi_m) \rightarrow \psi}{(K_i\varphi_1 \wedge \cdots \wedge K_i\varphi_m) \rightarrow K_i\psi},$$

which states that if the premise is a theorem, so is the conclusion.

Intuitively, RK_i says that the student on day i (or the i-th student) knows all the logical consequences of what he knows.

Step 1: Choosing the Formalism (reasoning system)

To formalize the *reasoning* in the paradoxes, we will use the minimal "normal" modal proof system **K**, extending propositional logic with the following rule for each $i \in \mathbb{N}$ (Chellas 1980, §4.1):

$$\mathsf{RK}_i \ \frac{(\varphi_1 \wedge \cdots \wedge \varphi_m) \to \psi}{(K_i\varphi_1 \wedge \cdots \wedge K_i\varphi_m) \to K_i\psi},$$

which states that if the premise is a theorem, so is the conclusion.

Intuitively, RK_i says that the student on day i (or the i-th student) knows all the logical consequences of what he knows.

This "logical omniscience" assumption is obviously false for real, finite agents, but it is standardly assumed for the students in the surprise exam and designated student paradoxes. In any case, let us wait and see if this idealization distorts our analysis.

Step 1: Choosing the Formalism (reasoning system)

To formalize the *reasoning* involved in the paradox, we will use a simple modal proof system, extending propositional logic with the following rule for each $i \in \mathbb{N}$ (Chellas 1980, §4.1):

$$\mathsf{RK}_i \ \frac{(\varphi_1 \wedge \cdots \wedge \varphi_m) \rightarrow \psi}{(K_i\varphi_1 \wedge \cdots \wedge K_i\varphi_m) \rightarrow K_i\psi},$$

which states that if the premise is a theorem, so is the conclusion.

Intuitively, RK_i says that the student on day i (or the i-th student) knows all the logical consequences of what she knows.

In the $m = 0$ case, RK_i is the standard rule of Necessitation (Nec_i), i.e., if ψ is a theorem, then $K_i\psi$ is a theorem, so the student on day i (or the i-th student) knows all the theorems.

Step 1: Choosing the Formalism (reasoning system)

To formalize the *reasoning* involved in the paradox, we will use a simple modal proof system, extending propositional logic with the following rule for each $i \in \mathbb{N}$ (Chellas 1980, §4.1):

$$\mathsf{RK}_i \ \frac{(\varphi_1 \wedge \cdots \wedge \varphi_m) \to \psi}{(K_i\varphi_1 \wedge \cdots \wedge K_i\varphi_m) \to K_i\psi},$$

which states that if the premise is a theorem, so is the conclusion.

Intuitively, RK_i says that the student on day i (or the i-th student) knows all the logical consequences of what she knows.

Later we will consider extensions of **K** with axiom schemas such as **T**: $K\varphi \to \varphi$. Given schemas $\Sigma_1, \ldots, \Sigma_n$, $\mathbf{K}\Sigma_1 \ldots \Sigma_n$ is the least extension of **K** that includes all instances of $\Sigma_1, \ldots, \Sigma_n$.

Step 1: Choosing the Formalism (reasoning system)

A formula β is *provable* in $\mathbf{K}\Sigma_1 \ldots \Sigma_n$ from a set of formulas Γ, written $\Gamma \vdash_{\mathbf{K}\Sigma_1 \ldots \Sigma_n} \beta$, iff there is a sequence $\langle \chi_1, \ldots, \chi_l \rangle$ of formulas with $\beta = \chi_l$ such that for all $1 \leq k \leq l$, either:

Step 1: Choosing the Formalism (reasoning system)

A formula β is *provable* in $\mathbf{K}\Sigma_1 \ldots \Sigma_n$ from a set of formulas Γ, written $\Gamma \vdash_{\mathbf{K}\Sigma_1 \ldots \Sigma_n} \beta$, iff there is a sequence $\langle \chi_1, \ldots, \chi_l \rangle$ of formulas with $\beta = \chi_l$ such that for all $1 \leq k \leq l$, either:

(i) χ_k is an instance of a propositional tautology;

Step 1: Choosing the Formalism (reasoning system)

A formula β is *provable* in $\mathbf{K}\Sigma_1 \ldots \Sigma_n$ from a set of formulas Γ, written $\Gamma \vdash_{\mathbf{K}\Sigma_1\ldots\Sigma_n} \beta$, iff there is a sequence $\langle \chi_1, \ldots, \chi_l \rangle$ of formulas with $\beta = \chi_l$ such that for all $1 \leq k \leq l$, either:

(i) χ_k is an instance of a propositional tautology;

(ii) χ_k is an instance of one of the axiom schemas $\Sigma_1, \ldots, \Sigma_n$;

Step 1: Choosing the Formalism (reasoning system)

A formula β is *provable* in $\mathbf{K}\Sigma_1 \ldots \Sigma_n$ from a set of formulas Γ, written $\Gamma \vdash_{\mathbf{K}\Sigma_1 \ldots \Sigma_n} \beta$, iff there is a sequence $\langle \chi_1, \ldots, \chi_l \rangle$ of formulas with $\beta = \chi_l$ such that for all $1 \leq k \leq l$, either:

(i) χ_k is an instance of a propositional tautology;

(ii) χ_k is an instance of one of the axiom schemas $\Sigma_1, \ldots, \Sigma_n$;

(iii) $\chi_k \in \Gamma$;

Step 1: Choosing the Formalism (reasoning system)

A formula β is *provable* in $\mathbf{K}\Sigma_1 \dots \Sigma_n$ from a set of formulas Γ, written $\Gamma \vdash_{\mathbf{K}\Sigma_1\dots\Sigma_n} \beta$, iff there is a sequence $\langle \chi_1, \dots, \chi_l \rangle$ of formulas with $\beta = \chi_l$ such that for all $1 \leq k \leq l$, either:

(i) χ_k is an instance of a propositional tautology;

(ii) χ_k is an instance of one of the axiom schemas $\Sigma_1, \dots, \Sigma_n$;

(iii) $\chi_k \in \Gamma$;

(iv) (RK) χ_k is $(K_i\varphi_1 \wedge \cdots \wedge K_i\varphi_m) \rightarrow K_i\psi$ for some $i \in \mathbb{N}$, and for some $j < k$, χ_j is $(\varphi_1 \wedge \cdots \wedge \varphi_m) \rightarrow \psi$ and $\vdash_{\mathbf{K}\Sigma_1\dots\Sigma_n} \chi_j$;

Step 1: Choosing the Formalism (reasoning system)

A formula β is *provable* in $\mathbf{K}\Sigma_1 \dots \Sigma_n$ from a set of formulas Γ, written $\Gamma \vdash_{\mathbf{K}\Sigma_1 \dots \Sigma_n} \beta$, iff there is a sequence $\langle \chi_1, \dots, \chi_l \rangle$ of formulas with $\beta = \chi_l$ such that for all $1 \leq k \leq l$, either:

(i) χ_k is an instance of a propositional tautology;

(ii) χ_k is an instance of one of the axiom schemas $\Sigma_1, \dots, \Sigma_n$;

(iii) $\chi_k \in \Gamma$;

(iv) (RK) χ_k is $(K_i\varphi_1 \wedge \cdots \wedge K_i\varphi_m) \rightarrow K_i\psi$ for some $i \in \mathbb{N}$, and for some $j < k$, χ_j is $(\varphi_1 \wedge \cdots \wedge \varphi_m) \rightarrow \psi$ and $\vdash_{\mathbf{K}\Sigma_1 \dots \Sigma_n} \chi_j$;

(v) (Modus Ponens) there are $i, j < k$ such that χ_i is $\chi_j \rightarrow \chi_k$.

Step 1: Choosing the Formalism (reasoning system)

A formula β is *provable* in $\mathbf{K}\Sigma_1 \ldots \Sigma_n$ from a set of formulas Γ, written $\Gamma \vdash_{\mathbf{K}\Sigma_1 \ldots \Sigma_n} \beta$, iff there is a sequence $\langle \chi_1, \ldots, \chi_l \rangle$ of formulas with $\beta = \chi_l$ such that for all $1 \leq k \leq l$, either:

(i) χ_k is an instance of a propositional tautology;

(ii) χ_k is an instance of one of the axiom schemas $\Sigma_1, \ldots, \Sigma_n$;

(iii) $\chi_k \in \Gamma$;

(iv) (RK) χ_k is $(K_i\varphi_1 \wedge \cdots \wedge K_i\varphi_m) \rightarrow K_i\psi$ for some $i \in \mathbb{N}$, and for some $j < k$, χ_j is $(\varphi_1 \wedge \cdots \wedge \varphi_m) \rightarrow \psi$ and $\vdash_{\mathbf{K}\Sigma_1 \ldots \Sigma_n} \chi_j$;

(v) (Modus Ponens) there are $i, j < k$ such that χ_i is $\chi_j \rightarrow \chi_k$.

If there is no such proof, we write $\Gamma \nvdash_{\mathbf{K}\Sigma_1 \ldots \Sigma_n} \beta$. As usual, β is a *theorem* of $\mathbf{K}\Sigma_1 \ldots \Sigma_n$ iff β is provable from $\emptyset$, i.e., $\vdash_{\mathbf{K}\Sigma_1 \ldots \Sigma_n} \beta$.

Step 1: Choosing the Formalism (reasoning system)

A formula β is *provable* in $\mathbf{K}\Sigma_1 \ldots \Sigma_n$ from a set of formulas Γ, written $\Gamma \vdash_{\mathbf{K}\Sigma_1 \ldots \Sigma_n} \beta$, iff there is a sequence $\langle \chi_1, \ldots, \chi_l \rangle$ of formulas with $\beta = \chi_l$ such that for all $1 \leq k \leq l$, either:

(i) χ_k is an instance of a propositional tautology;

(ii) χ_k is an instance of one of the axiom schemas $\Sigma_1, \ldots, \Sigma_n$;

(iii) $\chi_k \in \Gamma$;

(iv) (RK) χ_k is $(K_i\varphi_1 \wedge \cdots \wedge K_i\varphi_m) \rightarrow K_i\psi$ for some $i \in \mathbb{N}$, and for some $j < k$, χ_j is $(\varphi_1 \wedge \cdots \wedge \varphi_m) \rightarrow \psi$ and $\vdash_{\mathbf{K}\Sigma_1 \ldots \Sigma_n} \chi_j$;

(v) (Modus Ponens) there are $i, j < k$ such that χ_i is $\chi_j \rightarrow \chi_k$.

It is important to observe the requirement in (iv) that the formula χ_j to which the RK_i rule is applied must be a theorem of the logic.

Step 2: Formalizing the Assumptions ($n = 2$)

Starting with the $n = 2$ case, consider the following assumptions:

Step 2: Formalizing the Assumptions ($n = 2$)

Starting with the $n = 2$ case, consider the following assumptions:

(A) $K_1((p_1 \wedge \neg K_1 p_1) \vee (p_2 \wedge \neg K_2 p_2))$;

(B) $K_1(p_2 \rightarrow K_2 \neg p_1)$;

(C) $K_1 K_2(p_1 \vee p_2)$.

Step 2: Formalizing the Assumptions ($n = 2$)

Starting with the $n = 2$ case, consider the following assumptions:

(A) $K_1((p_1 \wedge \neg K_1 p_1) \vee (p_2 \wedge \neg K_2 p_2))$;

(B) $K_1(p_2 \rightarrow K_2 \neg p_1)$;

(C) $K_1 K_2(p_1 \vee p_2)$.

For the surprise exam, (A) states that the student knows on the morning of day 1 that the teacher's announcement is true.

Step 2: Formalizing the Assumptions ($n = 2$)

Starting with the $n = 2$ case, consider the following assumptions:

(A) $K_1((p_1 \wedge \neg K_1 p_1) \vee (p_2 \wedge \neg K_2 p_2))$;

(B) $K_1(p_2 \to K_2 \neg p_1)$;

(C) $K_1 K_2(p_1 \vee p_2)$.

For the surprise exam, (A) states that the student knows on the morning of day 1 that the teacher's announcement is true. (B) states that the student knows on the morning of day 1 that if the exam is on the afternoon of day 2, then the student will know on the morning of day 2 that it was not on day 1 (on the basis of memory).

Step 2: Formalizing the Assumptions ($n = 2$)

Starting with the $n = 2$ case, consider the following assumptions:

(A) $K_1((p_1 \wedge \neg K_1 p_1) \vee (p_2 \wedge \neg K_2 p_2))$;

(B) $K_1(p_2 \rightarrow K_2 \neg p_1)$;

(C) $K_1 K_2(p_1 \vee p_2)$.

For the surprise exam, (A) states that the student knows on the morning of day 1 that the teacher's announcement is true. (B) states that the student knows on the morning of day 1 that if the exam is on the afternoon of day 2, then the student will know on the morning of day 2 that it was not on day 1 (on the basis of memory). Finally, (C) states that the student knows on the morning of day 1 that she will know on the morning of day 2 the part of the teacher's announcement about an *exam*.

Step 2: Formalizing the Assumptions ($n = 2$)

Starting with the $n = 2$ case, consider the following assumptions:

(A) $K_1((p_1 \wedge \neg K_1 p_1) \vee (p_2 \wedge \neg K_2 p_2))$;

(B) $K_1(p_2 \rightarrow K_2 \neg p_1)$;

(C) $K_1 K_2(p_1 \vee p_2)$.

For the designated student, (A) states that student 1 knows that the teacher's announcement is true.

Step 2: Formalizing the Assumptions ($n = 2$)

Starting with the $n = 2$ case, consider the following assumptions:

(A) $K_1((p_1 \wedge \neg K_1 p_1) \vee (p_2 \wedge \neg K_2 p_2))$;

(B) $K_1(p_2 \rightarrow K_2 \neg p_1)$;

(C) $K_1 K_2(p_1 \vee p_2)$.

For the designated student, (A) states that student 1 knows that the teacher's announcement is true. (B) states that student 1 knows that if student 2 has the gold star, then student 2 knows that student 1 does not have the gold star (on the basis of seeing the silver star on student 1's back).

Step 2: Formalizing the Assumptions ($n = 2$)

Starting with the $n = 2$ case, consider the following assumptions:

(A) $K_1((p_1 \wedge \neg K_1 p_1) \vee (p_2 \wedge \neg K_2 p_2))$;

(B) $K_1(p_2 \rightarrow K_2 \neg p_1)$;

(C) $K_1 K_2(p_1 \vee p_2)$.

For the designated student, (A) states that student 1 knows that the teacher's announcement is true. (B) states that student 1 knows that if student 2 has the gold star, then student 2 knows that student 1 does not have the gold star (on the basis of seeing the silver star on student 1's back). (C) states that student 1 knows that student 2 knows that one of them has the gold star.

Step 3: Showing Inconsistency with a Proof ($n = 2$)

Let us first show: $\{(A), (B), (C)\} \vdash_{\mathbf{K}} K_1(p_1 \wedge \neg K_1 p_1)$

Step 3: Showing Inconsistency with a Proof ($n = 2$)

Let us first show: $\{(A), (B), (C)\} \vdash_{\mathbf{K}} K_1(p_1 \wedge \neg K_1 p_1)$

(A) $K_1((p_1 \wedge \neg K_1 p_1) \vee (p_2 \wedge \neg K_2 p_2))$ premise

(B) $K_1(p_2 \rightarrow K_2 \neg p_1)$ premise

(C) $K_1 K_2(p_1 \vee p_2)$ premise

Step 3: Showing Inconsistency with a Proof ($n = 2$)

Let us first show: $\{(A), (B), (C)\} \vdash_{\mathbf{K}} K_1(p_1 \wedge \neg K_1 p_1)$

(A) $K_1((p_1 \wedge \neg K_1 p_1) \vee (p_2 \wedge \neg K_2 p_2))$ premise

(B) $K_1(p_2 \rightarrow K_2 \neg p_1)$ premise

(C) $K_1 K_2(p_1 \vee p_2)$ premise

(1.1) $((p_1 \vee p_2) \wedge \neg p_1) \rightarrow p_2)$ propositional tautology

Step 3: Showing Inconsistency with a Proof ($n = 2$)

Let us first show: $\{(A), (B), (C)\} \vdash_{\mathbf{K}} K_1(p_1 \wedge \neg K_1 p_1)$

(A) $K_1((p_1 \wedge \neg K_1 p_1) \vee (p_2 \wedge \neg K_2 p_2))$ premise

(B) $K_1(p_2 \rightarrow K_2 \neg p_1)$ premise

(C) $K_1 K_2(p_1 \vee p_2)$ premise

(1.1) $((p_1 \vee p_2) \wedge \neg p_1) \rightarrow p_2)$ propositional tautology

(1.2) $(K_2(p_1 \vee p_2) \wedge K_2 \neg p_1) \rightarrow K_2 p_2$ from (1.1) by RK_2

Step 3: Showing Inconsistency with a Proof ($n = 2$)

Let us first show: $\{(A), (B), (C)\} \vdash_{\mathbf{K}} K_1(p_1 \wedge \neg K_1 p_1)$

(A) $K_1((p_1 \wedge \neg K_1 p_1) \vee (p_2 \wedge \neg K_2 p_2))$ premise

(B) $K_1(p_2 \to K_2 \neg p_1)$ premise

(C) $K_1 K_2(p_1 \vee p_2)$ premise

(1) $(K_2(p_1 \vee p_2) \wedge K_2 \neg p_1) \to K_2 p_2$ using PL and RK_2

Step 3: Showing Inconsistency with a Proof ($n = 2$)

Let us first show: $\{(A), (B), (C)\} \vdash_{\mathbf{K}} K_1(p_1 \wedge \neg K_1 p_1)$

(A) $K_1((p_1 \wedge \neg K_1 p_1) \vee (p_2 \wedge \neg K_2 p_2))$ premise

(B) $K_1(p_2 \rightarrow K_2 \neg p_1)$ premise

(C) $K_1 K_2(p_1 \vee p_2)$ premise

(1) $(K_2(p_1 \vee p_2) \wedge K_2 \neg p_1) \rightarrow K_2 p_2$ using PL and RK_2

(2) $K_1((K_2(p_1 \vee p_2) \wedge K_2 \neg p_1) \rightarrow K_2 p_2)$ from (1) by Nec_1

Step 3: Showing Inconsistency with a Proof ($n = 2$)

Let us first show: $\{(A), (B), (C)\} \vdash_{\mathbf{K}} K_1(p_1 \wedge \neg K_1 p_1)$

(A) $K_1((p_1 \wedge \neg K_1 p_1) \vee (p_2 \wedge \neg K_2 p_2))$ premise

(B) $K_1(p_2 \rightarrow K_2 \neg p_1)$ premise

(C) $K_1 K_2(p_1 \vee p_2)$ premise

(1) $(K_2(p_1 \vee p_2) \wedge K_2 \neg p_1) \rightarrow K_2 p_2$ using PL and RK_2

(2) $K_1((K_2(p_1 \vee p_2) \wedge K_2 \neg p_1) \rightarrow K_2 p_2)$ from (1) by Nec_1

(3) $K_1(K_2 \neg p_1 \rightarrow K_2 p_2)$ from (C) and (2) using PL and RK_1

Step 3: Showing Inconsistency with a Proof ($n = 2$)

Let us first show: $\{(A), (B), (C)\} \vdash_{\mathbf{K}} K_1(p_1 \wedge \neg K_1 p_1)$

(A) $K_1((p_1 \wedge \neg K_1 p_1) \vee (p_2 \wedge \neg K_2 p_2))$ premise

(B) $K_1(p_2 \rightarrow K_2 \neg p_1)$ premise

(C) $K_1 K_2(p_1 \vee p_2)$ premise

(1) $(K_2(p_1 \vee p_2) \wedge K_2 \neg p_1) \rightarrow K_2 p_2$ using PL and RK_2

(2) $K_1((K_2(p_1 \vee p_2) \wedge K_2 \neg p_1) \rightarrow K_2 p_2)$ from (1) by Nec_1

(3) $K_1(K_2 \neg p_1 \rightarrow K_2 p_2)$ from (C) and (2) using PL and RK_1

(4) $K_1 \neg(p_2 \wedge \neg K_2 p_2)$ from (B) and (3) using PL and RK_1

Step 3: Showing Inconsistency with a Proof ($n = 2$)

Let us first show: $\{(A), (B), (C)\} \vdash_{\mathbf{K}} K_1(p_1 \wedge \neg K_1 p_1)$

(A) $K_1((p_1 \wedge \neg K_1 p_1) \vee (p_2 \wedge \neg K_2 p_2))$ premise

(B) $K_1(p_2 \rightarrow K_2 \neg p_1)$ premise

(C) $K_1 K_2(p_1 \vee p_2)$ premise

(1) $(K_2(p_1 \vee p_2) \wedge K_2 \neg p_1) \rightarrow K_2 p_2$ using PL and RK_2

(2) $K_1((K_2(p_1 \vee p_2) \wedge K_2 \neg p_1) \rightarrow K_2 p_2)$ from (1) by Nec_1

(3) $K_1(K_2 \neg p_1 \rightarrow K_2 p_2)$ from (C) and (2) using PL and RK_1

(4) $K_1 \neg(p_2 \wedge \neg K_2 p_2)$ from (B) and (3) using PL and RK_1

(5) $K_1(p_1 \wedge \neg K_1 p_1)$ from (A) and (4) using PL and RK_1

Step 3: Showing Inconsistency with a Proof ($n = 2$)

Given $\{(A), (B), (C)\} \vdash_{\mathbf{K}} K_1(p_1 \wedge \neg K_1 p_1)$, although we haven't yet derived a contradiction, we have derived something paradoxical.

Step 3: Showing Inconsistency with a Proof ($n = 2$)

Given $\{(A), (B), (C)\} \vdash_{\mathbf{K}} K_1(p_1 \wedge \neg K_1 p_1)$, although we haven't yet derived a contradiction, we have derived something paradoxical.

If we just add the "factivity" axiom T_1, $K_1\varphi \to \varphi$, or the "weak factivity" axiom J_1, $K_1\neg K_1\varphi \to \neg K_1\varphi$ (e.g., reading K as belief instead of knowledge), then we can derive a contradiction:

$$\{(A), (B), (C)\} \vdash_{\mathbf{KT_1}} \bot \text{ and } \{(A), (B), (C)\} \vdash_{\mathbf{KJ_1}} \bot.$$

Step 3: Showing Inconsistency with a Proof ($n = 2$)

Given $\{(A), (B), (C)\} \vdash_{\mathbf{K}} K_1(p_1 \wedge \neg K_1 p_1)$, although we haven't yet derived a contradiction, we have derived something paradoxical.

If we just add the "factivity" axiom T_1, $K_1\varphi \to \varphi$, or the "weak factivity" axiom J_1, $K_1\neg K_1\varphi \to \neg K_1\varphi$ (e.g., reading K as belief instead of knowledge), then we can derive a contradiction:

$$\{(A), (B), (C)\} \vdash_{\mathbf{KT_1}} \bot \text{ and } \{(A), (B), (C)\} \vdash_{\mathbf{KJ_1}} \bot.$$

Thus, we must reject either (A), (B), (C), or the rule RK_i...

Normal Modal Logics

A polymodal logic extending propositional logic with a set $\{\Box_i\}_{i\in I}$ of unary sentential operators is *normal* iff (i) for all $i \in I$,

$$\mathsf{RK}_i \ \frac{(\varphi_1 \wedge \cdots \wedge \varphi_m) \to \psi}{(\Box_i\varphi_1 \wedge \cdots \wedge \Box_i\varphi_m) \to \Box_i\psi}$$

is an admissible rule and (ii) the logic is closed under uniform substitution: if φ is a theorem, so is the result of uniformly substituting formulas for the atomic sentences in φ.

The "Problem" of Logical Omniscience

The rule

$$\mathsf{RK}_i \; \frac{(\varphi_1 \wedge \cdots \wedge \varphi_m) \to \psi}{(K_i\varphi_1 \wedge \cdots \wedge K_i\varphi_m) \to K_i\psi}$$

reflects so-called (*synchronic*) *logical omniscience*: the agent knows (at time t) all the consequences of what she knows (at t).

The "Problem" of Logical Omniscience

The rule

$$\mathsf{RK}_i \quad \frac{(\varphi_1 \wedge \cdots \wedge \varphi_m) \rightarrow \psi}{(K_i\varphi_1 \wedge \cdots \wedge K_i\varphi_m) \rightarrow K_i\psi}$$

reflects so-called (*synchronic*) *logical omniscience*: the agent knows (at time t) all the consequences of what she knows (at t).

Given this, there are two ways to view K_i: as representing either the idealized (implicit, "virtual") knowledge of ordinary agents, or the ordinary knowledge of idealized agents. For discussion, see

R. Stalnaker.

1991. "The Problem of Logical Omniscience, I," *Synthese*.

2006. "On Logics of Knowledge and Belief," *Philosophical Studies*.

The "Problem" of Logical Omniscience

The rule

$$\mathsf{RK}_i \ \frac{(\varphi_1 \wedge \cdots \wedge \varphi_m) \to \psi}{(K_i\varphi_1 \wedge \cdots \wedge K_i\varphi_m) \to K_i\psi}$$

reflects so-called (*synchronic*) *logical omniscience*: the agent knows (at time t) all the consequences of what she knows (at t).

There is now a large literature on alternative frameworks for representing the knowledge of agents with bounded rationality, who do not always "put two and two together" and therefore lack the logical omniscience reflected by RK_i. See, for example:

J. Y. Halpern and R. Pucella. 2011. *Dealing with Logical Omniscience: Expressiveness and Pragmatics*. Artificial Intelligence.

Logical Omniscience

- From $\varphi \leftrightarrow \psi$ infer $K_i\varphi \leftrightarrow K_i\psi$

Logical Omniscience

- From $\varphi \leftrightarrow \psi$ infer $K_i\varphi \leftrightarrow K_i\psi$
- From $\varphi \rightarrow \psi$ infer $K_i\varphi \rightarrow K_i\psi$

Logical Omniscience

- From $\varphi \leftrightarrow \psi$ infer $K_i\varphi \leftrightarrow K_i\psi$
- From $\varphi \rightarrow \psi$ infer $K_i\varphi \rightarrow K_i\psi$
- $(K_i(\varphi \rightarrow \psi) \wedge K_i\varphi) \rightarrow K_i\psi$

Logical Omniscience

- From $\varphi \leftrightarrow \psi$ infer $K_i\varphi \leftrightarrow K_i\psi$
- From $\varphi \rightarrow \psi$ infer $K_i\varphi \rightarrow K_i\psi$
- $(K_i(\varphi \rightarrow \psi) \wedge K_i\varphi) \rightarrow K_i\psi$
- From φ infer $K_i\varphi$

Logical Omniscience

- From $\varphi \leftrightarrow \psi$ infer $K_i\varphi \leftrightarrow K_i\psi$
- From $\varphi \rightarrow \psi$ infer $K_i\varphi \rightarrow K_i\psi$
- $(K_i(\varphi \rightarrow \psi) \wedge K_i\varphi) \rightarrow K_i\psi$
- From φ infer $K_i\varphi$
- $K_i\top$

Logical Omniscience

- From $\varphi \leftrightarrow \psi$ infer $K_i\varphi \leftrightarrow K_i\psi$
- From $\varphi \rightarrow \psi$ infer $K_i\varphi \rightarrow K_i\psi$
- $(K_i(\varphi \rightarrow \psi) \wedge K_i\varphi) \rightarrow K_i\psi$
- From φ infer $K_i\varphi$
- $K_i\top$
- $(K_i\varphi \wedge K_i\psi) \rightarrow K_i(\varphi \wedge \psi)$

Dealing with Logical Omniscience

- *Syntactic approaches*: an agents knowledge is represented by a set of formulas (intuitively, the set of formulas she knows);

Dealing with Logical Omniscience

- *Syntactic approaches*: an agents knowledge is represented by a set of formulas (intuitively, the set of formulas she knows);
- *Awareness*: an agent knows φ if she is aware of φ and φ is true in all the worlds she considers possible;

Dealing with Logical Omniscience

- *Syntactic approaches*: an agents knowledge is represented by a set of formulas (intuitively, the set of formulas she knows);
- *Awareness*: an agent knows φ if she is aware of φ and φ is true in all the worlds she considers possible;
- *Algorithmic knowledge*: an agent knows φ if her knowledge algorithm returns "Yes" on a query of φ; and

Dealing with Logical Omniscience

- *Syntactic approaches*: an agents knowledge is represented by a set of formulas (intuitively, the set of formulas she knows);
- *Awareness*: an agent knows φ if she is aware of φ and φ is true in all the worlds she considers possible;
- *Algorithmic knowledge*: an agent knows φ if her knowledge algorithm returns "Yes" on a query of φ; and
- *Impossible worlds*: an agent may consider possible worlds that are logically inconsistent (for example, where p and $\neg p$ may both be true).

Dealing with Logical Omniscience

- *Syntactic approaches*: an agents knowledge is represented by a set of formulas (intuitively, the set of formulas she knows);
- *Awareness*: an agent knows φ if she is aware of φ and φ is true in all the worlds she considers possible;
- *Algorithmic knowledge*: an agent knows φ if her knowledge algorithm returns "Yes" on a query of φ; and
- *Impossible worlds*: an agent may consider possible worlds that are logically inconsistent (for example, where p and $\neg p$ may both be true).

Non-Normal Modal Logics

Dealing with Logical Omniscience

- *Syntactic approaches*: $\mathcal{M}, w \models K_i\varphi$ iff $\varphi \in \mathcal{C}_i(w)$

Dealing with Logical Omniscience

- *Syntactic approaches*: $\mathcal{M}, w \models K_i\varphi$ iff $\varphi \in \mathcal{C}_i(w)$
- *Awareness structures*: $\mathcal{M}, w \models K_i\varphi$ iff for all $v \in W$, if wR_iv then $\mathcal{M}, v \models \varphi$ and $\varphi \in \mathcal{A}_i(w)$

Dealing with Logical Omniscience

- *Syntactic approaches*: $\mathcal{M}, w \models K_i\varphi$ iff $\varphi \in \mathcal{C}_i(w)$
- *Awareness structures*: $\mathcal{M}, w \models K_i\varphi$ iff for all $v \in W$, if wR_iv then $\mathcal{M}, v \models \varphi$ and $\varphi \in \mathcal{A}_i(w)$
- *Algorithmic knowledge*: $\mathcal{M}, w \models K_i\varphi$ iff $\mathsf{A}_i(w, \varphi) = \text{Yes}$

Dealing with Logical Omniscience

- *Syntactic approaches*: $\mathcal{M}, w \models K_i\varphi$ iff $\varphi \in \mathcal{C}_i(w)$

- *Awareness structures*: $\mathcal{M}, w \models K_i\varphi$ iff for all $v \in W$, if wR_iv then $\mathcal{M}, v \models \varphi$ and $\varphi \in \mathcal{A}_i(w)$

- *Algorithmic knowledge*: $\mathcal{M}, w \models K_i\varphi$ iff $\mathsf{A}_i(w, \varphi) = \mathrm{Yes}$

- *Impossible worlds*: $\mathcal{M}, w \models K_i\varphi$ iff if $w \in N$, then for all $v \in W$, if wR_iv and $v \in N$ then $\mathcal{M}, v \models \varphi$

  $\mathcal{M}, w \models K_i\varphi$ iff if $w \notin N$, then $\varphi \in \mathcal{C}_i(w)$

Justification Logic (1)

$t : \varphi$: "t is a *justification/proof* for φ"

S. Artemov and M. Fitting. *Justification logic*. The Stanford Encyclopedia of Philosophy, 2012.

S. Artemov. *Explicit provability and constructive semantics*. The Bulletin of Symbolic Logic 7 (2001) 1 36.

M. Fitting. *The logic of proofs, semantically*. Annals of Pure and Applied Logic 132 (2005) 1 25.

Justification Logic (2)

$$t \ := c \mid x \mid t + s \mid\ !t \mid t \cdot s$$

$$\varphi := p \mid \varphi \wedge \psi \mid \neg\varphi \mid t : \varphi$$

Justification Logic (2)

$$t \;:= c \mid x \mid t + s \mid\; !t \mid t \cdot s$$

$$\varphi := p \mid \varphi \wedge \psi \mid \neg\varphi \mid t : \varphi$$

Justification Logic:

- $t : \varphi \to \varphi$
- $t : (\varphi \to \psi) \to (s : \varphi \to t \cdot s : \psi)$
- $t : \varphi \to (t + s) : \varphi$
- $t : \varphi \to (s + t) : \varphi$
- $t : \varphi \to !t : t : \varphi$

Justification Logic (2)

$$t \;:= c \mid x \mid t+s \mid\; !t \mid t \cdot s$$

$$\varphi := p \mid \varphi \wedge \psi \mid \neg\varphi \mid t : \varphi$$

Justification Logic:

- $t : \varphi \rightarrow \varphi$
- $t : (\varphi \rightarrow \psi) \rightarrow (s : \varphi \rightarrow t \cdot s : \psi)$
- $t : \varphi \rightarrow (t+s) : \varphi$
- $t : \varphi \rightarrow (s+t) : \varphi$
- $t : \varphi \rightarrow !t : t : \varphi$

Internalization: if $\vdash_{JL} \varphi$ then there is a proof polynomial t such that $\vdash_{JL} t : \varphi$

Realization Theorem: if $\vdash_{\mathbf{S4}} \varphi$ then there is a proof polynomial t such that $\vdash_{JL} t : \varphi$

Justification Logic (3)

Fitting Semantics: $\mathcal{M} = \langle W, R, \mathcal{E}, V \rangle$

- $W \neq \emptyset$
- $R \subseteq W \times W$
- $\mathcal{E} : W \times \text{ProofTerms} \to \wp(\mathcal{L}_{JL})$
- $V : \text{At} \to \wp(W)$

$\mathcal{M}, w \models t : \varphi$ iff for all v, if wRv then $\mathcal{M}, v \models \varphi$ and $\varphi \in \mathcal{E}(w, t)$

Justification Logic (3)

Monotonicity For all $w, v \in W$, if wRv then for all proof polynomials t, $\mathcal{E}(w,t) \subseteq \mathcal{E}(v,t)$.

Application For all proof polynomials s, t and for each $w \in W$, if $\varphi \to \psi \in \mathcal{E}(w,t)$ and $\varphi \in \mathcal{E}(w,s)$, then $\psi \in \mathcal{E}(w, t \cdot s)$

Proof Checker For all proof polynomials t and for each $w \in W$, if $\varphi \in \mathcal{E}(w,t)$, then $t : \varphi \in \mathcal{E}(w, !t)$.

Sum For all proof polynomials s, t and for each $w \in W$, $\mathcal{E}(w,s) \cup \mathcal{E}(w,t) \subseteq \mathcal{E}(w, s+t)$.

Approaches

- Lack of awareness
- Lack of computational power
- Imperfect understanding of the model