Ten Puzzles and Paradoxes about Knowledge and Belief

ESSLLI 2013, Düsseldorf

Wes Holliday **Eric Pacuit**

August 12, 2013

Ten Puzzles and Paradoxes

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

Ten Puzzles and Paradoxes

1. Surprise Exam
2. The Knower
3. Logical Omniscience/Knowledge Closure
4. Lottery Paradox & Preface Paradox
5. Margin of Error Paradox
6. Fitch's Paradox
7. Aumann's Agreeing to Disagree Theorem
8. Brandenburger-Keisler Paradox
9. Absent-Minded Driver
10. Backward Induction
11. A puzzle about the sure-thing principle
12. Modeling awareness

Example (Berkeley and Düsseldorf)

Let K_b stand for **agent** b **knows that** and K_d stand for **agent** d **knows that**. Suppose agent b, who lives in Berkeley, knows that agent d lives in Düsseldorf. Let r stand for 'it's raining in Düsseldorf'. Although b doesn't know whether it's raining in Düsseldorf, b knows that d knows whether it's raining there:

Example (Berkeley and Düsseldorf)

Let K_b stand for **agent** b **knows that** and K_d stand for **agent** d **knows that**. Suppose agent b, who lives in Berkeley, knows that agent d lives in Düsseldorf. Let r stand for 'it's raining in Düsseldorf'. Although b doesn't know whether it's raining in Düsseldorf, b knows that d knows whether it's raining there:

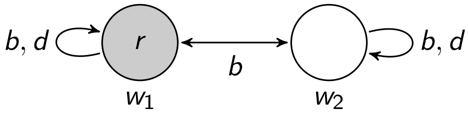
$$\neg(K_b r \vee K_b \neg r) \wedge K_b(K_d r \vee K_d \neg r).$$

Example (Berkeley and Düsseldorf)

Let K_b stand for **agent** b **knows that** and K_d stand for **agent** d **knows that**. Suppose agent b, who lives in Berkeley, knows that agent d lives in Düsseldorf. Let r stand for 'it's raining in Düsseldorf'. Although b doesn't know whether it's raining in Düsseldorf, b knows that d knows whether it's raining there:

$$\neg(K_b r \vee K_b \neg r) \wedge K_b(K_d r \vee K_d \neg r).$$

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

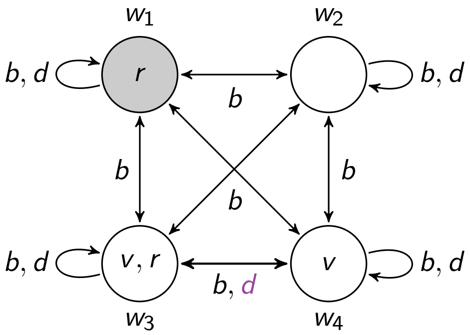


Now suppose that agent b doesn't know whether agent d has left Düsseldorf for a vacation. (Let v stand for 'd has left Düsseldorf on vacation'.) Agent b knows that if d is not on vacation, then d knows whether it's raining in Düsseldorf; but if d is on vacation, then d won't bother to follow the weather.

$$K_b(\neg v \to (K_d r \vee K_d \neg r)) \wedge K_b(v \to \neg(K_d r \vee K_d \neg r)).$$

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

$$K_b(\neg v \to (K_d r \vee K_d \neg r)) \wedge K_b(v \to \neg(K_d r \vee K_d \neg r)).$$



Definition (Truth)

Given a κ-agent model $\mathcal{M} = \langle W, \{R_a\}_{a \in \mathsf{Agt}}, V \rangle$ with $w \in W$ and $\varphi \in \mathcal{L}^{\kappa}_{\mathsf{EL}}$, we define $\mathcal{M}, w \vDash \varphi$ ("φ is true in $\mathcal{M}$ at w") as follows:

$$
\begin{array}{lll}
\mathcal{M}, w \vDash p & \text{iff} & w \in V(p); \\
\mathcal{M}, w \vDash \neg\varphi & \text{iff} & \mathcal{M}, w \nvDash \varphi; \\
\mathcal{M}, w \vDash (\varphi \wedge \psi) & \text{iff} & \mathcal{M}, w \vDash \varphi \text{ and } \mathcal{M}, w \vDash \psi;
\end{array}
$$

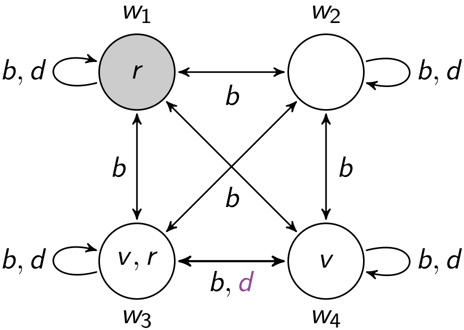
Definition (Truth)

Given a κ-agent model $\mathcal{M} = \langle W, \{R_a\}_{a \in \mathsf{Agt}}, V \rangle$ with $w \in W$ and $\varphi \in \mathcal{L}^{\kappa}_{\mathsf{EL}}$, we define $\mathcal{M}, w \vDash \varphi$ ("φ is true in $\mathcal{M}$ at w") as follows:

$$
\begin{array}{lll}
\mathcal{M}, w \vDash p & \text{iff} & w \in V(p); \\
\mathcal{M}, w \vDash \neg\varphi & \text{iff} & \mathcal{M}, w \nvDash \varphi; \\
\mathcal{M}, w \vDash (\varphi \wedge \psi) & \text{iff} & \mathcal{M}, w \vDash \varphi \text{ and } \mathcal{M}, w \vDash \psi; \\
\mathcal{M}, w \vDash K_a\varphi & \text{iff} & \forall v \in W \colon \text{if } wR_av \text{ then } \mathcal{M}, v \vDash \varphi;
\end{array}
$$

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

$$K_b(\neg v \rightarrow (K_d r \vee K_d \neg r)) \wedge K_b(v \rightarrow \neg(K_d r \vee K_d \neg r)).$$



Outline

- Modeling Knowledge
- The Prediction Paradox
    - The Surprise Exam Paradox
    - The Designated Student Paradox
    - The Knower Paradox
- Knowing What Follows
    - The Problem of Logical Omniscience
    - Epistemic Closure & the Skeptical Paradox
    - The Lottery & Preface Paradoxes

A Word about the Word 'Paradox'

What is a paradox? Roughly: an apparently good argument from apparently true premises to an apparently false/absurd conclusion.

A Word about the Word 'Paradox'

What is a paradox? Roughly: an apparently good argument from apparently true premises to an apparently false/absurd conclusion.

Following Quine, there are at least three classes of paradoxes:

A Word about the Word 'Paradox'

What is a paradox? Roughly: an apparently good argument from apparently true premises to an apparently false/absurd conclusion.

Following Quine, there are at least three classes of paradoxes:

- Veridical paradox: we've decided the conclusion is true despite any initial air of absurdity (e.g., the Barber Paradox).

A Word about the Word 'Paradox'

What is a paradox? Roughly: an apparently good argument from apparently true premises to an apparently false/absurd conclusion.

Following Quine, there are at least three classes of paradoxes:

- Veridical paradox: we've decided the conclusion is true despite any initial air of absurdity (e.g., the Barber Paradox).
- Falsidical paradox: we've decided the conclusion is false and spotted a fallacy in the argument (e.g., the "proof" of $2 = 1$).

A Word about the Word 'Paradox'

What is a paradox? Roughly: an apparently good argument from apparently true premises to an apparently false/absurd conclusion.

Following Quine, there are at least three classes of paradoxes:

- Veridical paradox: we've decided the conclusion is true despite any initial air of absurdity (e.g., the Barber Paradox).
- Falsidical paradox: we've decided the conclusion is false and spotted a fallacy in the argument (e.g., the "proof" of $2 = 1$).
- Genuine antinomy: we are unable to decide whether it is veridical or falsidical (e.g., the Liar paradox?).

A Word about the Word 'Paradox'

What is a paradox? Roughly: an apparently good argument from apparently true premises to an apparently false/absurd conclusion.

Following Quine, there are at least three classes of paradoxes:

- Veridical paradox: we've decided the conclusion is true despite any initial air of absurdity (e.g., the Barber Paradox).
- Falsidical paradox: we've decided the conclusion is false and spotted a fallacy in the argument (e.g., the "proof" of $2 = 1$).
- Genuine antinomy: we are unable to decide whether it is veridical or falsidical (e.g., the Liar paradox?).

W.V. Quine. "The Ways of Paradox." 1966.

The Prediction Paradox

Variants of our first paradox, the prediction paradox, were first discussed by logicians and philosophers in the 1940s.

The Prediction Paradox

Variants of our first paradox, the prediction paradox, were first discussed by logicians and philosophers in the 1940s.

Reportedly **Tarski** passed it along from Berkeley to Quine in the early 40s, and **Gödel** presented it at Princeton in '47.

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?

The Designated Student Paradox

Here is a version of Sorensen's *designated student paradox*:

The Designated Student Paradox

Here is a version of Sorensen's *designated student paradox*:

A teacher shows her class of $n \geq 2$ clever logicians one gold star and $n - 1$ silver stars. After lining the students up, single file, she walks behind each student and sticks one of the stars on his back. No student can see his own back, but each can see the backs of all students in front of him. The teacher announces that the student with the gold star will be **surprised** to learn that he has it.

The Designated Student Paradox

Here is a version of Sorensen's *designated student paradox*:

A teacher shows her class of $n \geq 2$ clever logicians one gold star and $n-1$ silver stars. After lining the students up, single file, she walks behind each student and sticks one of the stars on his back. No student can see his own back, but each can see the backs of all students in front of him. The teacher announces that the student with the gold star will be **surprised** to learn that he has it.

(This is clearly analogous to the surprise exam setup, but we have added a subtle but important difference. Think about it . . .)

The Designated Student Paradox

Student 1, at the front of the line, replies:

The Designated Student Paradox

Student 1, at the front of the line, replies:

- you can't give the gold star to student n, because then he'd see all silver stars and therefore know he has the gold star;

The Designated Student Paradox

Student 1, at the front of the line, replies:

- you can't give the gold star to student n, because then he'd see all silver stars and therefore know he has the gold star;
- you also can't give the gold star to student $n-1$, because then he'd see all silver stars and therefore know he has the gold star, having ruled out the possibility that student n has the gold star by the previous reasoning.

The Designated Student Paradox

Student 1, at the front of the line, replies:

- you can't give the gold star to student n, because then he'd see all silver stars and therefore know he has the gold star;
- you also can't give the gold star to student $n-1$, because then he'd see all silver stars and therefore know he has the gold star, having ruled out the possibility that student n has the gold star by the previous reasoning.
- you also can't give the gold star to student $n-2$, etc.

He concludes that the teacher's claim about a surprise is false.

The Designated Student Paradox

Student 1, at the front of the line, replies:

- you can't give the gold star to student n, because then he'd see all silver stars and therefore know he has the gold star;
- you also can't give the gold star to student $n-1$, because then he'd see all silver stars and therefore know he has the gold star, having ruled out the possibility that student n has the gold star by the previous reasoning.
- you also can't give the gold star to student $n-2$, etc.

He concludes that the teacher's claim about a surprise is false.

But then the students pull the stars off their backs and it is, say, student $n-1$ who has the gold star, and he is surprised.

The Designated Student Paradox

Student 1, at the front of the line, replies:

- you can't give the gold star to student n, because then he'd see all silver stars and therefore know he has the gold star;
- you also can't give the gold star to student $n-1$, because then he'd see all silver stars and therefore know he has the gold star, having ruled out the possibility that student n has the gold star by the previous reasoning.
- you also can't give the gold star to student $n-2$, etc.

He concludes that the teacher's claim about a surprise is false.

But then the students pull the stars off their backs and it is, say, student $n-1$ who has the gold star, and he is surprised.

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

The Surprise Exam & Designated Student Paradoxes

Many "solutions" of the surprise exam paradox and its variations have been given by philosophers and logicians in the last 60+ years.

The Surprise Exam & Designated Student Paradoxes

Many "solutions" of the surprise exam paradox and its variations have been given by philosophers and logicians in the last 60+ years.

We won't try to survey the solutions that have been given or argue for a particular solution here. Instead, we'll just try to get a clearer understanding of the paradox by formalizing it in epistemic logic.

The Surprise Exam & Designated Student Paradoxes

Many "solutions" of the surprise exam paradox and its variations have been given by philosophers and logicians in the last 60+ years.

We won't try to survey the solutions that have been given or argue for a particular solution here. Instead, we'll just try to get a clearer understanding of the paradox by formalizing it in epistemic logic.

One thing that formalization forces us to do is to make explicit a number of suppressed assumptions behind the clever student's reasoning, without which a paradox can't be generated.

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

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

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

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

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

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

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

Step 1: Choosing the Formalism (language)

To formalize the paradoxes, we use the epistemic language

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

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

Step 1: Choosing the Formalism (language)

To formalize the paradoxes, we use the epistemic language

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

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

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

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

Step 1: Choosing the Formalism (language)

To formalize the paradoxes, we use the epistemic language

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

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

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

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

For the designated student paradox, we read

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

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

Step 1: Choosing the Formalism (reasoning system)

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

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

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

Step 1: Choosing the Formalism (reasoning system)

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

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

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

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

Step 1: Choosing the Formalism (reasoning system)

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

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

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

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

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

Step 1: Choosing the Formalism (reasoning system)

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

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

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

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

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

Step 1: Choosing the Formalism (reasoning system)

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

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

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

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

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

Step 1: Choosing the Formalism (reasoning system)

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

Step 1: Choosing the Formalism (reasoning system)

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

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

Step 1: Choosing the Formalism (reasoning system)

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

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

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

Step 1: Choosing the Formalism (reasoning system)

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

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

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

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

Step 1: Choosing the Formalism (reasoning system)

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

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

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

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

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

Step 1: Choosing the Formalism (reasoning system)

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

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

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

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

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

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

Step 1: Choosing the Formalism (reasoning system)

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

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

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

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

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

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

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

Step 1: Choosing the Formalism (reasoning system)

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

(A) $K_1((p_1 \land \neg K_1 p_1) \lor (p_2 \land \neg K_2 p_2))$ premise

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

(C) $K_1 K_2(p_1 \lor p_2)$ 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) $K_1(p_2 \to K_2 \neg p_1)$ premise

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

(1.1) $((p_1 \vee p_2) \wedge \neg p_1) \to p_2)$ 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, $K_1\varphi \to \varphi$, or the "weak factivity" axiom J_1, $K_1\neg K_1\varphi \to \neg K_1\varphi$ (e.g., reading K as belief instead of knowledge), then we can derive a contradiction:

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

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

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

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

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

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

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

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

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

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

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):

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

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

(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).
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: $\{(A^3), (B^3)\} \vdash_{\mathbf{K4}_1^<} K_1(p \land \neg K_1 p_1)$

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

(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

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

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

(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

$(2,4)$ $K_1\neg S_2$ 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) $K_1(((p_2 \vee p_3) \to K_2 \neg p_1) \wedge (p_3 \to K_3 \neg(p_1 \vee p_2))$;

(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

$(2,4)$ $K_1 \neg S_2$ 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) $K_1(((p_2 \vee p_3) \rightarrow K_2 \neg p_1) \wedge (p_3 \rightarrow K_3 \neg (p_1 \vee p_2))$.

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

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

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:

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

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

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

Step 4: Showing Consistency with a Model

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

Let's now establish the previous claim about the consistency of (A^3), (B^3), (C^3), even with strong assumptions about knowledge.

Step 4: Showing Consistency with a Model

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

Let's now establish the previous claim about the consistency of (A^3), (B^3), (C^3), even with strong assumptions about knowledge.

Even adding to **K** the T schema, $K_i\varphi \to \varphi$, and the 5 schema, $\neg K_i\varphi \to K_i\neg K_i\varphi$, to obtain the strong system **S5**, we have:

$$\{(A^3), (B^3), (C^3)\} \nvdash_{\mathbf{S5}} \bot.$$

Step 4: Showing Consistency with a Model

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

Let's now establish the previous claim about the consistency of (A^3), (B^3), (C^3), even with strong assumptions about knowledge.

Even adding to **K** the T schema, $K_i\varphi \to \varphi$, and the 5 schema, $\neg K_i\varphi \to K_i \neg K_i \varphi$, to obtain the strong system **S5**, we have:

$$\{(A^3), (B^3), (C^3)\} \nvdash_{\mathbf{S5}} \bot.$$

To show this, we'll turn to models for the epistemic language.

Step 4: Showing Consistency with a Model

The logic **S5** is sound with respect to the class of relational models $\mathcal{M} = \langle W, \{R_i\}_{i \in \mathbb{N}}, V \rangle$ where each R_i is an *equivalence relation*, i.e., reflexive, symmetric, and transitive.

Step 4: Showing Consistency with a Model

The logic **S5** is sound with respect to the class of relational models $\mathcal{M} = \langle W, \{R_i\}_{i \in \mathbb{N}}, V \rangle$ where each R_i is an *equivalence relation*, i.e., reflexive, symmetric, and transitive.

Thus, if we can construct such a model in which (A^3), (B^3), and (C^3) are all true, then we have $\{(A^3), (B^3), (C^3)\} \nvdash_{\mathbf{S5}} \bot$.

Step 4: Showing Consistency with a Model

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

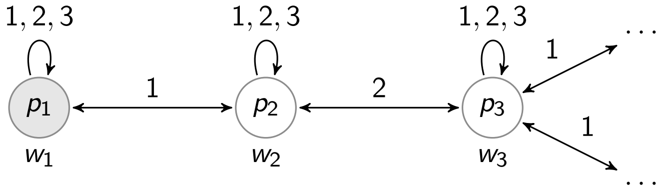


Figure: model establishing **S5**-consistency of $(A^3), (B^3), (C^3)$.

Observe that $K_1 \neg p_3 \to K_1 K_2 \neg p_3$, an instance of $4_1^<$, is false at w_1.

Summary of What We've Seen

Here's a summary of what we've seen:

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

Summary of What We've Seen

Here's a summary of what we've seen:

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

Summary of What We've Seen

Here's a summary of what we've seen:

- $\{(A^2), (B^2), (C^2)\} \vdash_{\mathbf{K}} K_1(p_1 \wedge \neg K_1)$;
- $\{(A^2), (B^2), (C^2)\} \vdash_{\mathbf{KJ_1}} \bot$ 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 of What We've Seen

Here's a summary of what we've seen:

- $\{(A^2), (B^2), (C^2)\} \vdash_{\mathbf{K}} K_1(p_1 \wedge \neg K_1)$;
- $\{(A^2), (B^2), (C^2)\} \vdash_{\mathbf{KJ_1}} \bot$ and $\{(A^2), (B^2), (C^2)\} \vdash_{\mathbf{KT_1}} \bot$;
- $\{(A^3), (B^3), (C^3)\} \nvdash_{\mathbf{S5}} \bot$.
- $\{(A^3), (B^3)\} \vdash_{\mathbf{K4_1^<}} K_1(p_1 \wedge \neg K_1)$;
- $\{(A^3), (B^3)\} \vdash_{\mathbf{KJ_1 4_1^<}} \bot$ and $\{(A^3), (B^3)\} \vdash_{\mathbf{KT_1 4_1^<}} \bot$;

With these facts, one can make a strong case that the culprit behind the paradoxes is the (mistaken) $4_1^<$ axiom, $K_1\varphi \to K_1 K_i \varphi$ $(i > 1)$. But we don't have time to explain this solution. See

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

The Knower

We can't resist mentioning another paradox that Kaplan and Montague consider a limiting case of the surprise exam paradox.

D. Kaplan and R. Montague. 1960. "A Paradox Regained," *NDJFL*.

The Knower

Suppose that instead of representing knowledge with a sentential operator added to the language of propositional logic, we do so with a predicate added to the language of Peano Arithmetic: where $\ulcorner\varphi\urcorner$ is the Gödel number of φ, $\mathsf{K}(\ulcorner\varphi\urcorner)$ means the agent knows φ.

The Knower

Suppose that instead of representing knowledge with a sentential operator added to the language of propositional logic, we do so with a predicate added to the language of Peano Arithmetic: where $\ulcorner\varphi\urcorner$ is the Gödel number of φ, $\mathsf{K}(\ulcorner\varphi\urcorner)$ means the agent knows φ.

Consider the following assumptions (schemas) for the K predicate:

(K1) $\mathsf{K}(\ulcorner\varphi\urcorner) \to \varphi$;

The Knower

Suppose that instead of representing knowledge with a sentential operator added to the language of propositional logic, we do so with a predicate added to the language of Peano Arithmetic: where $\ulcorner\varphi\urcorner$ is the Gödel number of φ, $\mathsf{K}(\ulcorner\varphi\urcorner)$ means the agent knows φ.

Consider the following assumptions (schemas) for the K predicate:

(K1) $\mathsf{K}(\ulcorner\varphi\urcorner) \to \varphi$;

(K2) $\mathsf{K}(\ulcorner\mathsf{K}(\ulcorner\varphi\urcorner) \to \varphi\urcorner)$;

The Knower

Suppose that instead of representing knowledge with a sentential operator added to the language of propositional logic, we do so with a predicate added to the language of Peano Arithmetic: where $\ulcorner\varphi\urcorner$ is the Gödel number of φ, $\mathsf{K}(\ulcorner\varphi\urcorner)$ means the agent knows φ.

Consider the following assumptions (schemas) for the K predicate:

(K1) $\mathsf{K}(\ulcorner\varphi\urcorner) \to \varphi$;

(K2) $\mathsf{K}(\ulcorner\mathsf{K}(\ulcorner\varphi\urcorner) \to \varphi\urcorner)$;

(K3) $(\mathsf{Prove}_{\mathbf{PA}}(\ulcorner\varphi\urcorner, \ulcorner\psi\urcorner) \wedge \mathsf{K}(\ulcorner\varphi\urcorner)) \to \mathsf{K}(\ulcorner\psi\urcorner)$.

The Knower

Suppose that instead of representing knowledge with a sentential operator added to the language of propositional logic, we do so with a predicate added to the language of Peano Arithmetic: where $\ulcorner\varphi\urcorner$ is the Gödel number of φ, $\mathsf{K}(\ulcorner\varphi\urcorner)$ means the agent knows φ.

Consider the following assumptions (schemas) for the K predicate:

(K1) $\mathsf{K}(\ulcorner\varphi\urcorner) \to \varphi$;

(K2) $\mathsf{K}(\ulcorner \mathsf{K}(\ulcorner\varphi\urcorner) \to \varphi\urcorner)$;

(K3) $(\mathsf{Prove}_{\mathbf{PA}}(\ulcorner\varphi\urcorner, \ulcorner\psi\urcorner) \wedge \mathsf{K}(\ulcorner\varphi\urcorner)) \to \mathsf{K}(\ulcorner\psi\urcorner)$.

Show: $\{(K1), (K2), (K3)\} \vdash_{\mathbf{PA}} \bot$.

The Knower

Suppose that instead of representing knowledge with a sentential operator added to the language of propositional logic, we do so with a predicate added to the language of Peano Arithmetic: where $\ulcorner\varphi\urcorner$ is the Gödel number of φ, $\mathsf{K}(\ulcorner\varphi\urcorner)$ means the agent knows φ.

Consider the following assumptions (schemas) for the K predicate:

(K1) $\mathsf{K}(\ulcorner\varphi\urcorner) \to \varphi$;

(K2) $\mathsf{K}(\ulcorner \mathsf{K}(\ulcorner\varphi\urcorner) \to \varphi\urcorner)$;

(K3) $(\mathsf{Prove}_{\mathbf{PA}}(\ulcorner\varphi\urcorner, \ulcorner\psi\urcorner) \wedge \mathsf{K}(\ulcorner\varphi\urcorner)) \to \mathsf{K}(\ulcorner\psi\urcorner)$.

Show: $\{(\mathrm{K1}), (\mathrm{K2}), (\mathrm{K3})\} \vdash_{\mathbf{PA}} \bot$.

Hint: by Gödel's fixed-point lemma, there is a sentence β such that

$$\vdash_{\mathbf{PA}} \beta \leftrightarrow K(\ulcorner\neg\beta\urcorner).$$

The Knower

Suppose that instead of representing knowledge with a sentential operator added to the language of propositional logic, we do so with a predicate added to the language of Peano Arithmetic: where $\ulcorner\varphi\urcorner$ is the Gödel number of φ, $\mathsf{K}(\ulcorner\varphi\urcorner)$ means the agent knows φ.

Consider the following assumptions (schemas) for the K predicate:

(K1) $\mathsf{K}(\ulcorner\varphi\urcorner) \to \varphi$;

(K2) $\mathsf{K}(\ulcorner \mathsf{K}(\ulcorner\varphi\urcorner) \to \varphi\urcorner)$;

(K3) $(\mathsf{Prove}_{\mathbf{PA}}(\ulcorner\varphi\urcorner, \ulcorner\psi\urcorner) \wedge \mathsf{K}(\ulcorner\varphi\urcorner)) \to \mathsf{K}(\ulcorner\psi\urcorner)$.

Show: $\{(K1), (K2), (K3)\} \vdash_{\mathbf{PA}} \bot$.

Question: What should we give up in response to this result?

The Knower

For further reading about the knower paradox, see:

S. Maitzen. 1998. "The Knower Paradox and Epistemic Closure," *Synthese*.

C.B. Cross. 2001. "The Paradox of the Knower without Epistemic Closure," *Mind*.

G. Uzquiano. 2004.
"The Paradox of the Knower without Epistemic Closure?" *Mind*.

C.B. Cross. 2004.
"More on the Paradox of the Knower without Epistemic Closure," *Mind*.

Normal Modal Logics

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

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

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

The "Problem" of Logical Omniscience

The rule

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

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

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

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

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

R. Stalnaker.

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

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

The "Problem" of Logical Omniscience

The rule

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

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

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.

Epistemic Closure & the Skeptical Paradox

The problem of logical omniscience must be distinguished from the problem of epistemic closure, which arises even if we assume that our agents are perfect logicians who always “put two and two together” and deduce the consequence of what they know.

Epistemic Closure & the Skeptical Paradox

The problem of logical omniscience must be distinguished from the problem of epistemic closure, which arises even if we assume that our agents are perfect logicians who always "put two and two together" and deduce the consequence of what they know.

The problem of epistemic closure is raised by the so-called *Skeptical Paradox*.

S. Cohen. 1988. "How to be a Fallibilist," *Philosophical Perspectives*.

K. DeRose. 1995. "Solving the Skeptical Problem," *Philosophical Review*.

Epistemic Closure & the Skeptical Paradox

Let p be a mundane proposition, e.g., Eric was born in the U.S., that we think our agent knows.

Epistemic Closure & the Skeptical Paradox

Let p be a mundane proposition, e.g., Eric was born in the U.S., that we think our agent knows.

Let SH be a "skeptical hypothesis" (or a disjunction of hypotheses) incompatible with the truth of p, but according to which everything would be indistinguishable from the actual world for the agent, e.g., Russell's hypothesis that the world was created 5 minutes ago with everyone having false memories of a long past.

Epistemic Closure & the Skeptical Paradox

Let p be a mundane proposition, e.g., Eric was born in the U.S., that we think our agent knows.

Let SH be a "skeptical hypothesis" (or a disjunction of hypotheses) incompatible with the truth of p, but according to which everything would be indistinguishable from the actual world for the agent, e.g., Russell's hypothesis that the world was created 5 minutes ago with everyone having false memories of a long past.

The skeptic argues that since the agent doesn't know $\neg SH$, but she does know the obvious fact that $p \rightarrow \neg SH$, it follows by RK_i that she doesn't know p; i.e., $(Kp \wedge K(p \rightarrow \neg SH)) \rightarrow K\neg SH$ implies

$$(\neg K\neg SH \wedge K(p \rightarrow \neg SH)) \rightarrow \neg Kp.$$

Since p was basically arbitrary, you don't know much of anything.

Epistemic Closure & the Skeptical Paradox

The skeptic argues that since the agent doesn't know $\neg SH$, but she does know the obvious fact that $p \to \neg SH$, it follows by RK_i that she doesn't know p; i.e., $(Kp \wedge K(p \to \neg SH)) \to K\neg SH$ implies

$$(\neg K \neg SH \wedge K(p \to \neg SH)) \to \neg Kp.$$

Since p was basically arbitrary, you don't know much of anything.

Three responses in defense of knowledge:

- Mooreanism: actually, you do know $\neg SH$ (How? One answer: because you know p and know that $p \to \neg SH$. Too cheap?)
- Deny closure: RK_i is invalid; for the strange case of SH versus p, we have Kp, $\neg K \neg SH$, and $K(p \to \neg SH)$.
- Contextualism: in a context where we're not worried about skepticism, we can truly claim Kp; in a context where we are, we can truly claim $\neg K \neg SH$; in every fixed context, RK_i holds.

Epistemic Closure & the Skeptical Paradox

(1) Mooreanism: actually, you do know $\neg SH$ (How? One answer: because you know p and know that $p \to \neg SH$. Too cheap?)

(2) Deny closure: RK_i is invalid; for the strange case of SH against p, the truth is Kp, $\neg K \neg SH$, $K(p \to \neg SH)$.

(3) Contextualism: in a context where we're not worried about skepticism, we can truly claim Kp; in a context where we are, we can truly claim $\neg K \neg SH$; in every fixed context, RK_i holds.

The third option leads naturally to questions about how context is supposed to change as we consider skeptical possibilities. For modeling of this in the framework of dynamic epistemic logic, see:

Wes Holliday (http://philosophy.berkeley.edu/holliday). 2012.

"Epistemic Logic, Relevant Alternatives, and the Dynamics of Context."

Epistemic Closure & the Skeptical Paradox

(2) Deny closure: RK_i is invalid; for the strange case of SH against p, the truth is Kp, $\neg K \neg SH$, $K(p \to \neg SH)$.

The second option leads naturally to questions about what closure principles do hold, if closure under known implication does not. For example, shouldn't $K(\varphi \wedge \psi) \to K\varphi$ still be valid?

Epistemic Closure & the Skeptical Paradox

(2) Deny closure: RK_i is invalid; for the strange case of SH against p, the truth is Kp, $\neg K \neg SH$, $K(p \rightarrow \neg SH)$.

The second option leads naturally to questions about what closure principles do hold, if closure under known implication does not. For example, shouldn't $K(\varphi \wedge \psi) \rightarrow K\varphi$ still be valid? See:

W. Holliday (http://philosophy.berkeley.edu/holliday). 2012.

"Epistemic Closure and Epistemic Logic I:

Relevant Alternatives and Subjunctivism."

Epistemic Closure & the Skeptical Paradox

(2) Deny closure: RK_i is invalid; for the strange case of SH against p, the truth is Kp, $\neg K \neg SH$, $K(p \to \neg SH)$.

The second option leads naturally to questions about what closure principles do hold, if closure under known implication does not. For example, shouldn't $K(\varphi \wedge \psi) \to K\varphi$ still be valid? See:

W. Holliday (http://philosophy.berkeley.edu/holliday). 2012.
"Epistemic Closure and Epistemic Logic I:
Relevant Alternatives and Subjunctivism."

There are open questions about the complete logics of some famous theories of knowledge. See Problem 8.12 in the above.

The Lottery and Preface Paradoxes

Another challenge to RK_i comes from the so-called *lottery paradox* and *preface paradox*.

H.E. Kyburg, Jr. 1961. *Probability and the Logic of Rational Belief*. Wesleyan University Press.

D.C. Makinson. 1965. "The Paradox of the Preface," *Analysis*.

The Lottery Paradox

Let $\Box_i$ stand for **it is rational for i to believe that**. Consider:

$$(\Box_i p \wedge \Box_i q) \rightarrow \Box_i(p \wedge q),$$

which is obviously derivable from the RK rule for $\Box_i$.

The Lottery Paradox

Let $\Box_i$ stand for **it is rational for *i* to believe that**. Consider:

$$(\Box_i p \wedge \Box_i q) \rightarrow \Box_i (p \wedge q),$$

which is obviously derivable from the RK rule for $\Box_i$.

There is a lottery with n tickets, of which one will be drawn.

- Let l_1 stand for 'lottery ticket 1 is the winning ticket'.
- Let l_2 stand for 'lottery ticket 2 is the winning ticket', etc.

The Lottery Paradox

Let $\Box_i$ stand for **it is rational for i to believe that**. Consider:

$$(\Box_i p \land \Box_i q) \rightarrow \Box_i (p \land q),$$

which is obviously derivable from the RK rule for $\Box_i$.

There is a lottery with n tickets, of which one will be drawn.

- Let l_1 stand for 'lottery ticket 1 is the winning ticket'.
- Let l_2 stand for 'lottery ticket 2 is the winning ticket', etc.

If n is large, it seems rational to believe of any individual ticket k that it is not the winning ticket: $\Box_i \neg l_k$.

The Lottery Paradox

Let $\Box_i$ stand for **it is rational for *i* to believe that**. Consider:

$$(\Box_i p \land \Box_i q) \to \Box_i (p \land q),$$

which is obviously derivable from the RK rule for $\Box_i$.

There is a lottery with n tickets, of which one will be drawn.

- Let l_1 stand for ‘lottery ticket 1 is the winning ticket’.
- Let l_2 stand for ‘lottery ticket 2 is the winning ticket’, etc.

If n is large, it seems rational to believe of any individual ticket k that it is not the winning ticket: $\Box_i \neg l_k$.Then according to the principle above, it’s rational to believe that no ticket will win: $\Box_i \neg l_1 \land \cdots \land \Box_i \neg l_n$.

The Lottery Paradox

Let $\Box_i$ stand for **it is rational for i to believe that**. Consider:

$$(\Box_i p \wedge \Box_i q) \to \Box_i (p \wedge q),$$

which is obviously derivable from the RK rule for $\Box_i$.

There is a lottery with n tickets, of which one will be drawn.

- Let l_1 stand for ‘lottery ticket 1 is the winning ticket’.
- Let l_2 stand for ‘lottery ticket 2 is the winning ticket’, etc.

If n is large, it seems rational to believe of any individual ticket k that it is not the winning ticket: $\Box_i \neg l_k$. Then according to the principle above, it’s rational to believe that no ticket will win: $\Box_i(\neg l_1 \wedge \cdots \wedge \neg l_n)$.

The Lottery Paradox

Let $\Box_i$ stand for **it is rational for i to believe that**. Consider:

$$(\Box_i p \wedge \Box_i q) \to \Box_i (p \wedge q),$$

which is obviously derivable from the RK rule for $\Box_i$.

There is a lottery with n tickets, of which one will be drawn.

- Let l_1 stand for 'lottery ticket 1 is the winning ticket'.
- Let l_2 stand for 'lottery ticket 2 is the winning ticket', etc.

If n is large, it seems rational to believe of any individual ticket k that it is not the winning ticket: $\Box_i \neg l_k$. Then according to the principle above, it's rational to believe that no ticket will win: $\Box_i(\neg l_1 \wedge \cdots \wedge \neg l_n)$. But it's certain that one will win!

Reasoning about High Probability

An uncontroversial example of a non-normal operator is **it is highly probable that**, for which $(\Box p \wedge \Box q) \rightarrow \Box(p \wedge q)$ is invalid.

Reasoning about High Probability

An uncontroversial example of a non-normal operator is **it is highly probable that**, for which $(\Box p \wedge \Box q) \rightarrow \Box(p \wedge q)$ is invalid.

Again, for any lottery ticket i, it is highly probably that i will loose. But then by repeated use of $(\Box p \wedge \Box q) \rightarrow \Box(p \wedge q)$, we could derive that it is highly probable that *all* tickets will loose, contradicting the fact that it is certain that one ticket will win.

The Preface Paradox

An author writes a book with n claims, $c_1, \ldots, c_n$, each of which the author checked carefully and therefore believes. Yet the author has written books before and realizes that errors are inevitable in any book; thus, in the preface he says something to the effect of "I thank ... for their help; but all the errors that remain are mine."

The Preface Paradox

An author writes a book with n claims, $c_1, \ldots, c_n$, each of which the author checked carefully and therefore believes. Yet the author has written books before and realizes that errors are inevitable in any book; thus, in the preface he says something to the effect of "I thank ... for their help; but all the errors that remain are mine."

It seems that we have here a situation in which

$$\Box_i c_1 \wedge \cdots \wedge \Box_i c_n$$

and

$$\Box_i \neg(c_1 \wedge \cdots \wedge c_n),$$

which means that the set of propositions believed is *inconsistent*.

The Preface Paradox

An author writes a book with n claims, $c_1, \ldots, c_n$, each of which the author checked carefully and therefore believes. Yet the author has written books before and realizes that errors are inevitable in any book; thus, in the preface he says something to the effect of "I thank . . . for their help; but all the errors that remain are mine."

It seems that we have here a situation in which

$$\Box_i c_1 \land \cdots \land \Box_i c_n$$

and

$$\Box_i \neg(c_1 \land \cdots \land c_n),$$

which means that the set of propositions believed is *inconsistent*.

But is there anything irrational about the author so described?