Introduction to Logics of Knowledge and Belief

Eric Pacuit

University of Maryland

pacuit.org

epacuit@umd.edu

April 15, 2019

Epistemic Logic

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

Epistemic Logic

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

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

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"

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"

Epistemic Logic

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

$$
\begin{aligned}
K_a(P \to Q)&: \text{"Ann knows that } P \text{ implies } Q\text{"}\\
K_a P \vee \neg K_a P&: \text{"either Ann does or does not know } P\text{"}\\
K_a P \vee K_a \neg P&: \text{"Ann knows whether } P \text{ is true"}\\
\neg K_a \neg P&: \text{"} P \text{ is an epistemic possibility for Ann"}
\end{aligned}
$$

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"

$\neg K_a \neg P$: "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.

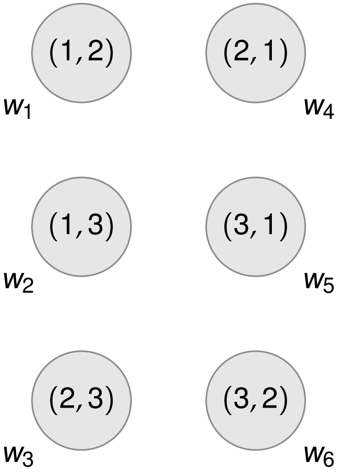
What are the relevant states?

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?

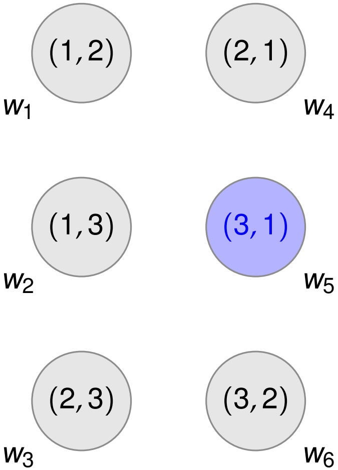


Example

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

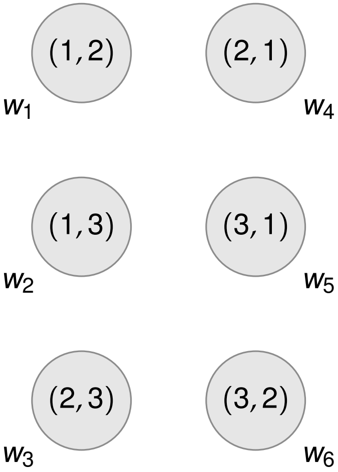


Example

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?

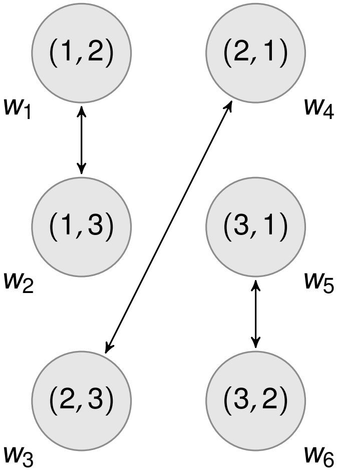


Example

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?

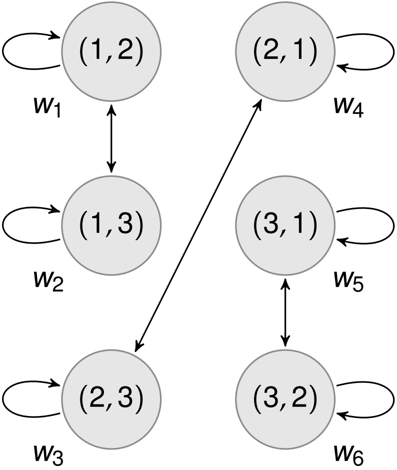


Example

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

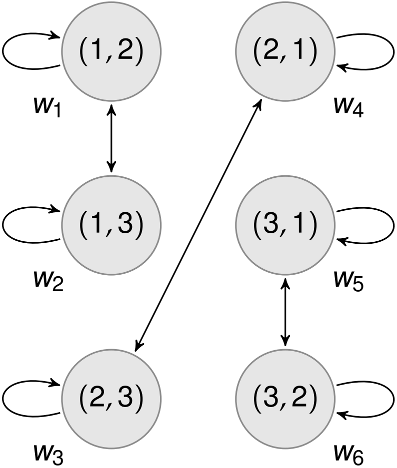
Suppose there are three cards: 1, 2 and 3.

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

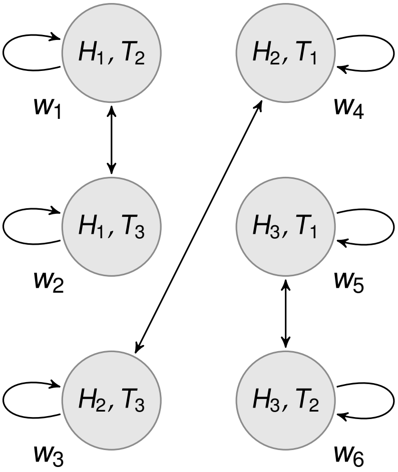
Suppose there are three cards: 1, 2 and 3.

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"

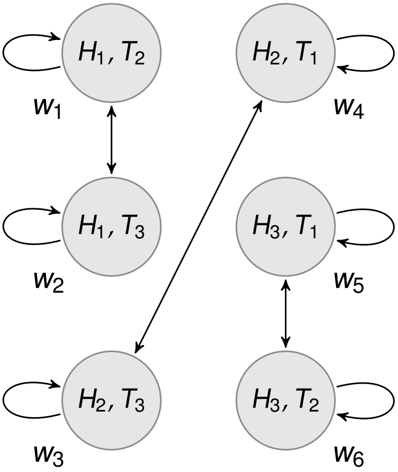
Eg., $V(H_1) = \{w_1, w_2\}$



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.

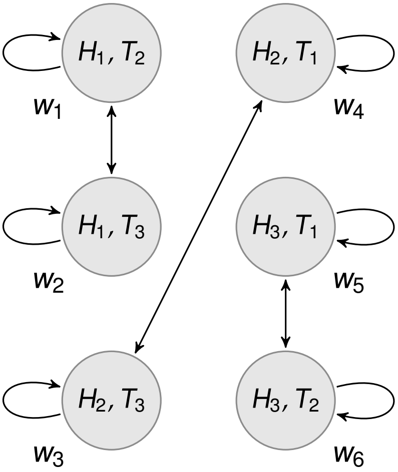


Example

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.

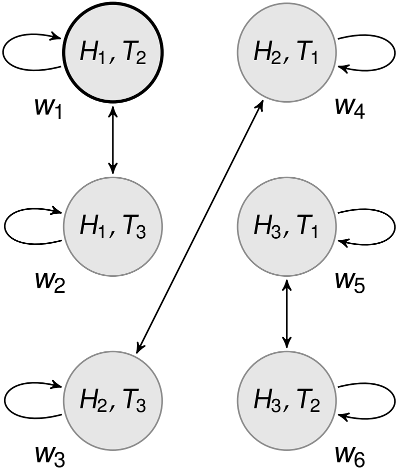


Example

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.

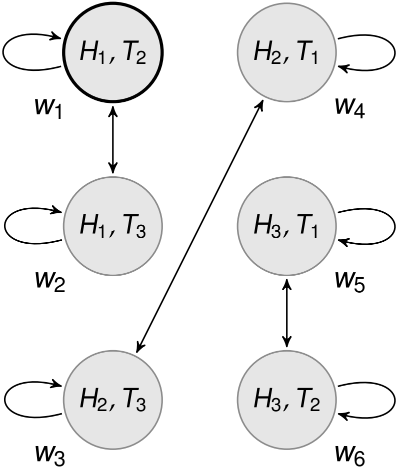


Example

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.

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

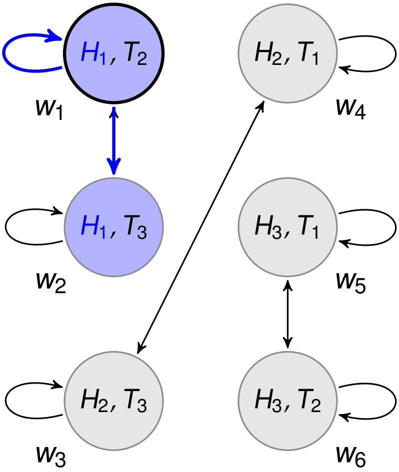


Example

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.

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



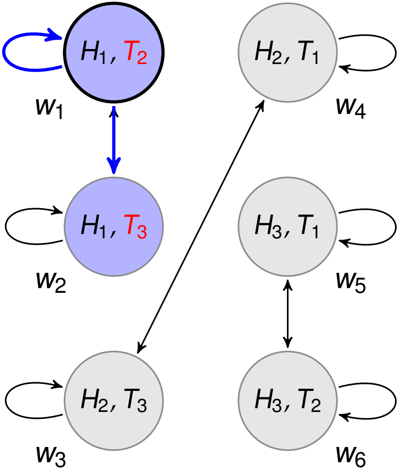
Example

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.

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

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

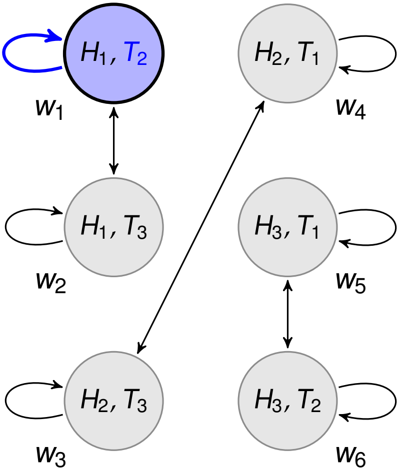


Example

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.

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

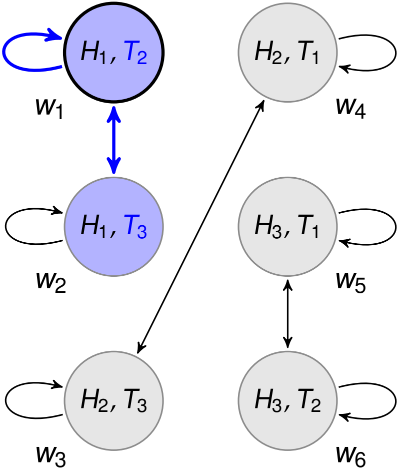


Example

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!

$K_a P$ means "Ann knows P"

$K_b P$ means "Bob knows P"

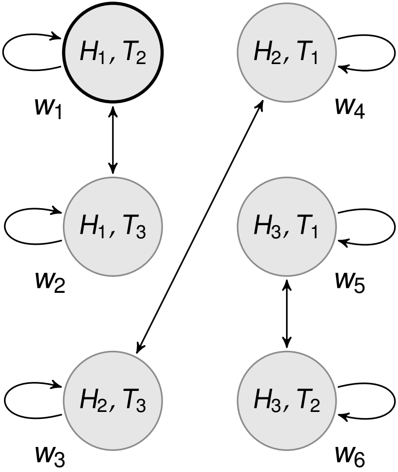
- $K_a K_b \varphi$: "Ann knows that Bob knows φ"
- $K_a(K_b\varphi \vee K_b\neg\varphi)$: "Ann knows that Bob knows whether φ
- $\neg K_b K_a K_b(\varphi)$: "Bob does not know that Ann knows that Bob knows that φ"

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.

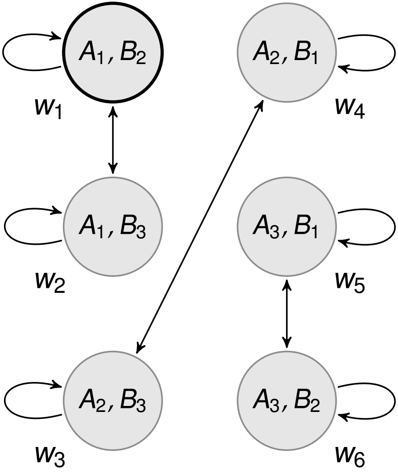


Example

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.

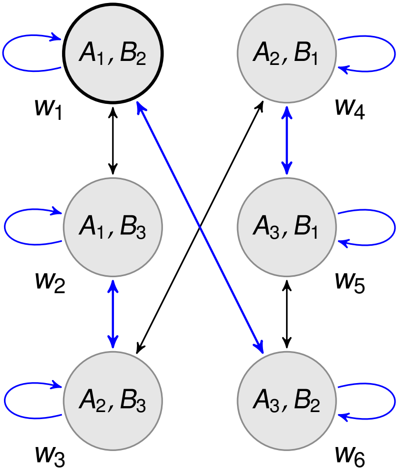


Example

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.

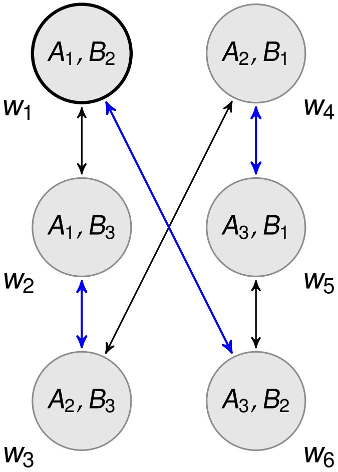


Example

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.



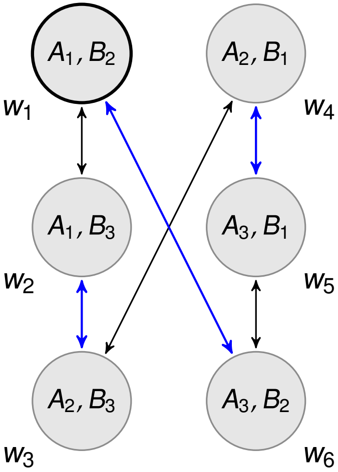
Example

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.

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



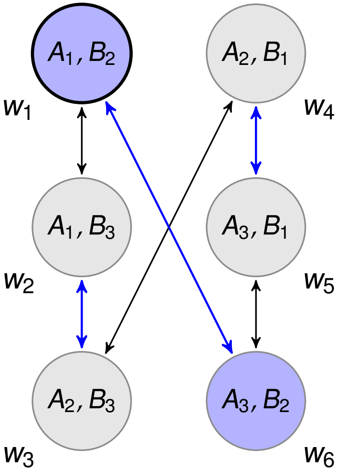
Example

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.

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



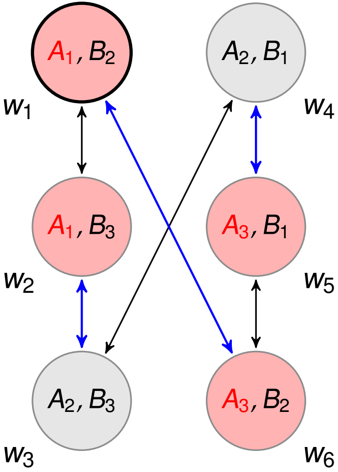
Example

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

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

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:

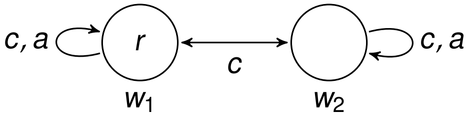
$$\neg(K_c r \vee K_c \neg r) \wedge K_c(K_a r \vee K_a \neg r).$$

College Park and Amsterdam

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*:

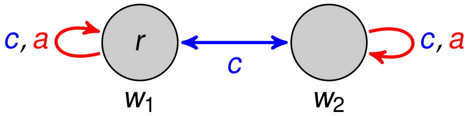


College Park and Amsterdam

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*:

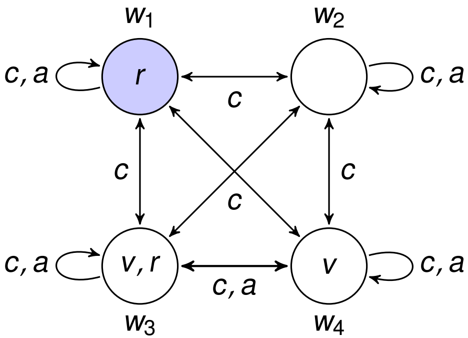


Now suppose that agent c doesn't know whether agent a has left Amsterdam for a vacation. (Let v stand for 'a has left Amsterdam on vacation'.) Agent c knows that if a is not on vacation, then a knows whether it's raining in Amsterdam; but if a is on vacation, then a won't bother to follow the weather.

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

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.

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



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

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

- $p \in \mathsf{At}$ 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

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

- $p \in \mathsf{At}$ 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".
- 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

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

$K_a(p \to q)$: "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

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

$K_a(p \to q)$: "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"

$L_a\varphi$: "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 $wR_a v$, 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 $p \in \mathsf{At}$)
- ✓ $\mathcal{M}, w \models \neg\varphi$ 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 $wR_a v$, then $\mathcal{M}, v \models \varphi$
- ✓ $\mathcal{M}, w \models L_a\varphi$ if there exists a $v \in W$ such that $wR_a v$ 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., $R_a(w) = \{v \mid wR_a v\} \cap [\![\varphi]\!]_{\mathcal{M}} = \{v \mid \mathcal{M}, v \models \varphi\} \neq \emptyset$

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

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

- $L_a\varphi$: "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 ($\neg K_a \neg\varphi$)".

Taking Stock

Multi-agent language: $\varphi \;:=\; p \mid \neg\varphi \mid \varphi \wedge \psi \mid \Box_i\varphi$

- $\Box_i\varphi$: "agent i knows that φ" (write $K_i\varphi$ for $\Box_i\varphi$)
- $\Box_i\varphi$: "agent i believes that φ" (write $B_i\varphi$ for $\Box_i\varphi$)

Kripke Models: $\mathcal{M} = \langle W, \{R_i\}_{i\in\mathcal{A}}, V\rangle$

Truth: $\mathcal{M}, w \models \Box_i\varphi$ iff for all $v \in W$, if wR_iv then $\mathcal{M}, v \models \varphi$

Modal Formula	Corresponding Property

Modal Formula	Corresponding Property
$\Box(\varphi \rightarrow \psi) \rightarrow (\Box\varphi \rightarrow \Box\psi)$	—

Modal Formula	Corresponding Property
$\Box(\varphi \to \psi) \to (\Box\varphi \to \Box\psi)$	—
$\Box\varphi \to \varphi$	Reflexive

Modal Formula	Corresponding Property
$\Box(\varphi \to \psi) \to (\Box\varphi \to \Box\psi)$	—
$\Box\varphi \to \varphi$	Reflexive
$\Box\varphi \to \Box\Box\varphi$	Transitive

Modal Formula	Corresponding Property
$\Box(\varphi \rightarrow \psi) \rightarrow (\Box\varphi \rightarrow \Box\psi)$	—
$\Box\varphi \rightarrow \varphi$	Reflexive
$\Box\varphi \rightarrow \Box\Box\varphi$	Transitive
$\neg\Box\varphi \rightarrow \Box\neg\Box\varphi$	Euclidean

Modal Formula	Corresponding Property
$\Box(\varphi \rightarrow \psi) \rightarrow (\Box\varphi \rightarrow \Box\psi)$	—
$\Box\varphi \rightarrow \varphi$	Reflexive
$\Box\varphi \rightarrow \Box\Box\varphi$	Transitive
$\neg\Box\varphi \rightarrow \Box\neg\Box\varphi$	Euclidean
$\neg\Box\bot$	Serial

The Logic **S5**

The logic **S5** contains the following axioms and rules:

Pc	Axiomatization of Propositional Calculus
K	$K(\varphi \rightarrow \psi) \rightarrow (K\varphi \rightarrow K\psi)$
T	$K\varphi \rightarrow \varphi$
4	$K\varphi \rightarrow KK\varphi$
5	$\neg K\varphi \rightarrow K\neg K\varphi$
MP	$\dfrac{\varphi \qquad \varphi \rightarrow \psi}{\psi}$
Nec	$\dfrac{\varphi}{K\psi}$

The Logic **S5**

The logic **S5** contains the following axioms and rules:

Pc	Axiomatization of Propositional Calculus
K	$K(\varphi \rightarrow \psi) \rightarrow (K\varphi \rightarrow K\psi)$
T	$K\varphi \rightarrow \varphi$
4	$K\varphi \rightarrow KK\varphi$
5	$\neg K\varphi \rightarrow K\neg K\varphi$
MP	$\dfrac{\varphi \qquad \varphi \rightarrow \psi}{\psi}$
Nec	$\dfrac{\varphi}{K\psi}$

Theorem

S5 *is sound and strongly complete with respect to the class of Kripke frames with equivalence relations.*

The Logic **KD45**

The logic **S5** contains the following axioms and rules:

Pc	Axiomatization of Propositional Calculus
K	$B(\varphi \rightarrow \psi) \rightarrow (B\varphi \rightarrow B\psi)$
D	$\neg B\bot \quad (B\varphi \rightarrow \neg B\neg\varphi)$
4	$B\varphi \rightarrow BB\varphi$
5	$\neg B\varphi \rightarrow B\neg B\varphi$
MP	$\dfrac{\varphi \qquad \varphi \rightarrow \psi}{\psi}$
Nec	$\dfrac{\varphi}{B\psi}$

The Logic **KD45**

The logic **S5** contains the following axioms and rules:

Pc	Axiomatization of Propositional Calculus
K	$B(\varphi \rightarrow \psi) \rightarrow (B\varphi \rightarrow B\psi)$
D	$\neg B\bot \quad (B\varphi \rightarrow \neg B\neg\varphi)$
4	$B\varphi \rightarrow BB\varphi$
5	$\neg B\varphi \rightarrow B\neg B\varphi$
MP	$\dfrac{\varphi \qquad \varphi \rightarrow \psi}{\psi}$
Nec	$\dfrac{\varphi}{B\psi}$

Theorem

KD45 *is sound and strongly complete with respect to the class of Kripke frames with pseudo-equivalence relations (reflexive, transitive and serial).*

Truth Axiom/Consistency

$$K\varphi \to \varphi$$

$$\neg B\bot$$

Negative Introspection

$$\neg\Box\varphi \rightarrow \Box\neg\Box\varphi$$

$(\Box = K, B)$

Why would an agent not know some fact φ? (i.e., why would $\neg K_i\varphi$ be true?)

Why would an agent not know some fact φ? (i.e., why would $\neg K_i\varphi$ be true?)

- The agent may or may not believe φ, but has not ruled out all the $\neg\varphi$-worlds

Why would an agent not know some fact φ? (i.e., why would $\neg K_i\varphi$ be true?)

- The agent may or may not believe φ, but has not ruled out all the $\neg\varphi$-worlds
- The agent may believe φ and ruled-out the $\neg\varphi$-worlds, but this was based on “bad” evidence, or was not justified, or the agent was “epistemically lucky” (e.g., Gettier cases),...

Why would an agent not know some fact φ? (i.e., why would $\neg K_i\varphi$ be true?)

- The agent may or may not believe φ, but has not ruled out all the $\neg\varphi$-worlds
- The agent may believe φ and ruled-out the $\neg\varphi$-worlds, but this was based on "bad" evidence, or was not justified, or the agent was "epistemically lucky" (e.g., Gettier cases),...
- The agent has not yet entertained possibilities relevant to the truth of φ (the agent is unaware of φ).

Positive Introspection

$$\Box\varphi \rightarrow \Box\Box\varphi$$

$(\Box = K, B)$

The KK Principle

More famous is the “KK principle” (or “positive introspection”):

$$4_i \quad K_i\varphi \to K_iK_i\varphi.$$

Hintikka, one of the inventors of epistemic logic, endorsed the 4 axiom—at least for what he considered a strong notion of knowledge, found in philosophy from Aristotle to Schopenhauer.

The KK Principle

More famous is the “KK principle” (or “positive introspection”):

$$4_i \quad K_i\varphi \to K_iK_i\varphi.$$

Hintikka, one of the inventors of epistemic logic, endorsed the 4 axiom—at least for what he considered a strong notion of knowledge, found in philosophy from Aristotle to Schopenhauer.

J. Hintikka. *Knowledge and Belief*. Cornell University Press, 1962.

Hintikka rejected arguments for 4 based on claims about agents introspective powers, or what he called “the myth of the self-illumination of certain mental activities” (67).

The KK Principle

More famous is the "KK principle" (or "positive introspection"):

$$4_i \quad K_i\varphi \rightarrow K_iK_i\varphi.$$

Hintikka, one of the inventors of epistemic logic, endorsed the 4 axiom—at least for what he considered a strong notion of knowledge, found in philosophy from Aristotle to Schopenhauer.

J. Hintikka. *Knowledge and Belief*. Cornell University Press, 1962.

Hintikka rejected arguments for 4 based on claims about agents introspective powers, or what he called "the myth of the self-illumination of certain mental activities" (67). Instead, his claim was that for a strong notion of knowledge, *knowing that one knows* "differs only in words" from *knowing* (§2.1-2.2).

How Many Modalities?

Fact. In **S5** and **KD45**, there are only three modalities ($\Box$, $\Diamond$, and the "empty modality")

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?

Wes Holliday. *“Simplifying the Surprise Exam.”*. UC Berkeley Working paper in Philosophy, 2016.

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 $m \in \mathbb{N}$:

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

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

(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).

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

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

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, $K_1\varphi \rightarrow \varphi$, or the "weak factivity" axiom J_1, $K_1\neg K_1\varphi \rightarrow \neg K_1\varphi$ (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...

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.

Comparison with $n = 3$ Case

The generalizations of (A), (B), and (C) to the $n = 3$ case are:

(A^3) $K_1((p_1 \wedge \neg K_1 p_1) \vee (p_2 \wedge \neg K_2 p_2) \vee (p_3 \wedge \neg K_3 p_3))$;

(B^3) $K_1(((p_2 \vee p_3) \rightarrow K_2 \neg p_1) \wedge (p_3 \rightarrow K_3 \neg (p_1 \vee p_2))$;

(C^3) $K_1(K_2(p_1 \vee p_2 \vee p_3) \wedge K_3(p_1 \vee p_2 \vee p_3))$.

Interestingly, as we will show later, these assumptions are *consistent* even if we make strong assumptions about knowledge.

Comparison with $n = 3$ Case

The generalizations of (A), (B), and (C) to the $n = 3$ case are:

(A^3) $K_1((p_1 \wedge \neg K_1 p_1) \vee (p_2 \wedge \neg K_2 p_2) \vee (p_3 \wedge \neg K_3 p_3))$;

(B^3) $K_1(((p_2 \vee p_3) \to K_2 \neg p_1) \wedge (p_3 \to K_3 \neg(p_1 \vee p_2))$;

(C^3) $K_1(K_2(p_1 \vee p_2 \vee p_3) \wedge K_3(p_1 \vee p_2 \vee p_3))$.

If you think about the clever student's reasoning, he assumes that if he knows something, then he will continue to know it (or, for the designated student, then the students behind him in line know it):

$$4_1^{<} \quad K_1\varphi \to K_1 K_i \varphi \quad i > 1$$

Comparison with $n = 3$ Case

The generalizations of (A), (B), and (C) to the $n = 3$ case are:

(A^3) $K_1((p_1 \wedge \neg K_1 p_1) \vee (p_2 \wedge \neg K_2 p_2) \vee (p_3 \wedge \neg K_3 p_3))$;

(B^3) $K_1(((p_2 \vee p_3) \to K_2 \neg p_1) \wedge (p_3 \to K_3 \neg(p_1 \vee p_2))$;

(C^3) $K_1(K_2(p_1 \vee p_2 \vee p_3) \wedge K_3(p_1 \vee p_2 \vee p_3))$.

Using the axiom

$$4_1^< \quad K_1\varphi \to K_1 K_i \varphi \quad i > 1,$$

we can get into trouble starting from (A^3) and (B^3).

Comparison with $n = 3$ Case

The generalizations of (A), (B), and (C) to the $n = 3$ case are:

(A^3) $K_1((p_1 \wedge \neg K_1 p_1) \vee (p_2 \wedge \neg K_2 p_2) \vee (p_3 \wedge \neg K_3 p_3))$;

(B^3) $K_1(((p_2 \vee p_3) \rightarrow K_2 \neg p_1) \wedge (p_3 \rightarrow K_3 \neg(p_1 \vee p_2))$;

(C^3) $K_1(K_2(p_1 \vee p_2 \vee p_3) \wedge K_3(p_1 \vee p_2 \vee p_3))$.

Using the axiom

$$4_1^< \quad K_1\varphi \rightarrow K_1 K_i \varphi \quad i > 1,$$

we can get into trouble starting from (A^3) and (B^3).
Indeed, the following result holds for any $n > 2$. See

Wes Holliday. "Simplifying the Surprise Exam." (email for manuscript)

Comparison with $n = 3$ Case

The generalizations of (A), (B), and (C) to the $n = 3$ case are:

(A^3) $K_1((p_1 \wedge \neg K_1 p_1) \vee (p_2 \wedge \neg K_2 p_2) \vee (p_3 \wedge \neg K_3 p_3))$;

(B^3) $K_1(((p_2 \vee p_3) \rightarrow K_2 \neg p_1) \wedge (p_3 \rightarrow K_3 \neg (p_1 \vee p_2))$;

(C^3) $K_1(K_2(p_1 \vee p_2 \vee p_3) \wedge K_3(p_1 \vee p_2 \vee p_3))$.

For convenience, let's use the following abbreviation for "surprise":

$$S_i := (p_i \wedge \neg K_i p_i).$$

Comparison with $n = 3$ Case

The generalizations of (A), (B), and (C) to the $n = 3$ case are:

(A^3) $K_1(S_1 \vee S_2 \vee S_3)$;

(B^3) $K_1(((p_2 \vee p_3) \rightarrow K_2\neg p_1) \wedge (p_3 \rightarrow K_3\neg(p_1 \vee p_2))$;

(C^3) $K_1(K_2(p_1 \vee p_2 \vee p_3) \wedge K_3(p_1 \vee p_2 \vee p_3))$.

For convenience, let's use the following abbreviation for "surprise":

$$S_i := (p_i \wedge \neg K_i p_i).$$

Let us now show: $\{(A^3), (B^3)\} \vdash_{\mathbf{K4}_1^<} K_1(p \wedge \neg K_1 p_1)$

Let us now show: $\{(A^3), (B^3)\} \vdash_{\mathbf{K4}_1^<} K_1(p \wedge \neg K_1 p_1)$

(A^3) $K_1(S_1 \vee S_2 \vee S_3)$;

(B^3) $K_1(((p_2 \vee p_3) \rightarrow K_2\neg p_1) \wedge (p_3 \rightarrow K_3\neg(p_1 \vee p_2))$;

Let us now show: $\{(A^3), (B^3)\} \vdash_{\mathbf{K4}_1^<} K_1(p \wedge \neg K_1 p_1)$

(A^3) $K_1(S_1 \vee S_2 \vee S_3)$;

(B^3) $K_1(((p_2 \vee p_3) \to K_2 \neg p_1) \wedge (p_3 \to K_3 \neg (p_1 \vee p_2))$;

(D^3) $K_1(K_2(S_1 \vee S_2 \vee S_3) \wedge K_3(p_1 \vee p_2 \vee p_3))$ from (A^3), $4_1^<$, RK_3, PL

Let us now show: $\{(A^3), (B^3)\} \vdash_{\mathbf{K4}_1^<} K_1(p \wedge \neg K_1 p_1)$

(A^3) $K_1(S_1 \vee S_2 \vee S_3)$;

(B^3) $K_1(((p_2 \vee p_3) \rightarrow K_2\neg p_1) \wedge (p_3 \rightarrow K_3\neg(p_1 \vee p_2)))$;

(D^3) $K_1(K_2(S_1 \vee S_2 \vee S_3) \wedge K_3(p_1 \vee p_2 \vee p_3))$ from (A^3), $4_1^<$, RK_3, PL

$(3,1)$ $(K_3(p_1 \vee p_2 \vee p_3) \wedge K_3\neg(p_1 \vee p_2)) \rightarrow K_3 p_3$ by PL and RK_3

Let us now show: $\{(A^3), (B^3)\} \vdash_{\mathbf{K4}_1^<} K_1(p \wedge \neg K_1 p_1)$

(A^3) $K_1(S_1 \vee S_2 \vee S_3)$;

(B^3) $K_1(((p_2 \vee p_3) \rightarrow K_2\neg p_1) \wedge (p_3 \rightarrow K_3\neg(p_1 \vee p_2))$;

(D^3) $K_1(K_2(S_1 \vee S_2 \vee S_3) \wedge K_3(p_1 \vee p_2 \vee p_3))$ from (A^3), $4_1^<$, RK_3, PL

$(3,1)$ $(K_3(p_1 \vee p_2 \vee p_3) \wedge K_3\neg(p_1 \vee p_2)) \rightarrow K_3 p_3$ by PL and RK_3

$(3,2)$ $K_1((K_3(p_1 \vee p_2 \vee p_3) \wedge K_3\neg(p_1 \vee p_2)) \rightarrow K_3 p_3)$ from $(3,1)$ by Nec_1

Let us now show: $\{(A^3), (B^3)\} \vdash_{\mathbf{K4}_1^<} K_1(p \wedge \neg K_1 p_1)$

(A^3) $K_1(S_1 \vee S_2 \vee S_3)$;

(B^3) $K_1(((p_2 \vee p_3) \rightarrow K_2\neg p_1) \wedge (p_3 \rightarrow K_3\neg(p_1 \vee p_2))$;

(D^3) $K_1(K_2(S_1 \vee S_2 \vee S_3) \wedge K_3(p_1 \vee p_2 \vee p_3))$ from (A^3), $4_1^<$, RK$_3$, PL

$(3,1)$ $(K_3(p_1 \vee p_2 \vee p_3) \wedge K_3\neg(p_1 \vee p_2)) \rightarrow K_3 p_3$ by PL and RK$_3$

$(3,2)$ $K_1((K_3(p_1 \vee p_2 \vee p_3) \wedge K_3\neg(p_1 \vee p_2)) \rightarrow K_3 p_3)$ from $(3,1)$ by Nec$_1$

$(3,3)$ $K_1(K_3\neg(p_1 \vee p_2) \rightarrow K_3 p_3)$ from (D^3), $(3,2)$ using RK$_1$ and PL

Let us now show: $\{(A^3), (B^3)\} \vdash_{\mathbf{K4}_1^<} K_1(p \wedge \neg K_1 p_1)$

(A^3) $K_1(S_1 \vee S_2 \vee S_3)$;

(B^3) $K_1(((p_2 \vee p_3) \rightarrow K_2\neg p_1) \wedge (p_3 \rightarrow K_3\neg(p_1 \vee p_2))$;

(D^3) $K_1(K_2(S_1 \vee S_2 \vee S_3) \wedge K_3(p_1 \vee p_2 \vee p_3))$ from (A^3), $4_1^<$, RK$_3$, PL

$(3,1)$ $(K_3(p_1 \vee p_2 \vee p_3) \wedge K_3\neg(p_1 \vee p_2)) \rightarrow K_3 p_3$ by PL and RK$_3$

$(3,2)$ $K_1((K_3(p_1 \vee p_2 \vee p_3) \wedge K_3\neg(p_1 \vee p_2)) \rightarrow K_3 p_3)$ from $(3,1)$ by Nec$_1$

$(3,3)$ $K_1(K_3\neg(p_1 \vee p_2) \rightarrow K_3 p_3)$ from (D^3), $(3,2)$ using RK$_1$ and PL

$(3,4)$ $K_1\neg S_3$ from (B^3), $(3,3)$ using RK$_1$ and PL

Let us now show: $\{(A^3), (B^3)\} \vdash_{\mathbf{K4}_1^<} K_1(p \wedge \neg K_1 p_1)$

(A^3) $K_1(S_1 \vee S_2 \vee S_3)$;

(B^3) $K_1(((p_2 \vee p_3) \rightarrow K_2\neg p_1) \wedge (p_3 \rightarrow K_3\neg(p_1 \vee p_2))$;

(D^3) $K_1(K_2(S_1 \vee S_2 \vee S_3) \wedge K_3(p_1 \vee p_2 \vee p_3))$ from (A^3), $4_1^<$, RK_3, PL

$(3,1)$ $(K_3(p_1 \vee p_2 \vee p_3) \wedge K_3\neg(p_1 \vee p_2)) \rightarrow K_3 p_3$ by PL and RK_3

$(3,2)$ $K_1((K_3(p_1 \vee p_2 \vee p_3) \wedge K_3\neg(p_1 \vee p_2)) \rightarrow K_3 p_3)$ from $(3,1)$ by Nec_1

$(3,3)$ $K_1(K_3\neg(p_1 \vee p_2) \rightarrow K_3 p_3)$ from (D^3), $(3,2)$ using RK_1 and PL

$(3,4)$ $K_1\neg S_3$ from (B^3), $(3,3)$ using RK_1 and PL

$(2,0)$ $K_1 K_2 \neg S_3$ from $(3,4)$ by $4_1^<$

Let us now show: $\{(A^3), (B^3)\} \vdash_{\mathbf{K4}_1^<} K_1(p \wedge \neg K_1 p_1)$

(A^3) $K_1(S_1 \vee S_2 \vee S_3)$;

(B^3) $K_1(((p_2 \vee p_3) \rightarrow K_2\neg p_1) \wedge (p_3 \rightarrow K_3\neg(p_1 \vee p_2))$;

(D^3) $K_1(K_2(S_1 \vee S_2 \vee S_3) \wedge K_3(p_1 \vee p_2 \vee p_3))$ from (A^3), $4_1^<$, RK_3, PL

$(3,1)$ $(K_3(p_1 \vee p_2 \vee p_3) \wedge K_3\neg(p_1 \vee p_2)) \rightarrow K_3 p_3$ by PL and RK_3

$(3,2)$ $K_1((K_3(p_1 \vee p_2 \vee p_3) \wedge K_3\neg(p_1 \vee p_2)) \rightarrow K_3 p_3)$ from $(3,1)$ by Nec_1

$(3,3)$ $K_1(K_3\neg(p_1 \vee p_2) \rightarrow K_3 p_3)$ from (D^3), $(3,2)$ using RK_1 and PL

$(3,4)$ $K_1\neg S_3$ from (B^3), $(3,3)$ using RK_1 and PL

$(2,0)$ $K_1 K_2\neg S_3$ from $(3,4)$ by $4_1^<$

$(2,1)$ $(K_2(S_1 \vee S_2 \vee S_3) \wedge K_2\neg p_1 \wedge K_2\neg S_3) \rightarrow K_2 p_2$ by PL and RK_2

Let us now show: $\{(A^3), (B^3)\} \vdash_{\mathbf{K4}_1^<} K_1(p \wedge \neg K_1 p_1)$

(A^3) $K_1(S_1 \vee S_2 \vee S_3)$;

(B^3) $K_1(((p_2 \vee p_3) \rightarrow K_2\neg p_1) \wedge (p_3 \rightarrow K_3\neg(p_1 \vee p_2)))$;

(D^3) $K_1(K_2(S_1 \vee S_2 \vee S_3) \wedge K_3(p_1 \vee p_2 \vee p_3))$ from (A^3), $4_1^<$, RK_3, PL

$(3,1)$ $(K_3(p_1 \vee p_2 \vee p_3) \wedge K_3\neg(p_1 \vee p_2)) \rightarrow K_3 p_3$ by PL and RK_3

$(3,2)$ $K_1((K_3(p_1 \vee p_2 \vee p_3) \wedge K_3\neg(p_1 \vee p_2)) \rightarrow K_3 p_3)$ from $(3,1)$ by Nec_1

$(3,3)$ $K_1(K_3\neg(p_1 \vee p_2) \rightarrow K_3 p_3)$ from (D^3), $(3,2)$ using RK_1 and PL

$(3,4)$ $K_1\neg S_3$ from (B^3), $(3,3)$ using RK_1 and PL

$(2,0)$ $K_1 K_2 \neg S_3$ from $(3,4)$ by $4_1^<$

$(2,1)$ $(K_2(S_1 \vee S_2 \vee S_3) \wedge K_2\neg p_1 \wedge K_2\neg S_3) \rightarrow K_2 p_2$ by PL and RK_2

$(2,2)$ $K_1((K_2(S_1 \vee S_2 \vee S_3) \wedge K_2\neg p_1 \wedge K_2\neg S_3) \rightarrow K_2 p_2)$ from $(2,1)$ by Nec_1

Let us now show: $\{(A^3),(B^3)\} \vdash_{\mathbf{K4}_1^<} K_1(p \wedge \neg K_1 p_1)$

(A^3) $K_1(S_1 \vee S_2 \vee S_3)$;

(B^3) $K_1(((p_2 \vee p_3) \to K_2\neg p_1) \wedge (p_3 \to K_3\neg(p_1 \vee p_2))$;

(D^3) $K_1(K_2(S_1 \vee S_2 \vee S_3) \wedge K_3(p_1 \vee p_2 \vee p_3))$ from (A^3), $4_1^<$, RK_3, PL

$(3,1)$ $(K_3(p_1 \vee p_2 \vee p_3) \wedge K_3\neg(p_1 \vee p_2)) \to K_3 p_3$ by PL and RK_3

$(3,2)$ $K_1((K_3(p_1 \vee p_2 \vee p_3) \wedge K_3\neg(p_1 \vee p_2)) \to K_3 p_3)$ from $(3,1)$ by Nec_1

$(3,3)$ $K_1(K_3\neg(p_1 \vee p_2) \to K_3 p_3)$ from (D^3), $(3,2)$ using RK_1 and PL

$(3,4)$ $K_1\neg S_3$ from (B^3), $(3,3)$ using RK_1 and PL

$(2,0)$ $K_1 K_2\neg S_3$ from $(3,4)$ by $4_1^<$

$(2,1)$ $(K_2(S_1 \vee S_2 \vee S_3) \wedge K_2\neg p_1 \wedge K_2\neg S_3) \to K_2 p_2$ by PL and RK_2

$(2,2)$ $K_1((K_2(S_1 \vee S_2 \vee S_3) \wedge K_2\neg p_1 \wedge K_2\neg S_3) \to K_2 p_2)$ from $(2,1)$ by Nec_1

$(2,3)$ $K_1(K_2\neg p_1 \to K_2 p_2)$ from (D^3), $(2,0)$, $(2,2)$ using RK_1 and PL

Let us now show: $\{(A^3), (B^3)\} \vdash_{\mathbf{K4}_1^<} K_1(p \wedge \neg K_1 p_1)$

(A^3) $K_1(S_1 \vee S_2 \vee S_3)$;

(B^3) $K_1(((p_2 \vee p_3) \rightarrow K_2\neg p_1) \wedge (p_3 \rightarrow K_3\neg(p_1 \vee p_2))$;

(D^3) $K_1(K_2(S_1 \vee S_2 \vee S_3) \wedge K_3(p_1 \vee p_2 \vee p_3))$ from (A^3), $4_1^<$, RK$_3$, PL

$(3,1)$ $(K_3(p_1 \vee p_2 \vee p_3) \wedge K_3\neg(p_1 \vee p_2)) \rightarrow K_3 p_3$ by PL and RK$_3$

$(3,2)$ $K_1((K_3(p_1 \vee p_2 \vee p_3) \wedge K_3\neg(p_1 \vee p_2)) \rightarrow K_3 p_3)$ from $(3,1)$ by Nec$_1$

$(3,3)$ $K_1(K_3\neg(p_1 \vee p_2) \rightarrow K_3 p_3)$ from (D^3), $(3,2)$ using RK$_1$ and PL

$(3,4)$ $K_1\neg S_3$ from (B^3), $(3,3)$ using RK$_1$ and PL

$(2,0)$ $K_1 K_2 \neg S_3$ from $(3,4)$ by $4_1^<$

$(2,1)$ $(K_2(S_1 \vee S_2 \vee S_3) \wedge K_2\neg p_1 \wedge K_2\neg S_3) \rightarrow K_2 p_2$ by PL and RK$_2$

$(2,2)$ $K_1((K_2(S_1 \vee S_2 \vee S_3) \wedge K_2\neg p_1 \wedge K_2\neg S_3) \rightarrow K_2 p_2)$ from $(2,1)$ by Nec$_1$

$(2,3)$ $K_1(K_2\neg p_1 \rightarrow K_2 p_2)$ from (D^3), $(2,0)$, $(2,2)$ using RK$_1$ and PL

$(2,4)$ $K_1\neg S_2$ from (B^3), $(2,3)$ using RK$_1$ and PL

Let us now show: $\{(A^3), (B^3)\} \vdash_{\mathbf{K4}_1^<} K_1(p \wedge \neg K_1 p_1)$

(A^3) $K_1(S_1 \vee S_2 \vee S_3)$;

(B^3) $K_1(((p_2 \vee p_3) \rightarrow K_2\neg p_1) \wedge (p_3 \rightarrow K_3\neg(p_1 \vee p_2))$;

(D^3) $K_1(K_2(S_1 \vee S_2 \vee S_3) \wedge K_3(p_1 \vee p_2 \vee p_3))$ from (A^3), $4_1^<$, RK_3, PL

$(3,1)$ $(K_3(p_1 \vee p_2 \vee p_3) \wedge K_3\neg(p_1 \vee p_2)) \rightarrow K_3 p_3$ by PL and RK_3

$(3,2)$ $K_1((K_3(p_1 \vee p_2 \vee p_3) \wedge K_3\neg(p_1 \vee p_2)) \rightarrow K_3 p_3)$ from $(3,1)$ by Nec_1

$(3,3)$ $K_1(K_3\neg(p_1 \vee p_2) \rightarrow K_3 p_3)$ from (D^3), $(3,2)$ using RK_1 and PL

$(3,4)$ $K_1\neg S_3$ from (B^3), $(3,3)$ using RK_1 and PL

$(2,0)$ $K_1 K_2\neg S_3$ from $(3,4)$ by $4_1^<$

$(2,1)$ $(K_2(S_1 \vee S_2 \vee S_3) \wedge K_2\neg p_1 \wedge K_2\neg S_3) \rightarrow K_2 p_2$ by PL and RK_2

$(2,2)$ $K_1((K_2(S_1 \vee S_2 \vee S_3) \wedge K_2\neg p_1 \wedge K_2\neg S_3) \rightarrow K_2 p_2)$ from $(2,1)$ by Nec_1

$(2,3)$ $K_1(K_2\neg p_1 \rightarrow K_2 p_2)$ from (D^3), $(2,0)$, $(2,2)$ using RK_1 and PL

$(2,4)$ $K_1\neg S_2$ from (B^3), $(2,3)$ using RK_1 and PL

$(2,5)$ $K_1 S_1$ from (A^3), $(3,4)$, $(2,4)$ using RK_1 and PL

Comparison with $n = 3$ Case

(A^3) $K_1((p_1 \wedge \neg K_1 p_1) \vee (p_2 \wedge \neg K_2 p_2) \vee (p_3 \wedge \neg K_3 p_3))$;

(B^3) $K_1(((p_2 \vee p_3) \to K_2 \neg p_1) \wedge (p_3 \to K_3 \neg (p_1 \vee p_2))$.

As before, given $\{(A^3), (B^3)\} \vdash_{\mathbf{K4}_1^<} K_1(p \wedge \neg K_1 p_1)$, we also have:

$$\{(A^3), (B^3)\} \vdash_{\mathbf{KT_14}_1^<} \perp \text{ and } \{(A^3), (B^3)\} \vdash_{\mathbf{KJ_14}_1^<} \perp.$$

Comparison with $n = 3$ Case

(A^3) $K_1((p_1 \wedge \neg K_1 p_1) \vee (p_2 \wedge \neg K_2 p_2) \vee (p_3 \wedge \neg K_3 p_3))$;

(B^3) $K_1(((p_2 \vee p_3) \to K_2 \neg p_1) \wedge (p_3 \to K_3 \neg (p_1 \vee p_2))$.

As before, given $\{(A^3), (B^3)\} \vdash_{\mathbf{K4}_1^<} K_1(p \wedge \neg K_1 p_1)$, we also have:

$$\{(A^3), (B^3)\} \vdash_{\mathbf{KT_1 4}_1^<} \bot \text{ and } \{(A^3), (B^3)\} \vdash_{\mathbf{KJ_1 4}_1^<} \bot.$$

Thus, we must reject (A^3), (B^3), the rule RK or the axiom

$$4_1^< \quad K_1\varphi \to K_1 K_i \varphi \quad i > 1.$$

Summary

- $\{(A^2), (B^2), (C^2)\} \vdash_{\mathbf{K}} K_1(p_1 \wedge \neg K_1);$
- $\{(A^2), (B^2), (C^2)\} \vdash_{\mathbf{KJ_1}} \perp$ and $\{(A^2), (B^2), (C^2)\} \vdash_{\mathbf{KT_1}} \perp;$

Summary

- $\{(A^2), (B^2), (C^2)\} \vdash_{\mathbf{K}} K_1(p_1 \wedge \neg K_1)$;
- $\{(A^2), (B^2), (C^2)\} \vdash_{\mathbf{KJ_1}} \bot$ and $\{(A^2), (B^2), (C^2)\} \vdash_{\mathbf{KT_1}} \bot$;
- $\{(A^3), (B^3), (C^3)\} \nvdash_{\mathbf{S5}} \bot$.

Summary

- $\{(A^2), (B^2), (C^2)\} \vdash_{\mathbf{K}} K_1(p_1 \wedge \neg K_1)$;
- $\{(A^2), (B^2), (C^2)\} \vdash_{\mathbf{KJ_1}} \bot$ and $\{(A^2), (B^2), (C^2)\} \vdash_{\mathbf{KT_1}} \bot$;
- $\{(A^3), (B^3), (C^3)\} \nvdash_{\mathbf{S5}} \bot$.
- $\{(A^3), (B^3)\} \vdash_{\mathbf{K4_1^<}} K_1(p_1 \wedge \neg K_1)$;
- $\{(A^3), (B^3)\} \vdash_{\mathbf{KJ_14_1^<}} \bot$ and $\{(A^3), (B^3)\} \vdash_{\mathbf{KT_14_1^<}} \bot$;

Summary

- $\{(A^2), (B^2), (C^2)\} \vdash_{\mathbf{K}} K_1(p_1 \wedge \neg K_1)$;
- $\{(A^2), (B^2), (C^2)\} \vdash_{\mathbf{KJ_1}} \bot$ and $\{(A^2), (B^2), (C^2)\} \vdash_{\mathbf{KT_1}} \bot$;
- $\{(A^3), (B^3), (C^3)\} \nvdash_{\mathbf{S5}} \bot$.
- $\{(A^3), (B^3)\} \vdash_{\mathbf{K4_1^<}} K_1(p_1 \wedge \neg K_1)$;
- $\{(A^3), (B^3)\} \vdash_{\mathbf{KJ_14_1^<}} \bot$ and $\{(A^3), (B^3)\} \vdash_{\mathbf{KT_14_1^<}} \bot$;

With these facts, one can make a strong case that the culprit behind the paradoxes is the (mistaken) $4_1^<$ axiom, $K_1\varphi \to K_1K_i\varphi$ $(i > 1)$....

Wes Holliday. *"Simplifying the Surprise Exam."*. UC Berkeley Working paper in Philosophy, 2016.

The “Problem” of Logical Omniscience

The rule

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

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

The “Problem” of Logical Omniscience

The rule

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

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

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

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

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

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

$\mathcal{M}, w \models t : \varphi$ 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

Summary

(Multi-agent) **S5** is a logic of “knowledge”

(Multi-agent) **KD45** is a logic of “belief”

Summary

(Multi-agent) **S5** is a logic of “knowledge”

(Multi-agent) **KD45** is a logic of “belief”

Two issues:

- Modeling awareness/unawareness
- Logics with both knowledge and belief operators

Unawareness

Why would an agent not know some fact φ? (i.e., why would $\neg K_i\varphi$ be true?)

Unawareness

Why would an agent not know some fact φ? (i.e., why would $\neg K_i\varphi$ be true?)

- The agent may or may not believe φ, but has not ruled out all the $\neg\varphi$-worlds

Unawareness

Why would an agent not know some fact φ? (i.e., why would $\neg K_i\varphi$ be true?)

- The agent may or may not believe φ, but has not ruled out all the $\neg\varphi$-worlds
- The agent may believe φ and ruled-out the $\neg\varphi$-worlds, but this was based on "bad" evidence, or was not justified, or the agent was "epistemically lucky" (e.g., Gettier cases),...

Unawareness

Why would an agent not know some fact φ? (i.e., why would $\neg K_i\varphi$ be true?)

- The agent may or may not believe φ, but has not ruled out all the $\neg\varphi$-worlds
- The agent may believe φ and ruled-out the $\neg\varphi$-worlds, but this was based on "bad" evidence, or was not justified, or the agent was "epistemically lucky" (e.g., Gettier cases),...
- The agent has not yet entertained possibilities relevant to the truth of φ (the agent is unaware of φ).

Can we model unawareness in state-space models?

Can we model unawareness in state-space models?

E. Dekel, B. Lipman and A. Rustichini. *Standard State-Space Models Preclude Unawareness*. Econometrica, 55:1, pp. 159 - 173 (1998).

Properties of Unawareness

1. $U\varphi \rightarrow (\neg K\varphi \wedge \neg K \neg K\varphi)$

Properties of Unawareness

1. $U\varphi \to (\neg K\varphi \wedge \neg K\neg K\varphi)$
2. $\neg KU\varphi$

Properties of Unawareness

1. $U\varphi \to (\neg K\varphi \wedge \neg K \neg K\varphi)$
2. $\neg KU\varphi$
3. $U\varphi \to UU\varphi$

Properties of Unawareness

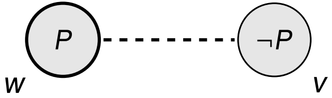
1. $U\varphi \to (\neg K\varphi \wedge \neg K\neg K\varphi)$
2. $\neg KU\varphi$
3. $U\varphi \to UU\varphi$

Theorem. In any logic where U satisfies the above axiom schemes, we have

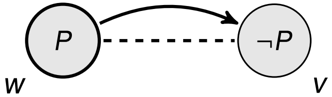
1. If K satisfies Necessitation (from φ infer $K\varphi$), then for all formulas φ, $\neg U\varphi$ is derivable (the agent is aware of everything); and
2. If K satisfies Monotonicity (from $\varphi \to \psi$ infer $K\varphi \to K\psi$), then for all φ and ψ, $U\varphi \to \neg K\psi$ is derivable (if the agent is unaware of something then the agent does not know anything).

B. Schipper. *Online Bibliography on Models of Unawareness*. `http://www.econ.ucdavis.edu/faculty/schipper/unaw.htm`.

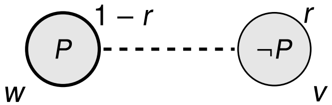
J. Halpern. *Alternative semantics for unawareness*. Games and Economic Behavior, 37, 321-339, 2001.



Ann does not know that P



Ann does not know that P, but she believes that $\neg P$



Ann does not know that P, but she believes that $\neg P$ is true to degree r.

Combining Logics of Knowledge and Belief

$\mathcal{M} = \langle W, \{\sim_i\}_{i\in\mathcal{A}}, \{R_i\}_{i\in\mathcal{A}}, V\rangle$ where

- $W \neq \emptyset$ is a set of states;
- each $\sim_i$ is an equivalence relation on W;
- each R_i is a serial, transitive, Euclidean relation on W; and
- V is a valuation function.

Combining Logics of Knowledge and Belief

$\mathcal{M} = \langle W, \{\sim_i\}_{i\in\mathcal{A}}, \{R_i\}_{i\in\mathcal{A}}, V\rangle$ where

- $W \neq \emptyset$ is a set of states;
- each $\sim_i$ is an equivalence relation on W;
- each R_i is a serial, transitive, Euclidean relation on W; and
- V is a valuation function.

What is the relationship between knowledge (K_i) and believe (B_i)?

- Each K_i is **S5**

Combining Logics of Knowledge and Belief

$\mathcal{M} = \langle W, \{\sim_i\}_{i \in \mathcal{A}}, \{R_i\}_{i \in \mathcal{A}}, V \rangle$ where

- $W \neq \emptyset$ is a set of states;
- each $\sim_i$ is an equivalence relation on W;
- each R_i is a serial, transitive, Euclidean relation on W; and
- V is a valuation function.

What is the relationship between knowledge (K_i) and believe (B_i)?

- Each K_i is **S5**
- Each B_i is **KD45**

Combining Logics of Knowledge and Belief

$\mathcal{M} = \langle W, \{\sim_i\}_{i\in\mathcal{A}}, \{R_i\}_{i\in\mathcal{A}}, V\rangle$ where

- $W \neq \emptyset$ is a set of states;
- each $\sim_i$ is an equivalence relation on W;
- each R_i is a serial, transitive, Euclidean relation on W; and
- V is a valuation function.

What is the relationship between knowledge (K_i) and believe (B_i)?

- Each K_i is **S5**
- Each B_i is **KD45**
- $K_i\varphi \rightarrow B_i\varphi$? "knowledge implies belief"

Combining Logics of Knowledge and Belief

$\mathcal{M} = \langle W, \{\sim_i\}_{i\in\mathcal{A}}, \{R_i\}_{i\in\mathcal{A}}, V\rangle$ where

- $W \neq \emptyset$ is a set of states;
- each $\sim_i$ is an equivalence relation on W;
- each R_i is a serial, transitive, Euclidean relation on W; and
- V is a valuation function.

What is the relationship between knowledge (K_i) and believe (B_i)?

- Each K_i is **S5**
- Each B_i is **KD45**
- $K_i\varphi \rightarrow B_i\varphi$? “knowledge implies belief”
- $B_i\varphi \rightarrow B_iK_i\varphi$? “positive certainty”

Combining Logics of Knowledge and Belief

$\mathcal{M} = \langle W, \{\sim_i\}_{i\in\mathcal{A}}, \{R_i\}_{i\in\mathcal{A}}, V\rangle$ where

- $W \neq \emptyset$ is a set of states;
- each $\sim_i$ is an equivalence relation on W;
- each R_i is a serial, transitive, Euclidean relation on W; and
- V is a valuation function.

What is the relationship between knowledge (K_i) and believe (B_i)?

- Each K_i is **S5**
- Each B_i is **KD45**
- $K_i\varphi \rightarrow B_i\varphi$? "knowledge implies belief"
- $B_i\varphi \rightarrow B_iK_i\varphi$? "positive certainty"
- $B_i\varphi \rightarrow K_iB_i\varphi$? "strong introspection"

An Issue

- Suppose that p is something you are certain of (you *believe* it with probability one), but is false: $\neg p \wedge Bp$

An Issue

- Suppose that p is something you are certain of (you *believe* it with probability one), but is false: $\neg p \wedge Bp$

- Assuming 1. B satisfies **KD45**, 2. K satisfies **S5**, 3. knowledge implies believe and 4. positive certainty leads to a contradiction.

An Issue

- Suppose that p is something you are certain of (you *believe* it with probability one), but is false: $\neg p \wedge Bp$
- Assuming 1. B satisfies **KD45**, 2. K satisfies **S5**, 3. knowledge implies believe and 4. positive certainty leads to a contradiction.
- $Bp \rightarrow BKp$

An Issue

- Suppose that p is something you are certain of (you *believe* it with probability one), but is false: $\neg p \wedge Bp$
- Assuming 1. B satisfies **KD45**, 2. K satisfies **S5**, 3. knowledge implies believe and 4. positive certainty leads to a contradiction.
- $Bp \to BKp$
- $\neg p \to \neg Kp$

An Issue

- Suppose that p is something you are certain of (you *believe* it with probability one), but is false: $\neg p \wedge Bp$
- Assuming 1. B satisfies **KD45**, 2. K satisfies **S5**, 3. knowledge implies believe and 4. positive certainty leads to a contradiction.
- $Bp \to BKp$
- $\neg p \to \neg Kp \to K\neg Kp$

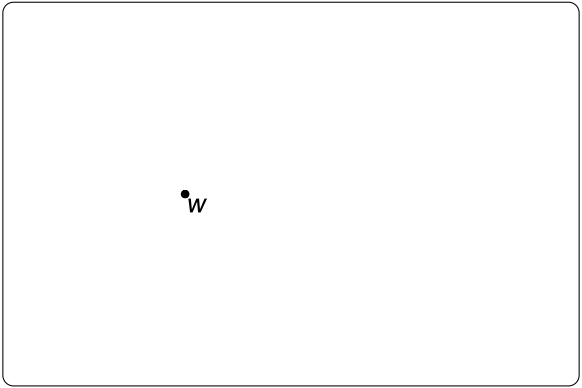
An Issue

- Suppose that p is something you are certain of (you *believe* it with probability one), but is false: $\neg p \wedge Bp$
- Assuming 1. B satisfies **KD45**, 2. K satisfies **S5**, 3. knowledge implies believe and 4. positive certainty leads to a contradiction.
- $Bp \rightarrow BKp$
- $\neg p \rightarrow \neg Kp \rightarrow K\neg Kp \rightarrow B\neg Kp$

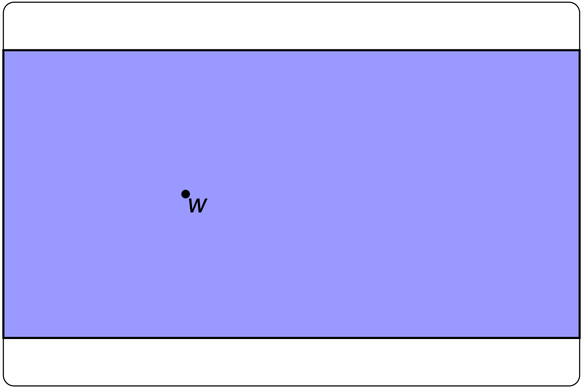
An Issue

- Suppose that p is something you are certain of (you *believe* it with probability one), but is false: $\neg p \wedge Bp$
- Assuming 1. B satisfies **KD45**, 2. K satisfies **S5**, 3. knowledge implies believe and 4. positive certainty leads to a contradiction.
- $Bp \rightarrow BKp$
- $\neg p \rightarrow \neg Kp \rightarrow K\neg Kp \rightarrow B\neg Kp$
- So, $BKp \wedge B\neg Kp$ also holds, but this contradictions $B\varphi \rightarrow \neg B\neg\varphi$.

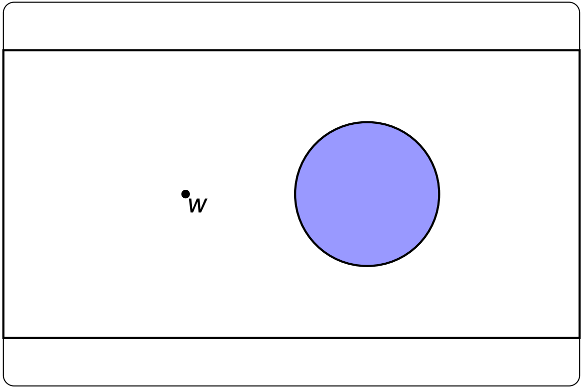
J. Halpern. *Should Knowledge Entail Belief?*. Journal of Philosophical Logic, 25:5, 1996, pp. 483-494.



- The set of states, with a distinguished state denoted the "actual world"



- The set of states, with a distinguished state denoted the "actual world"
- The agent's (hard) information (i.e., the states consistent with what the agent knows)



- The agent's (hard) information (i.e., the states consistent with what the agent knows)
- The agent's beliefs (soft information—-the states consistent with what the agent believes)

Digression on Belief Change, I

Digression on Belief Change, I

Consider the following beliefs of a rational agent:

- p_1 All Europeans swans are white.
- p_2 The bird caught in the trap is a swan.
- p_3 The bird caught in the trap comes from Sweden.
- p_4 Sweden is part of Europe.

Thus, the agent believes:

- q The bird caught in the trap is white.

Digression on Belief Change, I

Consider the following beliefs of a rational agent:

p_1 All Europeans swans are white.

p_2 The bird caught in the trap is a swan.

p_3 The bird caught in the trap comes from Sweden.

p_4 Sweden is part of Europe.

Thus, the agent believes:

q The bird caught in the trap is white.

Now suppose the rational agent—for example, You—learn that the bird caught in the trap is black ($\neg q$).

Digression on Belief Change, I

Consider the following beliefs of a rational agent:

- p_1 All Europeans swans are white.
- p_2 The bird caught in the trap is a swan.
- p_3 The bird caught in the trap comes from Sweden.
- p_4 Sweden is part of Europe.

Thus, the agent believes:

- q The bird caught in the trap is white.

Question: How should the agent incorporate $\neg q$ into his belief state to obtain a consistent belief state?

Digression on Belief Change, I

Consider the following beliefs of a rational agent:

- p_1 All Europeans swans are white.
- p_2 The bird caught in the trap is a swan.
- p_3 The bird caught in the trap comes from Sweden.
- p_4 Sweden is part of Europe.

Thus, the agent believes:

- q The bird caught in the trap is white.

Question: How should the agent incorporate $\neg q$ into his belief state to obtain a consistent belief state?
Problem: Logical considerations alone are insufficient to answer this question! Why??

Digression on Belief Change, I

Consider the following beliefs of a rational agent:

- p_1 All Europeans swans are white.
- p_2 The bird caught in the trap is a swan.
- p_3 The bird caught in the trap comes from Sweden.
- p_4 Sweden is part of Europe.

Thus, the agent believes:

- q The bird caught in the trap is white.

Question: How should the agent incorporate $\neg q$ into his belief state to obtain a consistent belief state?

Problem: Logical considerations alone are insufficient to answer this question! Why??

There are several logically consistent ways to incorporate $\neg q$!

Digression on Belief Change, II

What extralogical factors serve to determine what beliefs to give up and what beliefs to retain?

Digression on Belief Change, III

Belief revision is a matter of choice, and the choices are to be made in such a way that:

1. The resulting theory squares with the experience;
2. It is simple; and
3. The choices disturb the original theory as little as possible.

Digression on Belief Change, III

Belief revision is a matter of choice, and the choices are to be made in such a way that:

1. The resulting theory squares with the experience;
2. It is simple; and
3. The choices disturb the original theory as little as possible.

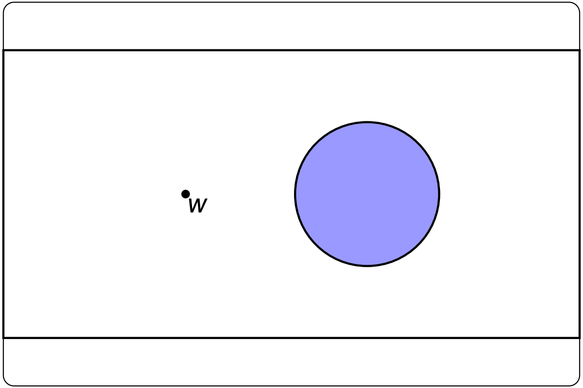
Research has relied on the following related guiding ideas:

1. When accepting a new piece of information, an agent should aim at a minimal change of his old beliefs.
2. If there are different ways to effect a belief change, the agent should give up those beliefs which are least entrenched.

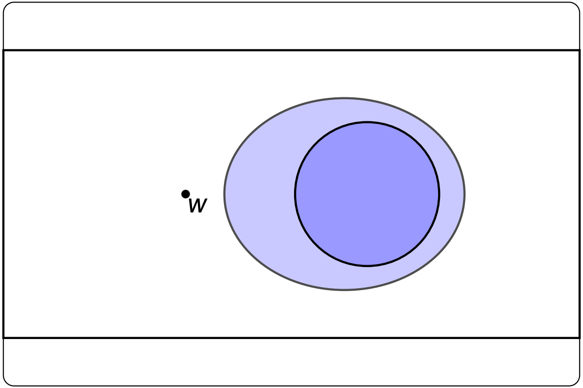
Digression: Belief Revision

A.P. Pedersen and H. Arló-Costa. *"Belief Revision"*. In Continuum Companion to Philosophical Logic. Continuum Press, 2011.

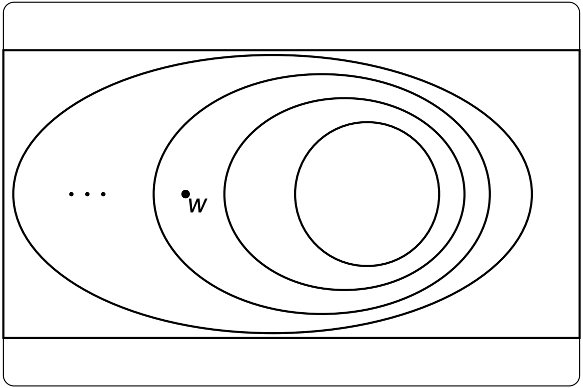
Hans Rott. *Change, Choice and Inference: A Study of Belief Revision and Nonmonotonic Reasoning*. Oxford University Press, 2001.



- The agent's (hard) information (i.e., the states consistent with what the agent knows)
- The agent's beliefs (soft information—-the states consistent with what the agent believes)



- The agent's beliefs (soft information—-the states consistent with what the agent believes)
- The agent's "contingency plan": when the stronger beliefs fail, go with the weaker ones.



- The agent's beliefs (soft information—-the states consistent with what the agent believes)
- The agent's "contingency plan": when the stronger beliefs fail, go with the weaker ones.

Sphere Models

Sphere Models

Let W be a set of states, A system of spheres $\mathcal{F} \subseteq \wp(W)$ such that:

- For each $S, S' \in \mathcal{F}$, either $S \subseteq S'$ or $S' \subseteq S$
- For any $P \subseteq W$ there is a smallest $S \in \mathcal{F}$ (according to the subset relation) such that $P \cap S \neq \emptyset$
- The spheres are non-empty $\bigcap \mathcal{F} \neq \emptyset$ and cover the entire information cell $\bigcup \mathcal{F} = W$ (or $[w] = \{v \mid w \sim v\}$)

Let $\mathcal{F}$ be a system of spheres on W: for $w, v \in W$, let

$$w \preceq_{\mathcal{F}} v \text{ iff for all } S \in \mathcal{F}, \text{ if } v \in S \text{ then } w \in S$$

Then, $\preceq_{\mathcal{F}}$ is reflexive, transitive, and well-founded.

$w \preceq_{\mathcal{F}} v$ means that: no matter what the agent learns in the future, as long as world v is still consistent with her beliefs and w is still epistemically possible, then w is also consistent with her beliefs.

Plausibility Models

Epistemic Models: $\mathcal{M} = \langle W, \{\sim_i\}_{i\in\mathcal{A}}, V\rangle$

Truth: $\mathcal{M}, w \models \varphi$ 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 $\mathcal{M}, w \models \psi$
- $\mathcal{M}, w \models K_i\varphi$ if for each $v \in W$, if $w \sim_i v$, then $\mathcal{M}, v \models \varphi$

Plausibility Models

Epistemic-Plausibility Models: $\mathcal{M} = \langle W, \{\sim_i\}_{i\in\mathcal{A}}, \{\preceq_i\}_{i\in\mathcal{A}}, V\rangle$

Truth: $\mathcal{M}, w \models \varphi$ 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 $\mathcal{M}, w \models \psi$
- $\mathcal{M}, w \models K_i\varphi$ if for each $v \in W$, if $w \sim_i v$, then $\mathcal{M}, v \models \varphi$

Plausibility Models

Epistemic-Plausibility Models: $\mathcal{M} = \langle W, \{\sim_i\}_{i\in\mathcal{A}}, \{\preceq_i\}_{i\in\mathcal{A}}, V\rangle$

Plausibility Relation: $\preceq_i \subseteq W \times W$. $w \preceq_i v$ means

"w is at least as plausible as v."

Plausibility Models

Epistemic-Plausibility Models: $\mathcal{M} = \langle W, \{\sim_i\}_{i\in\mathcal{A}}, \{\preceq_i\}_{i\in\mathcal{A}}, V\rangle$

Plausibility Relation: $\preceq_i \subseteq W \times W$. $w \preceq_i v$ means

"w is at least as plausible as v."

Properties of $\preceq_i$: reflexive, transitive, and *well-founded*.

Plausibility Models

Epistemic-Plausibility Models: $\mathcal{M} = \langle W, \{\sim_i\}_{i\in\mathcal{A}}, \{\preceq_i\}_{i\in\mathcal{A}}, V\rangle$

Plausibility Relation: $\preceq_i \subseteq W \times W$. $w \preceq_i v$ means

"w is at least as plausible as v."

Properties of $\preceq_i$: reflexive, transitive, and *well-founded*.

Most Plausible: For $X \subseteq W$, let

$$Min_{\preceq_i}(X) = \{v \in W \mid v \preceq_i w \text{ for all } w \in X\}$$

Plausibility Models

Epistemic-Plausibility Models: $\mathcal{M} = \langle W, \{\sim_i\}_{i\in\mathcal{A}}, \{\preceq_i\}_{i\in\mathcal{A}}, V\rangle$

Plausibility Relation: $\preceq_i \subseteq W \times W$. $w \preceq_i v$ means

"w is at least as plausible as v."

Properties of $\preceq_i$: reflexive, transitive, and *well-founded*.

Most Plausible: For $X \subseteq W$, let

$$Min_{\preceq_i}(X) = \{v \in W \mid v \preceq_i w \text{ for all } w \in X\}$$

Assumptions:

1. *plausibility implies possibility*: if $w \preceq_i v$ then $w \sim_i v$.
2. *locally-connected*: if $w \sim_i v$ then either $w \preceq_i v$ or $v \preceq_i w$.

Plausibility Models

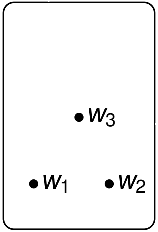
Epistemic-Plausibility Models: $\mathcal{M} = \langle W, \{\sim_i\}_{i\in\mathcal{A}}, \{\preceq_i\}_{i\in\mathcal{A}}, V\rangle$

Truth: $\mathcal{M}, w \models \varphi$ 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 $\mathcal{M}, w \models \psi$
- $\mathcal{M}, w \models K_i\varphi$ if for each $v \in W$, if $w \sim_i v$, then $\mathcal{M}, v \models \varphi$
- $\mathcal{M}, w \models B_i\varphi$ if for each $v \in Min_{\preceq_i}([w]_i)$, $\mathcal{M}, v \models \varphi$
  $[w]_i = \{v \mid w \sim_i v\}$ is the agent's **information cell.**

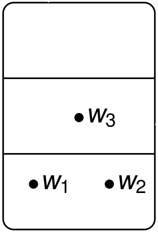
Beliefs via Plausibility

- $W = \{w_1, w_2, w_3\}$



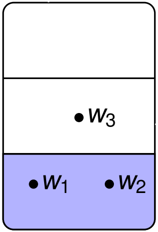
Beliefs via Plausibility

- $W = \{w_1, w_2, w_3\}$
- $w_1 \preceq w_2$ and $w_2 \preceq w_1$ (w_1 and w_2 are equi-plausbile)
- $w_1 \prec w_3$ ($w_1 \preceq w_3$ and $w_3 \not\preceq w_1$)
- $w_2 \prec w_3$ ($w_2 \preceq w_3$ and $w_3 \not\preceq w_2$)

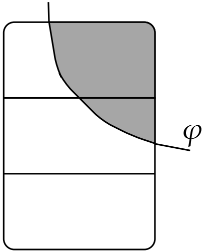


Beliefs via Plausibility

- $W = \{w_1, w_2, w_3\}$
- $w_1 \preceq w_2$ and $w_2 \preceq w_1$ (w_1 and w_2 are equi-plausbile)
- $w_1 \prec w_3$ ($w_1 \preceq w_3$ and $w_3 \not\preceq w_1$)
- $w_2 \prec w_3$ ($w_2 \preceq w_3$ and $w_3 \not\preceq w_2$)
- $\{w_1, w_2\} \subseteq Min_{\preceq}([w_i])$

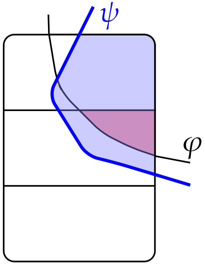


Beliefs via Plausibility



Conditional Belief: $B^{\varphi}\psi$

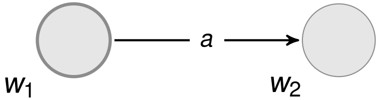
Beliefs via Plausibility



Conditional Belief: $B^{\varphi}\psi$

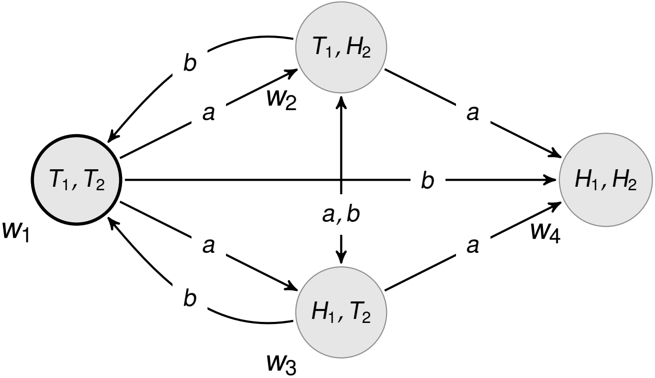
$$Min_{\preceq}(\llbracket\varphi\rrbracket_{\mathcal{M}}) \subseteq \llbracket\psi\rrbracket_{\mathcal{M}}$$

Example

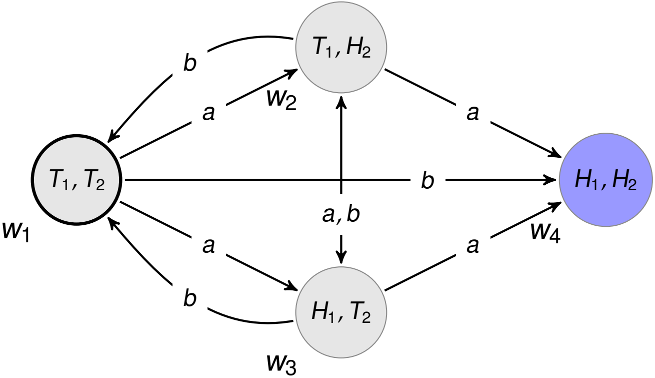


$$w_2 \preceq_a w_1$$

Example

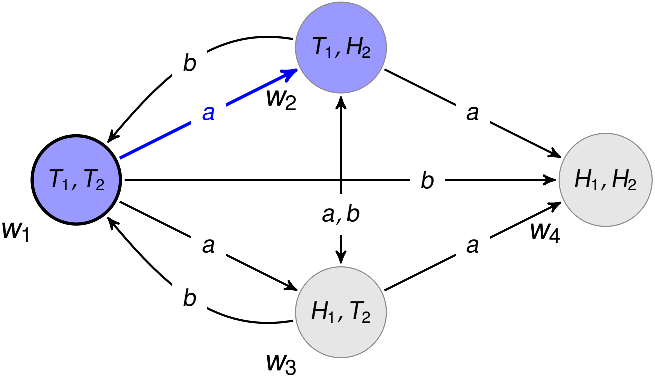


Example



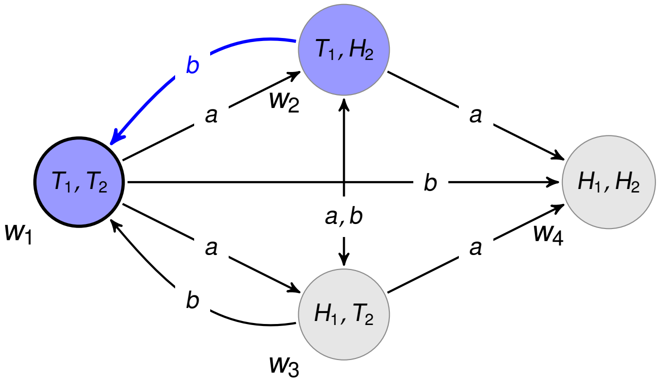
- $w_1 \models B_a(H_1 \wedge H_2) \wedge B_b(H_1 \wedge H_2)$

Example



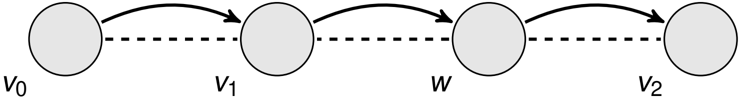
- $w_1 \models B_a(H_1 \wedge H_2) \wedge B_b(H_1 \wedge H_2)$
- $w_1 \models B_a^{T_1} H_2$

Example

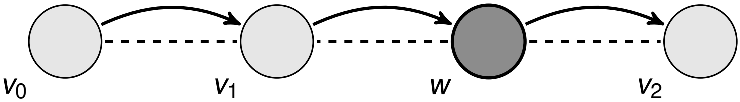


- $w_1 \models B_a(H_1 \wedge H_2) \wedge B_b(H_1 \wedge H_2)$
- $w_1 \models B_a^{T_1} H_2$
- $w_1 \models B_b^{T_1} T_2$

Grades of Doxastic Strength

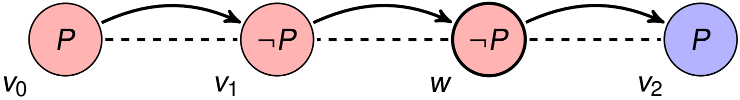


Grades of Doxastic Strength



Suppose that w is the current state.

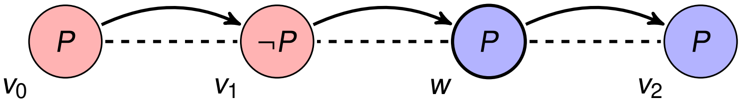
Grades of Doxastic Strength



Suppose that w is the current state.

- **Belief** (BP)

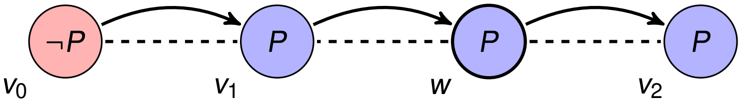
Grades of Doxastic Strength



Suppose that w is the current state.

- **Belief** (BP)
- **Robust Belief** ($[\preceq]P$)

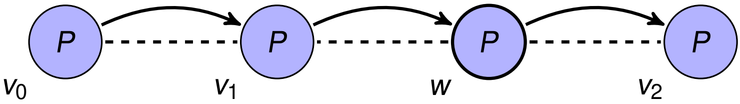
Grades of Doxastic Strength



Suppose that w is the current state.

- **Belief** (BP)
- **Robust Belief** ($[\preceq]P$)
- **Strong Belief** ($B^s P$)

Grades of Doxastic Strength



Suppose that w is the current state.

- **Belief** (BP)
- **Robust Belief** ($[\preceq]P$)
- **Strong Belief** ($B^s P$)
- **Knowledge** (KP)

Is $B\varphi \rightarrow B^{\psi}\varphi$ valid?

Is $B\varphi \rightarrow B^{\psi}\varphi$ valid?

Is $B^{\alpha}\varphi \rightarrow B^{\alpha\wedge\beta}\varphi$ valid?

Is $B\varphi \rightarrow B^{\psi}\varphi$ valid?

Is $B^{\alpha}\varphi \rightarrow B^{\alpha\wedge\beta}\varphi$ valid?

Is $B\varphi \rightarrow B^{\psi}\varphi \vee B^{\neg\psi}\varphi$ valid?

Is $B\varphi \rightarrow B^{\psi}\varphi$ valid?

Is $B^{\alpha}\varphi \rightarrow B^{\alpha\wedge\beta}\varphi$ valid?

Is $B\varphi \rightarrow B^{\psi}\varphi \vee B^{\neg\psi}\varphi$ valid?

Exercise: Prove that B, B^{φ} and B^{s} are definable in the language with K and $[\leq]$ modalities.

$\mathcal{M}, w \models B^{\varphi}\psi$ if for each $v \in Min_{\leq}([w] \cap [\![\varphi]\!])$, $\mathcal{M}, v \models \varphi$
where $[\![\varphi]\!] = \{w \mid \mathcal{M}, w \models \varphi\}$ and $[w] = \{v \mid w \sim v\}$

$\mathcal{M}, w \models B^{\varphi}\psi$ if for each $v \in Min_{\leq}([w] \cap [\![\varphi]\!])$, $\mathcal{M}, v \models \varphi$
where $[\![\varphi]\!] = \{w \mid \mathcal{M}, w \models \varphi\}$ and $[w] = \{v \mid w \sim v\}$

Core Logical Principles:

1. $B^{\varphi}\varphi$
2. $B^{\varphi}\psi \rightarrow B^{\varphi}(\psi \vee \chi)$
3. $(B^{\varphi}\psi_1 \wedge B^{\varphi}\psi_2) \rightarrow B^{\varphi}(\psi_1 \wedge \psi_2)$
4. $(B^{\varphi_1}\psi \wedge B^{\varphi_2}\psi) \rightarrow B^{\varphi_1 \vee \varphi_2}\psi$
5. $(B^{\varphi}\psi \wedge B^{\psi}\varphi) \rightarrow (B^{\varphi}\chi \leftrightarrow B^{\psi}\chi)$

J. Burgess. *Quick completeness proofs for some logics of conditionals*. *Notre Dame Journal of Formal Logic 22*, 76 – 84, 1981.

Types of Beliefs: Logical Characterizations

- $\mathcal{M}, w \models K_i\varphi$ iff $\mathcal{M}, w \models B_i^{\psi}\varphi$ for all ψ
  i knows φ iff *i* continues to believe φ given any new information

Types of Beliefs: Logical Characterizations

- $\mathcal{M}, w \models K_i\varphi$ iff $\mathcal{M}, w \models B_i^{\psi}\varphi$ for all ψ
  i knows φ iff *i* continues to believe φ given any new information

- $\mathcal{M}, w \models [\preceq_i]\varphi$ iff $\mathcal{M}, w \models B_i^{\psi}\varphi$ for all ψ with $\mathcal{M}, w \models \psi$.
  i robustly believes φ iff *i* continues to believe φ given any true formula.

Types of Beliefs: Logical Characterizations

- $\mathcal{M}, w \models K_i\varphi$ iff $\mathcal{M}, w \models B_i^{\psi}\varphi$ for all ψ
  i knows φ iff *i* continues to believe φ given any new information

- $\mathcal{M}, w \models [\preceq_i]\varphi$ iff $\mathcal{M}, w \models B_i^{\psi}\varphi$ for all ψ with $\mathcal{M}, w \models \psi$.
  i robustly believes φ iff *i* continues to believe φ given any true formula.

- $\mathcal{M}, w \models B_i^s\varphi$ iff $\mathcal{M}, w \models B_i\varphi$ and $\mathcal{M}, w \models B_i^{\psi}\varphi$ for all ψ with $\mathcal{M}, w \models \neg K_i(\psi \rightarrow \neg\varphi)$.
  i strongly believes φ iff *i* believes φ and continues to believe φ given any evidence (truthful or not) that is not known to contradict φ.

Additional Axioms

Success: $B_i^{\varphi}\varphi$

Additional Axioms

Success:	$B_i^{\varphi}\varphi$
Knowledge entails belief	$K_i\varphi \rightarrow B_i^{\psi}\varphi$

Additional Axioms

Success:	$B_i^{\varphi}\varphi$	
Knowledge entails belief	$K_i\varphi \rightarrow B_i^{\psi}\varphi$	
Full introspection:	$B_i^{\varphi}\psi \rightarrow K_iB_i^{\varphi}\psi$ and $\neg B_i^{\varphi}\psi \rightarrow K_i\neg B_i^{\varphi}\psi$	

Additional Axioms

Success:	$B_i^{\varphi}\varphi$
Knowledge entails belief	$K_i\varphi \rightarrow B_i^{\psi}\varphi$
Full introspection:	$B_i^{\varphi}\psi \rightarrow K_iB_i^{\varphi}\psi$ and $\neg B_i^{\varphi}\psi \rightarrow K_i\neg B_i^{\varphi}\psi$
Cautious Monotonicity:	$(B_i^{\varphi}\alpha \wedge B_i^{\varphi}\beta) \rightarrow B_i^{\varphi\wedge\beta}\alpha$

Additional Axioms

Success:	$B_i^{\varphi}\varphi$
Knowledge entails belief	$K_i\varphi \rightarrow B_i^{\psi}\varphi$
Full introspection:	$B_i^{\varphi}\psi \rightarrow K_iB_i^{\varphi}\psi$ and $\neg B_i^{\varphi}\psi \rightarrow K_i\neg B_i^{\varphi}\psi$
Cautious Monotonicity:	$(B_i^{\varphi}\alpha \wedge B_i^{\varphi}\beta) \rightarrow B_i^{\varphi\wedge\beta}\alpha$
Rational Monotonicity:	$(B_i^{\varphi}\alpha \wedge \neg B_i^{\varphi}\neg\beta) \rightarrow B_i^{\varphi\wedge\beta}\alpha$

Additional Axioms

Success:	$B_i^{\varphi}\varphi$
Knowledge entails belief	$K_i\varphi \to B_i^{\psi}\varphi$
Full introspection:	$B_i^{\varphi}\psi \to K_i B_i^{\varphi}\psi$ and $\neg B_i^{\varphi}\psi \to K_i \neg B_i^{\varphi}\psi$
Cautious Monotonicity:	$(B_i^{\varphi}\alpha \wedge B_i^{\varphi}\beta) \to B_i^{\varphi\wedge\beta}\alpha$
Rational Monotonicity:	$(B_i^{\varphi}\alpha \wedge \neg B_i^{\varphi}\neg\beta) \to B_i^{\varphi\wedge\beta}\alpha$

Fitch's Paradox

Fitch (1963) derived an unexpected consequence from the thesis, advocated by some anti-realists, that *every truth is knowable*:

Fitch's Paradox

Fitch (1963) derived an unexpected consequence from the thesis, advocated by some anti-realists, that *every truth is knowable*:

(VT) $q \to \Diamond Kq$,

where $\Diamond$ is a *possibility* operator (more on this later).

Fitch's Paradox

Fitch (1963) derived an unexpected consequence from the thesis, advocated by some anti-realists, that *every truth is knowable*:

(VT) $q \to \Diamond Kq$,

where $\Diamond$ is a *possibility* operator (more on this later).

Fitch make two modest assumptions for K, $K\varphi \to \varphi$ (T) and $K(\varphi \wedge \psi) \to (K\varphi \wedge K\psi)$ (M), and two modest assumptions for $\Diamond$:

Fitch's Paradox

Fitch (1963) derived an unexpected consequence from the thesis, advocated by some anti-realists, that *every truth is knowable*:

(VT) $q \to \Diamond Kq$,

where $\Diamond$ is a *possibility* operator (more on this later).

Fitch make two modest assumptions for K, $K\varphi \to \varphi$ (T) and $K(\varphi \wedge \psi) \to (K\varphi \wedge K\psi)$ (M), and two modest assumptions for $\Diamond$:

- $\Diamond$ is the dual of $\Box$ for *necessity*, so $\neg\Diamond\varphi$ follows from $\Box\neg\varphi$.

Fitch's Paradox

Fitch (1963) derived an unexpected consequence from the thesis, advocated by some anti-realists, that *every truth is knowable*:

(VT) $q \rightarrow \Diamond Kq$,

where $\Diamond$ is a *possibility* operator (more on this later).

Fitch make two modest assumptions for K, $K\varphi \rightarrow \varphi$ (T) and $K(\varphi \wedge \psi) \rightarrow (K\varphi \wedge K\psi)$ (M), and two modest assumptions for $\Diamond$:

- $\Diamond$ is the dual of $\Box$ for *necessity*, so $\neg\Diamond\varphi$ follows from $\Box\neg\varphi$.
- $\Box$ obeys the rule of Necessitation: if φ is a theorem, so is $\Box\varphi$.

Fitch's Paradox

For an arbitrary p, consider the following instance of (VT):

(0) $(p \wedge \neg Kp) \rightarrow \Diamond K(p \wedge \neg Kp)$

Fitch's Paradox

For an arbitrary p, consider the following instance of (VT):

(0) $(p \wedge \neg Kp) \rightarrow \Diamond K(p \wedge \neg Kp)$

Here is the proof for Fitch's Paradox:

Fitch's Paradox

For an arbitrary p, consider the following instance of (VT):

(0) $(p \wedge \neg Kp) \rightarrow \Diamond K(p \wedge \neg Kp)$

Here is the proof for Fitch's Paradox:

(1) $K(p \wedge \neg Kp) \rightarrow (Kp \wedge K\neg Kp)$ instance of M axiom

Fitch's Paradox

For an arbitrary p, consider the following instance of (VT):

(0) $(p \wedge \neg Kp) \to \Diamond K(p \wedge \neg Kp)$

Here is the proof for Fitch's Paradox:

(1) $K(p \wedge \neg Kp) \to (Kp \wedge K\neg Kp)$ instance of M axiom

(2) $K\neg Kp \to \neg Kp$ instance of T axiom

Fitch's Paradox

For an arbitrary p, consider the following instance of (VT):

(0) $(p \wedge \neg Kp) \rightarrow \Diamond K(p \wedge \neg Kp)$

Here is the proof for Fitch's Paradox:

(1) $K(p \wedge \neg Kp) \rightarrow (Kp \wedge K\neg Kp)$ instance of M axiom

(2) $K\neg Kp \rightarrow \neg Kp$ instance of T axiom

(3) $K(p \wedge \neg Kp) \rightarrow (Kp \wedge \neg Kp)$ from (1) and (2) by PL

Fitch's Paradox

For an arbitrary p, consider the following instance of (VT):

(0) $(p \wedge \neg Kp) \rightarrow \Diamond K(p \wedge \neg Kp)$

Here is the proof for Fitch's Paradox:

(1) $K(p \wedge \neg Kp) \rightarrow (Kp \wedge K\neg Kp)$ instance of M axiom

(2) $K\neg Kp \rightarrow \neg Kp$ instance of T axiom

(3) $K(p \wedge \neg Kp) \rightarrow (Kp \wedge \neg Kp)$ from (1) and (2) by PL

(4) $\neg K(p \wedge \neg Kp)$ from (3) by PL

Fitch's Paradox

For an arbitrary p, consider the following instance of (VT):

(0) $(p \wedge \neg Kp) \rightarrow \Diamond K(p \wedge \neg Kp)$

Here is the proof for Fitch's Paradox:

(1) $K(p \wedge \neg Kp) \rightarrow (Kp \wedge K\neg Kp)$ instance of M axiom

(2) $K\neg Kp \rightarrow \neg Kp$ instance of T axiom

(3) $K(p \wedge \neg Kp) \rightarrow (Kp \wedge \neg Kp)$ from (1) and (2) by PL

(4) $\neg K(p \wedge \neg Kp)$ from (3) by PL

(5) $\Box\neg K(p \wedge \neg Kp)$ from (4) by $\Box$-Necessitation

Fitch's Paradox

For an arbitrary p, consider the following instance of (VT):

(0) $(p \wedge \neg Kp) \rightarrow \Diamond K(p \wedge \neg Kp)$

Here is the proof for Fitch's Paradox:

(1) $K(p \wedge \neg Kp) \rightarrow (Kp \wedge K\neg Kp)$ instance of M axiom

(2) $K\neg Kp \rightarrow \neg Kp$ instance of T axiom

(3) $K(p \wedge \neg Kp) \rightarrow (Kp \wedge \neg Kp)$ from (1) and (2) by PL

(4) $\neg K(p \wedge \neg Kp)$ from (3) by PL

(5) $\Box\neg K(p \wedge \neg Kp)$ from (4) by $\Box$-Necessitation

(6) $\neg\Diamond K(p \wedge \neg Kp)$ from (5) by $\Box$ - $\Diamond$ Duality

Fitch's Paradox

For an arbitrary p, consider the following instance of (VT):

(0) $(p \wedge \neg Kp) \rightarrow \Diamond K(p \wedge \neg Kp)$

Here is the proof for Fitch's Paradox:

(1) $K(p \wedge \neg Kp) \rightarrow (Kp \wedge K\neg Kp)$ instance of M axiom

(2) $K\neg Kp \rightarrow \neg Kp$ instance of T axiom

(3) $K(p \wedge \neg Kp) \rightarrow (Kp \wedge \neg Kp)$ from (1) and (2) by PL

(4) $\neg K(p \wedge \neg Kp)$ from (3) by PL

(5) $\Box\neg K(p \wedge \neg Kp)$ from (4) by $\Box$-Necessitation

(6) $\neg\Diamond K(p \wedge \neg Kp)$ from (5) by $\Box$ - $\Diamond$ Duality

(7) $\neg(p \wedge \neg Kp)$ from (0) by PL

Fitch's Paradox

For an arbitrary p, consider the following instance of (VT):

(0) $(p \wedge \neg Kp) \rightarrow \Diamond K(p \wedge \neg Kp)$

Here is the proof for Fitch's Paradox:

(1) $K(p \wedge \neg Kp) \rightarrow (Kp \wedge K\neg Kp)$ instance of M axiom

(2) $K\neg Kp \rightarrow \neg Kp$ instance of T axiom

(3) $K(p \wedge \neg Kp) \rightarrow (Kp \wedge \neg Kp)$ from (1) and (2) by PL

(4) $\neg K(p \wedge \neg Kp)$ from (3) by PL

(5) $\Box\neg K(p \wedge \neg Kp)$ from (4) by $\Box$-Necessitation

(6) $\neg\Diamond K(p \wedge \neg Kp)$ from (5) by $\Box$ - $\Diamond$ Duality

(7) $\neg(p \wedge \neg Kp)$ from (0) by PL

(8) $p \rightarrow Kp$ from (7) by classical PL

Fitch's Paradox

For an arbitrary p, consider the following instance of (VT):

(0) $(p \wedge \neg Kp) \to \Diamond K(p \wedge \neg Kp)$

Here is the proof for Fitch's Paradox:

(1) $K(p \wedge \neg Kp) \to (Kp \wedge K\neg Kp)$ instance of M axiom

(2) $K\neg Kp \to \neg Kp$ instance of T axiom

(3) $K(p \wedge \neg Kp) \to (Kp \wedge \neg Kp)$ from (1) and (2) by PL

(4) $\neg K(p \wedge \neg Kp)$ from (3) by PL

(5) $\Box\neg K(p \wedge \neg Kp)$ from (4) by $\Box$-Necessitation

(6) $\neg\Diamond K(p \wedge \neg Kp)$ from (5) by $\Box$ - $\Diamond$ Duality

(7) $\neg(p \wedge \neg Kp)$ from (0) by PL

(8) $p \to Kp$ from (7) by classical PL

Since p was arbitrary, we have shown that *every truth is known*.

The Question

Fitch's Paradox leaves us with **the question**: what must we require in addition to the truth of φ to ensure the knowability of φ?

The Question

Fitch's Paradox leaves us with **the question**: what must we require in addition to the truth of φ to ensure the knowability of φ?

There is a fairly large literature on knowability and related issues. See, e.g.:

J. Salerno. 2009. *New Essays on the Knowability Paradox*, OUP

J. van Benthem. 2004. "What One May Come to Know," *Analysis*.

P. Balbiani et al. 2008. "'Knowable' as 'Known after an Announcement,'" *Review of Symbolic Logic*.

Dynamic Epistemic Logic

The key idea of dynamic epistemic logic is that we can represent changes in agents' epistemic states by *transforming models*.

Dynamic Epistemic Logic

The key idea of dynamic epistemic logic is that we can represent changes in agents' epistemic states by *transforming models*.

In the simplest case, we model an agent's acquisition of knowledge by the elimination of possibilities from an initial epistemic model.

Finding out that φ

$$\mathcal{M} = \langle W, \{\sim_i\}_{i\in\mathcal{A}}, \{\preceq_i\}_{i\in\mathcal{A}}, V\rangle$$

$$\Big\Downarrow$$

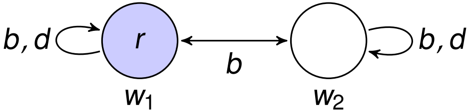
Find out that φ

$$\Big\Downarrow$$

$$\mathcal{M}' = \langle W', \{\sim_i'\}_{i\in\mathcal{A}}, \{\preceq_i'\}_{i\in\mathcal{A}}, V|_{W'}\rangle$$

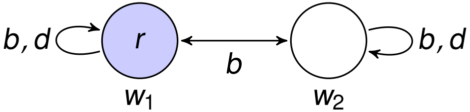
Example: College Park and Amsterdam

Recall the College Park agent who doesn't know whether it's raining in Amsterdam, whose epistemic state is represented by the model:



Example: College Park and Amsterdam

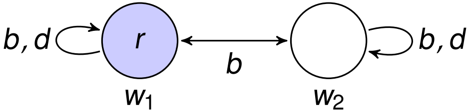
Recall the College Park agent who doesn't know whether it's raining in Amsterdam, whose epistemic state is represented by the model:



What happens when the Amsterdam agent calls the College Park agent on the phone and says, "It's raining in Amsterdam"?

Example: College Park and Amsterdam

Recall the College Park agent who doesn't know whether it's raining in Amsterdam, whose epistemic state is represented by the model:

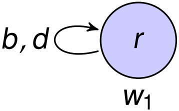


What happens when the Amsterdam agent calls the College Park agent on the phone and says, "It's raining in Amsterdam"?

We model the change in *b*'s epistemic state by eliminating all epistemic possibilities in which it's *not* raining in Amsterdam.

Example: College Park and Amsterdam

Recall the College Park agent who doesn't know whether it's raining in Amsterdam, whose epistemic state is represented by the model:



What happens when the Amsterdam agent calls the College Park agent on the phone and says, "It's raining in Amsterdam"?

We model the change in b's epistemic state by eliminating all epistemic possibilities in which it's *not* raining in Amsterdam.

Model Update

We can easily give a formal definition that captures the idea of knowledge acquisition as the elimination of possibilities.

Model Update

We can easily give a formal definition that captures the idea of knowledge acquisition as the elimination of possibilities.

Given $\mathcal{M} = \langle W, \{R_a \mid a \in \text{Agt}\}, V\rangle$, the *updated model* $\mathcal{M}_{|\varphi}$ is obtained by deleting from $\mathcal{M}$ all worlds in which φ was false.

Model Update

We can easily give a formal definition that captures the idea of knowledge acquisition as the elimination of possibilities.

Given $\mathcal{M} = \langle W, \{R_a \mid a \in \text{Agt}\}, V\rangle$, the *updated model* $\mathcal{M}_{|\varphi}$ is obtained by deleting from $\mathcal{M}$ all worlds in which φ was false.

Formally, $\mathcal{M}_{|\varphi} = \langle W_{|\varphi}, \{R_{a_{|\varphi}} \mid a \in \text{Agt}\}, V_{|\varphi}\rangle$ is the model s.th.:

$$W_{|\varphi} = \{v \in W \mid \mathcal{M}, v \models \varphi\};$$

Model Update

We can easily give a formal definition that captures the idea of knowledge acquisition as the elimination of possibilities.

Given $\mathcal{M} = \langle W, \{R_a \mid a \in \text{Agt}\}, V\rangle$, the *updated model* $\mathcal{M}_{|\varphi}$ is obtained by deleting from $\mathcal{M}$ all worlds in which φ was false.

Formally, $\mathcal{M}_{|\varphi} = \langle W_{|\varphi}, \{R_{a_{|\varphi}} \mid a \in \text{Agt}\}, V_{|\varphi}\rangle$ is the model s.th.:

$$W_{|\varphi} = \{v \in W \mid \mathcal{M}, v \models \varphi\};$$

$R_{a_{|\varphi}}$ is the restriction of R_a to $W_{|\varphi}$;

Model Update

We can easily give a formal definition that captures the idea of knowledge acquisition as the elimination of possibilities.

Given $\mathcal{M} = \langle W, \{R_a \mid a \in \text{Agt}\}, V\rangle$, the *updated model* $\mathcal{M}_{|\varphi}$ is obtained by deleting from $\mathcal{M}$ all worlds in which φ was false.

Formally, $\mathcal{M}_{|\varphi} = \langle W_{|\varphi}, \{R_{a_{|\varphi}} \mid a \in \text{Agt}\}, V_{|\varphi}\rangle$ is the model s.th.:

$W_{|\varphi} = \{v \in W \mid \mathcal{M}, v \models \varphi\}$;

$R_{a_{|\varphi}}$ is the restriction of R_a to $W_{|\varphi}$;

$V_{|\varphi}(p)$ is the intersection of $V(p)$ and $W_{|\varphi}$.

Model Update

We can easily give a formal definition that captures the idea of knowledge acquisition as the elimination of possibilities.

Given $\mathcal{M} = \langle W, \{R_a \mid a \in \mathsf{Agt}\}, V\rangle$, the *updated model* $\mathcal{M}_{|\varphi}$ is obtained by deleting from $\mathcal{M}$ all worlds in which φ was false.

Formally, $\mathcal{M}_{|\varphi} = \langle W_{|\varphi}, \{R_{a_{|\varphi}} \mid a \in \mathsf{Agt}\}, V_{|\varphi}\rangle$ is the model s.th.:

$W_{|\varphi} = \{v \in W \mid \mathcal{M}, v \models \varphi\}$;

$R_{a_{|\varphi}}$ is the restriction of R_a to $W_{|\varphi}$;

$V_{|\varphi}(p)$ is the intersection of $V(p)$ and $W_{|\varphi}$.

In the single-agent case, this models the agent learning φ. In the multi-agent case, this models all agents *publicly* learning φ.

Public Announcement Logic

One of the **big ideas** of dynamic epistemic logic is to add to our formal language operators that can describe the kinds of model updates that we just saw for the College Park and Amsterdam example.

Public Announcement Logic

One of the **big ideas** of dynamic epistemic logic is to add to our formal language operators that can describe the kinds of model updates that we just saw for the College Park and Amsterdam example.

The language of Public Announcement Logic (PAL) is given by:

$$\varphi ::= p \mid \neg\varphi \mid (\varphi \wedge \varphi) \mid K_a\varphi \mid [!\varphi]\varphi$$

Public Announcement Logic

One of the **big ideas** of dynamic epistemic logic is to add to our formal language operators that can describe the kinds of model updates that we just saw for the College Park and Amsterdam example.

The language of Public Announcement Logic (PAL) is given by:

$$\varphi ::= p \mid \neg\varphi \mid (\varphi \wedge \varphi) \mid K_a\varphi \mid [!\varphi]\varphi$$

Read $[!\varphi]\psi$ as "after (every) true announcement of φ, ψ."

Public Announcement Logic

One of the **big ideas** of dynamic epistemic logic is to add to our formal language operators that can describe the kinds of model updates that we just saw for the College Park and Amsterdam example.

The language of Public Announcement Logic (PAL) is given by:

$$\varphi ::= p \mid \neg\varphi \mid (\varphi \wedge \varphi) \mid K_a\varphi \mid [!\varphi]\varphi$$

Read $[!\varphi]\psi$ as "after (every) true announcement of φ, ψ."

Read $\langle !\varphi\rangle\psi := \neg[!\varphi]\neg\psi$ as "after a true announcement of φ, ψ."

Public Announcement Logic

Read $[!\varphi]\psi$ as "after (every) true announcement of φ, ψ."

Read $\langle !\varphi\rangle\psi := \neg[!\varphi]\neg\psi$ as "after a true announcement of φ, ψ."

The truth clause for the dynamic operator $[!\varphi]$ is:

Public Announcement Logic

Read $[!\varphi]\psi$ as "after (every) true announcement of φ, ψ."

Read $\langle !\varphi\rangle\psi := \neg[!\varphi]\neg\psi$ as "after a true announcement of φ, ψ."

The truth clause for the dynamic operator $[!\varphi]$ is:

- $\mathcal{M}, w \models [!\varphi]\psi$ iff $\mathcal{M}, w \models \varphi$ *implies* $\mathcal{M}_{|\varphi}, w \models \psi$.

Public Announcement Logic

Read $[!\varphi]\psi$ as "after (every) true announcement of φ, ψ."

Read $\langle !\varphi\rangle\psi := \neg[!\varphi]\neg\psi$ as "after a true announcement of φ, ψ."

The truth clause for the dynamic operator $[!\varphi]$ is:

- $\mathcal{M}, w \models [!\varphi]\psi$ iff $\mathcal{M}, w \models \varphi$ *implies* $\mathcal{M}_{|\varphi}, w \models \psi$.

So if φ is false, $[!\varphi]\psi$ is vacuously true.

Public Announcement Logic

Read $[!\varphi]\psi$ as "after (every) true announcement of φ, ψ."

Read $\langle!\varphi\rangle\psi := \neg[!\varphi]\neg\psi$ as "after a true announcement of φ, ψ."

The truth clause for the dynamic operator $[!\varphi]$ is:

- $\mathcal{M}, w \models [!\varphi]\psi$ iff $\mathcal{M}, w \models \varphi$ *implies* $\mathcal{M}_{|\varphi}, w \models \psi$.

So if φ is false, $[!\varphi]\psi$ is vacuously true. Here is the $\langle!\varphi\rangle$ clause:

Public Announcement Logic

Read $[!\varphi]\psi$ as "after (every) true announcement of φ, ψ."

Read $\langle !\varphi\rangle\psi := \neg[!\varphi]\neg\psi$ as "after a true announcement of φ, ψ."

The truth clause for the dynamic operator $[!\varphi]$ is:

- $\mathcal{M}, w \models [!\varphi]\psi$ iff $\mathcal{M}, w \models \varphi$ *implies* $\mathcal{M}_{|\varphi}, w \models \psi$.

So if φ is false, $[!\varphi]\psi$ is vacuously true. Here is the $\langle !\varphi\rangle$ clause:

- $\mathcal{M}, w \models \langle !\varphi\rangle\psi$ iff $\mathcal{M}, w \models \varphi$ *and* $\mathcal{M}_{|\varphi}, w \models \psi$.

Public Announcement Logic

Read $[!\varphi]\psi$ as "after (every) true announcement of φ, ψ."

Read $\langle !\varphi \rangle\psi := \neg[!\varphi]\neg\psi$ as "after a true announcement of φ, ψ."

The truth clause for the dynamic operator $[!\varphi]$ is:

- $\mathcal{M}, w \models [!\varphi]\psi$ iff $\mathcal{M}, w \models \varphi$ *implies* $\mathcal{M}_{|\varphi}, w \models \psi$.

So if φ is false, $[!\varphi]\psi$ is vacuously true. Here is the $\langle !\varphi \rangle$ clause:

- $\mathcal{M}, w \models \langle !\varphi \rangle\psi$ iff $\mathcal{M}, w \models \varphi$ *and* $\mathcal{M}_{|\varphi}, w \models \psi$.

Big Idea: we evaluate $[!\varphi]\psi$ and $\langle !\varphi \rangle\psi$ not by looking at *other worlds in the same model*, but rather by looking at a new model.

Public Announcement Logic

Suppose $\mathcal{M} = \langle W, \{\sim_i\}_{i\in\mathcal{A}}, \{\preceq_i\}_{i\in\mathcal{A}}, V\rangle$ is a multi-agent Kripke Model

$$\mathcal{M}, w \models [\psi]\varphi \text{ iff } \mathcal{M}, w \models \psi \text{ implies } \mathcal{M}|_\psi, w \models \varphi$$

where $\mathcal{M}|_\psi = \langle W', \{\sim_i'\}_{i\in\mathcal{A}}, \{\preceq_i'\}_{i\in\mathcal{A}}, V'\rangle$ with

- $W' = W \cap \{w \mid \mathcal{M}, w \models \psi\}$
- For each i, $\sim_i' = \sim_i \cap (W' \times W')$
- For each i, $\preceq_i' = \preceq_i \cap (W' \times W')$
- for all $p \in \text{At}$, $V'(p) = V(p) \cap W'$

Public Announcement Logic

$$[\psi]p \quad \leftrightarrow \quad (\psi \rightarrow p)$$

Public Announcement Logic

$$\begin{array}{rcl} [\psi]p & \leftrightarrow & (\psi \rightarrow p) \\ [\psi]\neg\varphi & \leftrightarrow & (\psi \rightarrow \neg[\psi]\varphi) \end{array}$$

Public Announcement Logic

$$
\begin{aligned}
[\psi]p &\leftrightarrow (\psi \rightarrow p) \\
[\psi]\neg\varphi &\leftrightarrow (\psi \rightarrow \neg[\psi]\varphi) \\
[\psi](\varphi \wedge \chi) &\leftrightarrow ([\psi]\varphi \wedge [\psi]\chi)
\end{aligned}
$$

Public Announcement Logic

$$
\begin{array}{rcl}
[\psi]p & \leftrightarrow & (\psi \rightarrow p) \\
[\psi]\neg\varphi & \leftrightarrow & (\psi \rightarrow \neg[\psi]\varphi) \\
[\psi](\varphi \wedge \chi) & \leftrightarrow & ([\psi]\varphi \wedge [\psi]\chi) \\
[\psi][\varphi]\chi & \leftrightarrow & [\psi \wedge [\psi]\varphi]\chi
\end{array}
$$

Public Announcement Logic

$$
\begin{array}{rcl}
[\psi]p & \leftrightarrow & (\psi \rightarrow p) \\
[\psi]\neg\varphi & \leftrightarrow & (\psi \rightarrow \neg[\psi]\varphi) \\
[\psi](\varphi \wedge \chi) & \leftrightarrow & ([\psi]\varphi \wedge [\psi]\chi) \\
[\psi][\varphi]\chi & \leftrightarrow & [\psi \wedge [\psi]\varphi]\chi \\
[\psi]K_i\varphi & \leftrightarrow & (\psi \rightarrow K_i(\psi \rightarrow [\psi]\varphi))
\end{array}
$$

Public Announcement Logic

$$
\begin{array}{rcl}
[\psi]p & \leftrightarrow & (\psi \rightarrow p) \\
[\psi]\neg\varphi & \leftrightarrow & (\psi \rightarrow \neg[\psi]\varphi) \\
[\psi](\varphi \wedge \chi) & \leftrightarrow & ([\psi]\varphi \wedge [\psi]\chi) \\
[\psi][\varphi]\chi & \leftrightarrow & [\psi \wedge [\psi]\varphi]\chi \\
[\psi]K_i\varphi & \leftrightarrow & (\psi \rightarrow K_i(\psi \rightarrow [\psi]\varphi))
\end{array}
$$

Theorem Every formula of Public Announcement Logic is equivalent to a formula of Epistemic Logic.

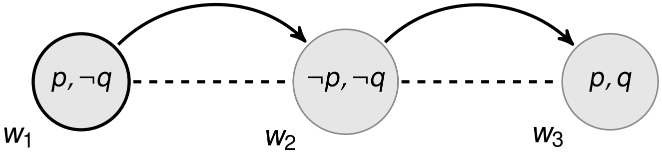
- $[q]Kq$

- $[q]Kq$

- $Kp \rightarrow [q]Kp$

- $[q]Kq$
- $Kp \rightarrow [q]Kp$
- $B\varphi \rightarrow [\psi]B\varphi$

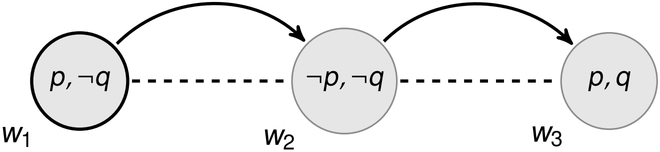
- $[q]Kq$
- $Kp \rightarrow [q]Kp$
- $B\varphi \rightarrow [\psi]B\varphi$



- $[q]Kq$

- $Kp \rightarrow [q]Kp$

- $B\varphi \rightarrow [\psi]B\varphi$



- $[\varphi]\varphi$

Public Announcement vs. Conditional Belief

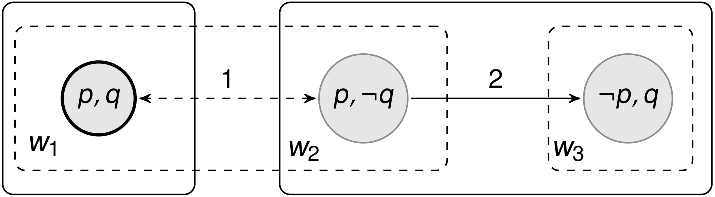
Are $[\varphi]B\psi$ and $B^{\varphi}\psi$ different?

Public Announcement vs. Conditional Belief

Are $[\varphi]B\psi$ and $B^{\varphi}\psi$ different? Yes!

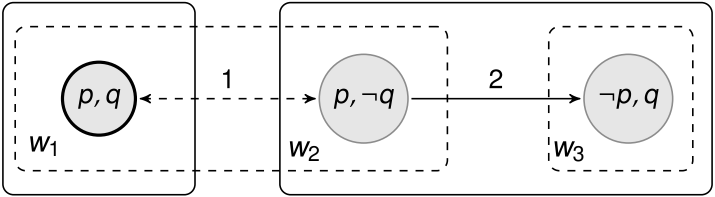
Public Announcement vs. Conditional Belief

Are $[\varphi]B\psi$ and $B^{\varphi}\psi$ different? Yes!



Public Announcement vs. Conditional Belief

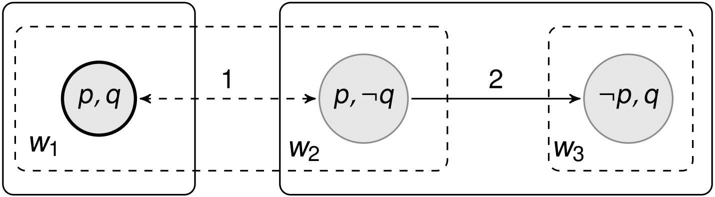
Are $[\varphi]B\psi$ and $B^{\varphi}\psi$ different? Yes!



- $w_1 \models B_1 B_2 q$

Public Announcement vs. Conditional Belief

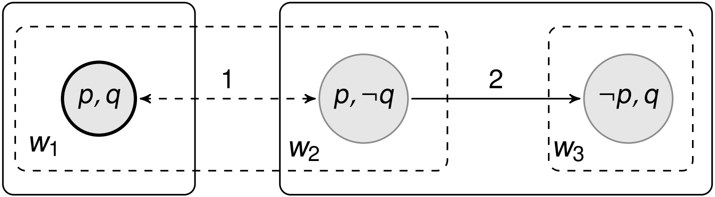
Are $[\varphi]B\psi$ and $B^{\varphi}\psi$ different? Yes!



- $w_1 \models B_1 B_2 q$
- $w_1 \models B_1^p B_2 q$

Public Announcement vs. Conditional Belief

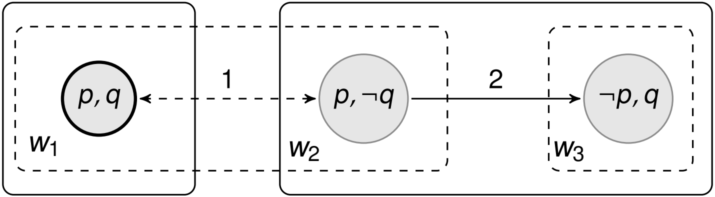
Are $[\varphi]B\psi$ and $B^{\varphi}\psi$ different? Yes!



- $w_1 \models B_1 B_2 q$
- $w_1 \models B_1^p B_2 q$
- $w_1 \models [p]\neg B_1 B_2 q$

Public Announcement vs. Conditional Belief

Are $[\varphi]B\psi$ and $B^{\varphi}\psi$ different? Yes!



- $w_1 \models B_1 B_2 q$
- $w_1 \models B_1^p B_2 q$
- $w_1 \models [p]\neg B_1 B_2 q$
- More generally, $B_i^p(p \wedge \neg K_i p)$ is satisfiable but $[p]B_i(p \wedge \neg K_i p)$ is not.

The Logic of Public Observation

- $[\varphi]K\psi \leftrightarrow (\varphi \rightarrow K(\varphi \rightarrow [\varphi]\psi))$

The Logic of Public Observation

- $[\varphi]K\psi \leftrightarrow (\varphi \rightarrow K(\varphi \rightarrow [\varphi]\psi))$

- $[\varphi][\preceq]\psi \leftrightarrow (\varphi \rightarrow [\preceq](\varphi \rightarrow [\varphi]\psi))$

The Logic of Public Observation

- $[\varphi]K\psi \leftrightarrow (\varphi \rightarrow K(\varphi \rightarrow [\varphi]\psi))$

- $[\varphi][\preceq]\psi \leftrightarrow (\varphi \rightarrow [\preceq](\varphi \rightarrow [\varphi]\psi))$

- **Belief**: $[\varphi]B\psi \not\leftrightarrow (\varphi \rightarrow B(\varphi \rightarrow [\varphi]\psi))$

The Logic of Public Observation

- $[\varphi]K\psi \leftrightarrow (\varphi \rightarrow K(\varphi \rightarrow [\varphi]\psi))$

- $[\varphi][\preceq]\psi \leftrightarrow (\varphi \rightarrow [\preceq](\varphi \rightarrow [\varphi]\psi))$

- **Belief**: $[\varphi]B\psi \not\leftrightarrow (\varphi \rightarrow B(\varphi \rightarrow [\varphi]\psi))$

  $[\varphi]B\psi \leftrightarrow (\varphi \rightarrow B^{\varphi}[\varphi]\psi)$

The Logic of Public Observation

- $[\varphi]K\psi \leftrightarrow (\varphi \rightarrow K(\varphi \rightarrow [\varphi]\psi))$

- $[\varphi][\preceq]\psi \leftrightarrow (\varphi \rightarrow [\preceq](\varphi \rightarrow [\varphi]\psi))$

- **Belief**: $[\varphi]B\psi \not\leftrightarrow (\varphi \rightarrow B(\varphi \rightarrow [\varphi]\psi))$

  $[\varphi]B\psi \leftrightarrow (\varphi \rightarrow B^{\varphi}[\varphi]\psi)$
  $[\varphi]B^{\alpha}\psi \leftrightarrow (\varphi \rightarrow B^{\varphi \wedge [\varphi]\alpha}[\varphi]\psi)$

Group Knowledge

Example (1)

Suppose there are two friends Ann and Bob are on a bus separated by a crowd.

Example (1)

Suppose there are two friends Ann and Bob are on a bus separated by a crowd. Before the bus comes to the next stop a mutual friend from outside the bus yells “get off at the next stop to get a drink?”.

Example (1)

Suppose there are two friends Ann and Bob are on a bus separated by a crowd. Before the bus comes to the next stop a mutual friend from outside the bus yells “get off at the next stop to get a drink?”.

Say Ann is standing near the front door and Bob near the back door.

Example (1)

Suppose there are two friends Ann and Bob are on a bus separated by a crowd. Before the bus comes to the next stop a mutual friend from outside the bus yells “get off at the next stop to get a drink?”.

Say Ann is standing near the front door and Bob near the back door. When the bus comes to a stop, will they get off?

Example (1)

Suppose there are two friends Ann and Bob are on a bus separated by a crowd. Before the bus comes to the next stop a mutual friend from outside the bus yells "get off at the next stop to get a drink?".

Say Ann is standing near the front door and Bob near the back door. When the bus comes to a stop, will they get off?

D. Lewis. *Convention*. 1969.

M. Chwe. *Rational Ritual*. 2001.

"*Common Knowledge*" is informally described as what any fool would know, given a certain situation: It encompasses what is relevant, agreed upon, established by precedent, assumed, being attended to, salient, or in the conversational record.

"*Common Knowledge*" is informally described as what any fool would know, given a certain situation: It encompasses what is relevant, agreed upon, established by precedent, assumed, being attended to, salient, or in the conversational record.

It is not Common Knowledge who "defined" Common Knowledge!

The first formal definition of common knowledge?

M. Friedell. *On the Structure of Shared Awareness*. Behavioral Science (1969).

R. Aumann. *Agreeing to Disagree*. Annals of Statistics (1976).

The first formal definition of common knowledge?

M. Friedell. *On the Structure of Shared Awareness*. Behavioral Science (1969).

R. Aumann. *Agreeing to Disagree*. Annals of Statistics (1976).

The first rigorous analysis of common knowledge

D. Lewis. *Convention, A Philosophical Study*. 1969.

The first formal definition of common knowledge?

M. Friedell. *On the Structure of Shared Awareness.* Behavioral Science (1969).

R. Aumann. *Agreeing to Disagree.* Annals of Statistics (1976).

The first rigorous analysis of common knowledge

D. Lewis. *Convention, A Philosophical Study.* 1969.

Fixed-point definition: $\gamma :=$ *i* and *j* know that (φ and γ)

G. Harman. *Review of* Linguistic Behavior. Language (1977).

J. Barwise. *Three views of Common Knowledge.* TARK (1987).

The first formal definition of common knowledge?

M. Friedell. *On the Structure of Shared Awareness*. Behavioral Science (1969).

R. Aumann. *Agreeing to Disagree*. Annals of Statistics (1976).

The first rigorous analysis of common knowledge

D. Lewis. *Convention, A Philosophical Study*. 1969.

Fixed-point definition: $\gamma :=$ *i* and *j* know that (φ and γ)

G. Harman. *Review of* Linguistic Behavior. Language (1977).

J. Barwise. *Three views of Common Knowledge*. TARK (1987).

Shared situation: There is a *shared situation s* such that (1) *s* entails φ, (2) *s* entails everyone knows φ, plus other conditions

H. Clark and C. Marshall. *Definite Reference and Mutual Knowledge*. 1981.

M. Gilbert. *On Social Facts*. Princeton University Press (1989).

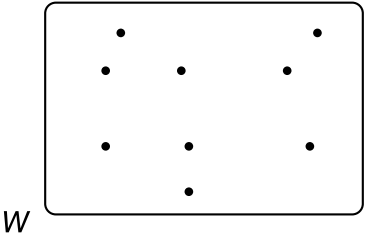
P. Vanderschraaf and G. Sillari. *"Common Knowledge", The Stanford Encyclopedia of Philosophy (2009)*.
http://plato.stanford.edu/entries/common-knowledge/.

The "Standard" Account

R. Aumann. *Agreeing to Disagree*. Annals of Statistics (1976).

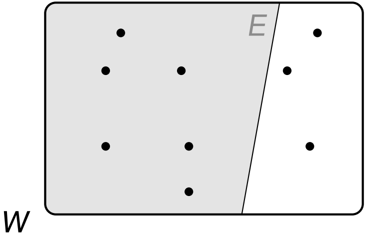
R. Fagin, J. Halpern, Y. Moses and M. Vardi. *Reasoning about Knowledge*. MIT Press, 1995.

The “Standard” Account



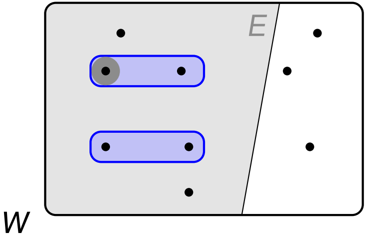
W is a set of **states** or **worlds**.

The “Standard” Account



An **event**/**proposition** is any (definable) subset $E \subseteq W$

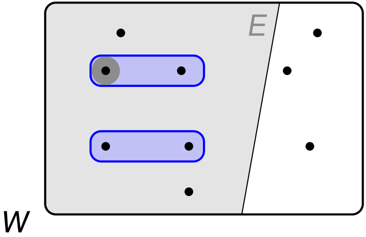
The “Standard” Account



At each state, agents are assigned a set of states they *consider possible* (according to their information).

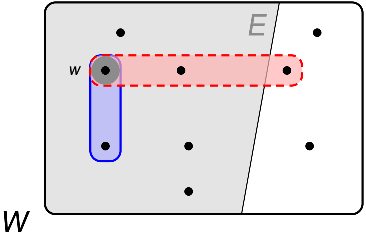
The information may be (in)correct, partitional, ...

The "Standard" Account



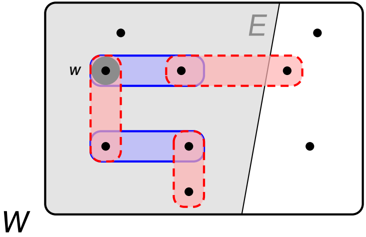
Knowledge Function: $K_i : \wp(W) \rightarrow \wp(W)$ where $K_i(E) = \{w \mid R_i(w) \subseteq E\}$

The “Standard” Account



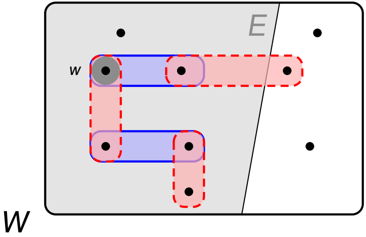
$w \in K_A(E)$ and $w \notin K_B(E)$

The “Standard” Account



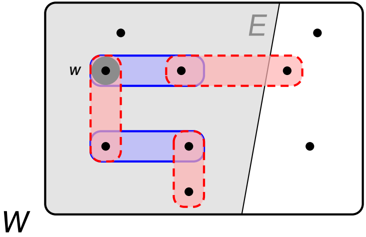
The model also describes the agents’ **higher-order knowledge/beliefs**

The "Standard" Account



Everyone Knows: $K(E) = \bigcap_{i \in \mathcal{A}} K_i(E)$, $K^0(E) = E$, $K^m(E) = K(K^{m-1}(E))$

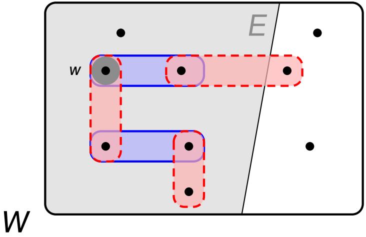
The “Standard” Account



Common Knowledge: $C : \wp(W) \to \wp(W)$ with

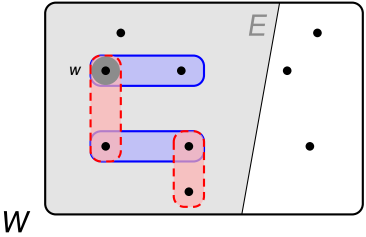
$$C(E) = \bigcap_{m \geq 0} K^m(E)$$

The "Standard" Account



$w \in K(E)$ $\qquad$ $w \notin C(E)$

The “Standard” Account



$w \in C(E)$

Fact. For all $i \in \mathcal{A}$ and $E \subseteq W$, $K_iC(E) = C(E)$.

Fact. For all $i \in \mathcal{A}$ and $E \subseteq W$, $K_i C(E) = C(E)$.

Suppose you are told "Ann and Bob are going together,"' and respond "sure, that's common knowledge." What you mean is not only that everyone knows this, but also that the announcement is pointless, occasions no surprise, reveals nothing new; in effect, that the situation after the announcement does not differ from that before. ...the event "Ann and Bob are going together" — call it E — is common knowledge if and only if some event — call it F — happened that entails E and also entails all players' knowing F (like all players met Ann and Bob at an intimate party). *(Aumann, pg. 271, footnote 8)*

Fact. For all $i \in \mathcal{A}$ and $E \subseteq W$, $K_i C(E) = C(E)$.

An event F is **self-evident** if $K_i(F) = F$ for all $i \in \mathcal{A}$.

Fact. An event E is commonly known iff some self-evident event that entails E obtains.

Fact. For all $i \in \mathcal{A}$ and $E \subseteq W$, $K_iC(E) = C(E)$.

An event F is **self-evident** if $K_i(F) = F$ for all $i \in \mathcal{A}$.

Fact. An event E is commonly known iff some self-evident event that entails E obtains.

Fact. $w \in C(E)$ if every finite path starting at w ends in a state in E

The following axiomatize common knowledge:

- $C(\varphi \to \psi) \to (C\varphi \to C\psi)$
- $C\varphi \to (\varphi \wedge EC\varphi)$ (Fixed-Point)
- $C(\varphi \to E\varphi) \to (\varphi \to C\varphi)$ (Induction)

An Example

Two players Ann and Bob are told that the following will happen. Some positive integer n will be chosen and *one* of n, $n + 1$ will be written on Ann's forehead, the other on Bob's. Each will be able to see the other's forehead, but not his/her own.

An Example

Two players Ann and Bob are told that the following will happen. Some positive integer n will be chosen and *one* of n, $n + 1$ will be written on Ann's forehead, the other on Bob's. Each will be able to see the other's forehead, but not his/her own.

Suppose the number are (2,3).

An Example

Two players Ann and Bob are told that the following will happen. Some positive integer n will be chosen and *one* of n, $n+1$ will be written on Ann's forehead, the other on Bob's. Each will be able to see the other's forehead, but not his/her own.

Suppose the number are (2,3).

Do the agents know there numbers are less than 1000?

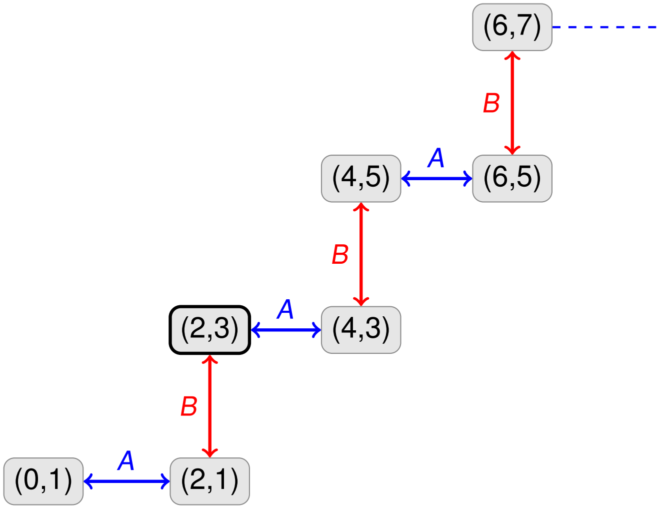
An Example

Two players Ann and Bob are told that the following will happen. Some positive integer n will be chosen and *one* of n, $n+1$ will be written on Ann's forehead, the other on Bob's. Each will be able to see the other's forehead, but not his/her own.

Suppose the number are (2,3).

Do the agents know there numbers are less than 1000?

Is it common knowledge that their numbers are less than 1000?


(6,7)
B
(4,5)
A
(6,5)
B
(2,3)
A
(4,3)
B
(0,1)
A
(2,1)

The Fixed-Point Definition

The Fixed-Point Definition

$$f_E(X) = K(E \cap X) = \bigcap_{i \in \mathcal{A}} K_i(E \cap X)$$

The Fixed-Point Definition

$$f_E(X) = K(E \cap X) = \bigcap_{i \in \mathcal{A}} K_i(E \cap X)$$

- $C(E)$ is a fixed point of f_E: $f_E(C(E))$

The Fixed-Point Definition

$$f_E(X) = K(E \cap X) = \bigcap_{i \in \mathcal{A}} K_i(E \cap X)$$

- $C(E)$ is a fixed point of f_E: $f_E(C(E)) = K(E \cap C(E))$

The Fixed-Point Definition

$$f_E(X) = K(E \cap X) = \bigcap_{i \in \mathcal{A}} K_i(E \cap X)$$

- $C(E)$ is a fixed point of f_E: $f_E(C(E)) = K(E \cap C(E)) = K(C(E)) = \bigcap_{i \in \mathcal{A}} K_i(C(E))$

The Fixed-Point Definition

$f_E(X) = K(E \cap X) = \bigcap_{i \in \mathcal{A}} K_i(E \cap X)$

- $C(E)$ is a fixed point of f_E: $f_E(C(E)) = K(E \cap C(E)) = K(C(E)) = \bigcap_{i \in \mathcal{A}} K_i(C(E)) = \bigcap_{i \in \mathcal{A}} C(E) = C(E)$

The Fixed-Point Definition

$$f_E(X) = K(E \cap X) = \bigcap_{i \in \mathcal{A}} K_i(E \cap X)$$

- $C(E)$ is a fixed point of f_E: $f_E(C(E)) = K(E \cap C(E)) = K(C(E)) = \bigcap_{i \in \mathcal{A}} K_i(C(E)) = \bigcap_{i \in \mathcal{A}} C(E) = C(E)$
- The are other fixed points of f_E: $f_E(\bot) = \bot$

The Fixed-Point Definition

$$f_E(X) = K(E \cap X) = \bigcap_{i \in \mathcal{A}} K_i(E \cap X)$$

- $C(E)$ is a fixed point of f_E: $f_E(C(E)) = K(E \cap C(E)) = K(C(E)) = \bigcap_{i \in \mathcal{A}} K_i(C(E)) = \bigcap_{i \in \mathcal{A}} C(E) = C(E)$
- The are other fixed points of f_E: $f_E(\bot) = \bot$
- f_E is monotonic:

The Fixed-Point Definition

$$f_E(X) = K(E \cap X) = \bigcap_{i\in\mathcal{A}} K_i(E \cap X)$$

- $C(E)$ is a fixed point of f_E: $f_E(C(E)) = K(E \cap C(E)) = K(C(E)) = \bigcap_{i\in\mathcal{A}} K_i(C(E)) = \bigcap_{i\in\mathcal{A}} C(E) = C(E)$
- The are other fixed points of f_E: $f_E(\perp) = \perp$
- f_E is monotonic: $A \subseteq B$ implies $E \cap A \subseteq E \cap B$.

The Fixed-Point Definition

$f_E(X) = K(E \cap X) = \bigcap_{i \in \mathcal{A}} K_i(E \cap X)$

- $C(E)$ is a fixed point of f_E: $f_E(C(E)) = K(E \cap C(E)) = K(C(E)) = \bigcap_{i \in \mathcal{A}} K_i(C(E)) = \bigcap_{i \in \mathcal{A}} C(E) = C(E)$
- The are other fixed points of f_E: $f_E(\perp) = \perp$
- f_E is monotonic: $A \subseteq B$ implies $E \cap A \subseteq E \cap B$. Then $f_E(E \cap A) = K(E \cap A) \subseteq K(E \cap B) = f_E(E \cap B)$

The Fixed-Point Definition

$$f_E(X) = K(E \cap X) = \bigcap_{i \in \mathcal{A}} K_i(E \cap X)$$

- $C(E)$ is a fixed point of f_E: $f_E(C(E)) = K(E \cap C(E)) = K(C(E)) = \bigcap_{i \in \mathcal{A}} K_i(C(E)) = \bigcap_{i \in \mathcal{A}} C(E) = C(E)$
- The are other fixed points of f_E: $f_E(\bot) = \bot$
- f_E is monotonic: $A \subseteq B$ implies $E \cap A \subseteq E \cap B$. Then $f_E(E \cap A) = K(E \cap A) \subseteq K(E \cap B) = f_E(E \cap B)$
- (Tarski) Every monotone operator has a greatest (and least) fixed point

The Fixed-Point Definition

$$f_E(X) = K(E \cap X) = \bigcap_{i \in \mathcal{A}} K_i(E \cap X)$$

- $C(E)$ is a fixed point of f_E: $f_E(C(E)) = K(E \cap C(E)) = K(C(E)) = \bigcap_{i \in \mathcal{A}} K_i(C(E)) = \bigcap_{i \in \mathcal{A}} C(E) = C(E)$
- The are other fixed points of f_E: $f_E(\perp) = \perp$
- f_E is monotonic: $A \subseteq B$ implies $E \cap A \subseteq E \cap B$. Then $f_E(E \cap A) = K(E \cap A) \subseteq K(E \cap B) = f_E(E \cap B)$
- (Tarski) Every monotone operator has a greatest (and least) fixed point
- Let $K^*(E)$ be the greatest fixed point of f_E.

The Fixed-Point Definition

$$f_E(X) = K(E \cap X) = \bigcap_{i \in \mathcal{A}} K_i(E \cap X)$$

- $C(E)$ is a fixed point of f_E: $f_E(C(E)) = K(E \cap C(E)) = K(C(E)) = \bigcap_{i \in \mathcal{A}} K_i(C(E)) = \bigcap_{i \in \mathcal{A}} C(E) = C(E)$
- The are other fixed points of f_E: $f_E(\bot) = \bot$
- f_E is monotonic: $A \subseteq B$ implies $E \cap A \subseteq E \cap B$. Then $f_E(E \cap A) = K(E \cap A) \subseteq K(E \cap B) = f_E(E \cap B)$
- (Tarski) Every monotone operator has a greatest (and least) fixed point
- Let $K^*(E)$ be the greatest fixed point of f_E.
- **Fact.** $K^*(E) = C(E)$.

The Fixed-Point Definition

Separating the fixed-point/iteration definition of common knowledge/belief:

J. Barwise. *Three views of Common Knowledge*. TARK (1987).

J. van Benthem and D. Saraenac. *The Geometry of Knowledge*. Aspects of Universal Logic (2004).

A. Heifetz. *Iterative and Fixed Point Common Belief*. Journal of Philosophical Logic (1999).

Some Issues

Some Issues

- What *does* a group know/believe/accept? vs. what *can* a group (come to) know/believe/accept?

Some Issues

- What *does* a group know/believe/accept? vs. what *can* a group (come to) know/believe/accept?

C. List. *Group knowledge and group rationality: a judgment aggregation perspective*. Episteme (2008).

Some Issues

- What *does* a group know/believe/accept? vs. what *can* a group (come to) know/believe/accept?

C. List. *Group knowledge and group rationality: a judgment aggregation perspective*. Episteme (2008).

- Other "group informational attitudes": distributed knowledge, common belief, ...

Some Issues

- What *does* a group know/believe/accept? vs. what *can* a group (come to) know/believe/accept?

C. List. *Group knowledge and group rationality: a judgment aggregation perspective*. Episteme (2008).

- Other "group informational attitudes": distributed knowledge, common belief, ...

- Where does common knowledge come from?

Some Issues

✓ What *does* a group know/believe/accept? vs. what *can* a group (come to) know/believe/accept?

C. List. *Group knowledge and group rationality: a judgment aggregation perspective*. Episteme (2008).

- Other “group informational attitudes”: distributed knowledge, common belief, …

- Where does common knowledge come from?

Distributed Knowledge

$$D_G(E) = \{w \mid \left(\bigcap_{i \in G} R_i(w)\right) \subseteq E\}$$

Distributed Knowledge

$$D_G(E) = \{w \mid \left(\bigcap_{i\in G} R_i(w)\right) \subseteq E\}$$

- $K_A(p) \wedge K_B(p \to q) \to D_{A,B}(q)$
- $D_G(\varphi) \to \bigwedge_{i\in G} K_i\varphi$

Distributed Knowledge

$$D_G(E) = \{w \mid \left(\bigcap_{i \in G} R_i(w)\right) \subseteq E\}$$

- $K_A(p) \wedge K_B(p \rightarrow q) \rightarrow D_{A,B}(q)$
- $D_G(\varphi) \rightarrow \bigwedge_{i \in G} K_i\varphi$

F. Roelofsen. *Distributed Knowledge*. Journal of Applied Nonclassical Logic (2006).

Distributed Knowledge

$$D_G(E) = \{w \mid \left(\bigcap_{i \in G} R_i(w)\right) \subseteq E\}$$

- $K_A(p) \wedge K_B(p \to q) \to D_{A,B}(q)$
- $D_G(\varphi) \to \bigwedge_{i \in G} K_i\varphi$

F. Roelofsen. *Distributed Knowledge*. Journal of Applied Nonclassical Logic (2006).

$w \in K_G(E)$ iff $R_G(w) \subseteq E$ (without necessarily $R_G(w) = \bigcap_{i \in G} R_i(w)$)

A. Baltag and S. Smets. *Correlated Knowledge: an Epistemic-Logic view on Quantum Entanglement*. Int. Journal of Theoretical Physics (2010).

Ingredients of a Logical Analysis of Rational Agency

- ⇒ informational attitudes (eg., knowledge, belief, certainty)
- ⇒ time, actions and ability
- ⇒ motivational attitudes (eg., preferences)
- ⇒ group notions (e.g., common knowledge and coalitional ability)
- ⇒ normative attitudes (eg., obligations)

Ingredients of a Logical Analysis of Rational Agency

✓ informational attitudes (eg., knowledge, belief, certainty)

$\Rightarrow$ time, actions and ability

$\Rightarrow$ motivational attitudes (eg., preferences)

✓ group notions (e.g., common knowledge)

$\Rightarrow$ normative attitudes (eg., obligations)

Robert Aumann. *Agreeing to Disagree*. Annals of Statistics **4** (1976).

Theorem. Suppose that n agents share a common prior and have different private information. If there is common knowledge in the group of the posterior probabilities, then the posteriors must be equal.

Robert Aumann. *Agreeing to Disagree*. Annals of Statistics **4** (1976).

Theorem. Suppose that n agents share a common prior and have different private information. If there is common knowledge in the group of the posterior probabilities, then the posteriors must be equal.

S. Morris. *The common prior assumption in economic theory*. Economics and Philosophy, 11, pgs. 227 - 254, 1995.

Generalized Aumann's Theorem

Qualitative versions: *like-minded individuals cannot agree to make different decisions*.

M. Bacharach. *Some Extensions of a Claim of Aumann in an Axiomatic Model of Knowledge*. Journal of Economic Theory (1985).

J.A.K. Cave. *Learning to Agree*. Economic Letters (1983).

D. Samet. *Agreeing to disagree: The non-probabilistic case*. Games and Economic Behavior, Vol. 69, 2010, 169-174.

The Framework

Knowledge Structure: $\langle W, \{\Pi_i\}_{i\in\mathcal{A}}\rangle$ where each Π_i is a partition on W ($\Pi_i(w)$ is the cell in Π_i containing w).

Decision Function: Let D be a nonempty set of **decisions.** A decision function for $i \in \mathcal{A}$ is a function $\mathbf{d}_i : W \to D$. A vector $\mathbf{d} = (d_1, \ldots, d_n)$ is a decision function profile. Let $[\mathbf{d}_i = d] = \{w \mid \mathbf{d}_i(w) = d\}$.

The Framework

Knowledge Structure: $\langle W, \{\Pi_i\}_{i \in \mathcal{A}} \rangle$ where each Π_i is a partition on W ($\Pi_i(w)$ is the cell in Π_i containing w).

Decision Function: Let D be a nonempty set of **decisions.** A decision function for $i \in \mathcal{A}$ is a function $\mathbf{d}_i : W \to D$. A vector $\mathbf{d} = (d_1, \ldots, d_n)$ is a decision function profile. Let $[\mathbf{d}_i = d] = \{w \mid \mathbf{d}_i(w) = d\}$.

(A1) Each agent knows her own decision:

$$[\mathbf{d}_i = d] \subseteq K_i([\mathbf{d}_i = d])$$

Comparing Knowledge

$[j \succeq i]$: agent j is at least as knowledgeable as agent i.

$$[j \succeq i] := \bigcap_{E \in \wp(W)} (K_i(E) \Rightarrow K_j(E)) = \bigcap_{E \in \wp(W)} (\neg K_i(E) \cup K_j(E))$$

Comparing Knowledge

$[j \succeq i]$: agent j is at least as knowledgeable as agent i.

$$[j \succeq i] := \bigcap_{E \in \wp(W)} (K_i(E) \Rightarrow K_j(E)) = \bigcap_{E \in \wp(W)} (\neg K_i(E) \cup K_j(E))$$

$w \in [j \succeq i]$ then j knows at w every event that i knows there.

Comparing Knowledge

$[j \succeq i]$: agent j is at least as knowledgeable as agent i.

$$[j \succeq i] := \bigcap_{E \in \wp(W)} (K_i(E) \Rightarrow K_j(E)) = \bigcap_{E \in \wp(W)} (\neg K_i(E) \cup K_j(E))$$

$w \in [j \succeq i]$ then j knows at w every event that i knows there.

$[j \sim i] = [j \succeq i] \cap [i \succeq j]$

The Sure-Thing Principle

A businessman contemplates buying a certain piece of property. He considers the outcome of the next presidential election relevant.

The Sure-Thing Principle

A businessman contemplates buying a certain piece of property. He considers the outcome of the next presidential election relevant. So, to clarify the matter to himself, he asks whether he would buy if he knew *that the Democratic candidate were going to win, and decides that he would.*

The Sure-Thing Principle

A businessman contemplates buying a certain piece of property. He considers the outcome of the next presidential election relevant. So, to clarify the matter to himself, he asks whether he would buy if he knew *that the Democratic candidate were going to win, and decides that he would. Similarly, he considers whether he would buy if he* knew *a Republican candidate were going to win, and again he finds that he would.*

The Sure-Thing Principle

A businessman contemplates buying a certain piece of property. He considers the outcome of the next presidential election relevant. So, to clarify the matter to himself, he asks whether he would buy if he knew *that the Democratic candidate were going to win, and decides that he would. Similarly, he considers whether he would buy if he* knew *a Republican candidate were going to win, and again he finds that he would. Seeing that he would buy in either event, he decides that he should buy, even though he does not know which event obtains, or will obtain, as we would ordinarily say.* *(Savage, 1954)*

The sure-thing principle cannot appropriately be accepted as a postulate...because it would introduce new undefined technical terms referring to knowledge and possibility that would render it mathematically useless without still more postulates governing these terms. It will be preferable to regard the principle as a loose one that suggests certain formal postulates well articulated with P1 [the transitivity of preferences] (Savage, 1954)

Sure-Thing Principle

Should I study or have a beer?

Sure-Thing Principle

Should I study or have a beer? Either I pass or I won't pass the exam.

Sure-Thing Principle

Should I study or have a beer? Either I pass or I won't pass the exam. If I pass, it is better to drink and pass, so I should drink. If I fail, it is better to drink and fail, so I should drink.

Sure-Thing Principle

Should I study or have a beer? Either I pass or I won't pass the exam. If I pass, it is better to drink and pass, so I should drink. If I fail, it is better to drink and fail, so I should drink. I should drink in either case, so I should have a drink.

Sure-Thing Principle

It is not the logical principle $\varphi \to \chi$ and $\psi \to \chi$ then $\varphi \vee \psi \to \chi$.

Sure-Thing Principle

It is not the logical principle $\varphi \rightarrow \chi$ and $\psi \rightarrow \chi$ then $\varphi \vee \psi \rightarrow \chi$.
There is a book I want to read which was written by one of two authors.

Sure-Thing Principle

It is not the logical principle $\varphi \rightarrow \chi$ and $\psi \rightarrow \chi$ then $\varphi \vee \psi \rightarrow \chi$. There is a book I want to read which was written by one of two authors. If I know it is written by author *A* then I will read it. If I know it is written by author *B* then I will read it.

Sure-Thing Principle

It is not the logical principle $\varphi \to \chi$ and $\psi \to \chi$ then $\varphi \vee \psi \to \chi$. There is a book I want to read which was written by one of two authors. If I know it is written by author A then I will read it. If I know it is written by author B then I will read it. If I know it is written by either author A or author B then I may not choose to read the book.

Sure-Thing Principle

R. Aumann, S. Hart and M. Perry. *Conditioning and the Sure-Thing Principle*. manuscript, 2005.

J. Pearl. *The Sure-Thing Principle*. Journal of Causal Inference, Causal, Casual, and Curious Section, 4(1):81-86, 2016.

Branden Fitelson. *Confirmation, Causation, and Simpson's Paradox*. Episteme, 2017.

The Nixon Diamond

You're told (from a reliable source) that Nixon is a republican, which suggests that he is a Hawk. You're also told (from a reliable source) that Nixon is a Quaker, which suggests that he is a Dove.

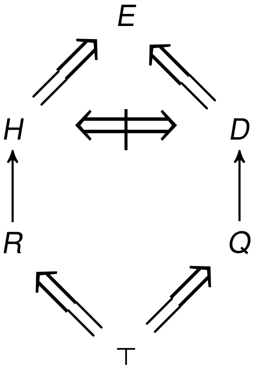
The Nixon Diamond

You're told (from a reliable source) that Nixon is a republican, which suggests that he is a Hawk. You're also told (from a reliable source) that Nixon is a Quaker, which suggests that he is a Dove. Either being a Hawk or a Dove implies having extreme political views.

The Nixon Diamond

You're told (from a reliable source) that Nixon is a republican, which suggests that he is a Hawk. You're also told (from a reliable source) that Nixon is a Quaker, which suggests that he is a Dove. Either being a Hawk or a Dove implies having extreme political views. Should you conclude that Nixon has extreme political views?

Floating Conclusions



J. Horty. *Skepticism and floating conclusions*. Artificial Intelligence, 135, pp. 55 - 72, 2002.

Your parents have 1M inheritance which will is split between you mother and father (each may give you 0.5M).

Your parents have 1M inheritance which will is split between you mother and father (each may give you 0.5M). Your brother (a reliable source) says that you will receive the money from your Mother (but not your Father).

Your parents have 1M inheritance which will is split between you mother and father (each may give you 0.5M). Your brother (a reliable source) says that you will receive the money from your Mother (but not your Father). Your sister (a reliable source) says that you will receive the money from your Father (but not your Mother).

Your parents have 1M inheritance which will is split between you mother and father (each may give you 0.5M). Your brother (a reliable source) says that you will receive the money from your Mother (but not your Father). Your sister (a reliable source) says that you will receive the money from your Father (but not your Mother). You want to buy a yacht which requires a large deposit and you can only afford it provided you inherit the money.

Your parents have 1M inheritance which will is split between you mother and father (each may give you 0.5M). Your brother (a reliable source) says that you will receive the money from your Mother (but not your Father). Your sister (a reliable source) says that you will receive the money from your Father (but not your Mother). You want to buy a yacht which requires a large deposit and you can only afford it provided you inherit the money. Should you make a deposit on the yacht?

Interpersonal Sure-Thing Principle (ISTP)

For any pair of agents i and j and decision d,

$$K_i([j \succeq i] \cap [\mathbf{d}_j = d]) \subseteq [\mathbf{d}_i = d]$$

Interpersonal Sure-Thing Principle (ISTP): Illustration

Suppose that Alice and Bob, two detectives who graduated the same police academy, are assigned to investigate a murder case.

Interpersonal Sure-Thing Principle (ISTP): Illustration

Suppose that Alice and Bob, two detectives who graduated the same police academy, are assigned to investigate a murder case. If they are exposed to different evidence, they may reach different decisions.

Interpersonal Sure-Thing Principle (ISTP): Illustration

Suppose that Alice and Bob, two detectives who graduated the same police academy, are assigned to investigate a murder case. If they are exposed to different evidence, they may reach different decisions. Yet, being the students of the same academy, the method by which they arrive at their conclusions is the same.

Interpersonal Sure-Thing Principle (ISTP): Illustration

Suppose that Alice and Bob, two detectives who graduated the same police academy, are assigned to investigate a murder case. If they are exposed to different evidence, they may reach different decisions. Yet, being the students of the same academy, the method by which they arrive at their conclusions is the same. Suppose now that detective Bob, a father of four who returns home every day at five oclock, collects all the information about the case at hand together with detective Alice.

Interpersonal Sure-Thing Principle (ISTP): Illustration

However, Alice, single and a workaholic, continues to collect more information every day until the wee hours of the morning — information which she does not necessarily share with Bob.

Interpersonal Sure-Thing Principle (ISTP): Illustration

However, Alice, single and a workaholic, continues to collect more information every day until the wee hours of the morning — information which she does not necessarily share with Bob. Obviously, Bob knows that Alice is at least as knowledgeable as he is.

Interpersonal Sure-Thing Principle (ISTP): Illustration

However, Alice, single and a workaholic, continues to collect more information every day until the wee hours of the morning — information which she does not necessarily share with Bob. Obviously, Bob knows that Alice is at least as knowledgeable as he is. Suppose that he also knows what Alices decision is.

Interpersonal Sure-Thing Principle (ISTP): Illustration

However, Alice, single and a workaholic, continues to collect more information every day until the wee hours of the morning — information which she does not necessarily share with Bob. Obviously, Bob knows that Alice is at least as knowledgeable as he is. Suppose that he also knows what Alices decision is. Since Alice uses the same investigation method as Bob, he knows that had he been in possession of the more extensive knowledge that Alice has collected, he would have made the same decision as she did. Thus, this is indeed his decision.

Implications of ISTP

Proposition. If the decision function profile $\mathbf{d}$ satisfies ISTP, then

$$[i \sim j] \subseteq \bigcup_{d \in D}([\mathbf{d}_i = d] \cap [\mathbf{d}_j = d])$$

ISTP Expandability

Agent i is an **epistemic dummy** if it is always the case that all the agents are at least as knowledgeable as i. That is, for each agent j,

$$[j \geq i] = W$$

A decision function profile $\mathbf{d}$ on $\langle W, \Pi_1, \ldots, \Pi_n \rangle$ is **ISTP expandable** if for any expanded structure $\langle W, \Pi_1, \ldots, \Pi_{n+1} \rangle$ where $n+1$ is an epistemic dummy, there exists a decision function $\mathbf{d}_{n+1}$ such that $(\mathbf{d}_1, \mathbf{d}_2, \ldots, \mathbf{d}_{n+1})$ satisfies ISTP.

ISTP Expandability: Illustration

Suppose that after making their decisions, Alice and Bob are told that another detective, one E.P. Dummy, who graduated the very same police academy, had also been assigned to investigate the same case.

ISTP Expandability: Illustration

Suppose that after making their decisions, Alice and Bob are told that another detective, one E.P. Dummy, who graduated the very same police academy, had also been assigned to investigate the same case. In principle, they would need to review their decisions in light of the third detectives knowledge: knowing what they know about the third detective, his usual sources of information, for example, may impinge upon their decision.

ISTP Expandability: Illustration

But this is not so in the case of detective Dummy. It is commonly known that the only information source of this detective, known among his colleagues as the couch detective, is the TV set.

ISTP Expandability: Illustration

But this is not so in the case of detective Dummy. It is commonly known that the only information source of this detective, known among his colleagues as the couch detective, is the TV set. Thus, it is commonly known that every detective is at least as knowledgeable as Dummy.

ISTP Expandability: Illustration

But this is not so in the case of detective Dummy. It is commonly known that the only information source of this detective, known among his colleagues as the couch detective, is the TV set. Thus, it is commonly known that every detective is at least as knowledgeable as Dummy. The news that he had been assigned to the same case is completely irrelevant to the conclusions that Alice and Bob have reached. Obviously, based on the information he gets from the media, Dummy also makes a decision.

ISTP Expandability: Illustration

But this is not so in the case of detective Dummy. It is commonly known that the only information source of this detective, known among his colleagues as the couch detective, is the TV set. Thus, it is commonly known that every detective is at least as knowledgeable as Dummy. The news that he had been assigned to the same case is completely irrelevant to the conclusions that Alice and Bob have reached. Obviously, based on the information he gets from the media, Dummy also makes a decision. We may assume that the decisions made by the three detectives satisfy the ISTP, for exactly the same reason we assumed it for the two detectives decisions

Generalized Agreement Theorem

If **d** is an ISTP expandable decision function profile on a partition structure $\langle W, \Pi_1, \ldots, \Pi_n \rangle$, then for any decisions $d_1, \ldots, d_n$ which are not identical, $C(\bigcap_i [\mathbf{d}_i = d_i]) = \emptyset$.