# Knowledge-of-own-factivity, the definition of surprise, and a solution to the Surprise Examination paradox

Samuel Allen Alexander

The U.S. Securities and Exchange Commission

**Alessandro Aldini** and Pierluigi Graziani

1506
UNIVERSITÀ
DEGLI STUDI
DI URBINO
CARLO BO

CIFMA, 27 September 2022

# Introduction

Agenda

- Knowledge, factivity of knowledge, and knowledge of factivity in epistemic logic
- An application: the Surprise Examination paradox
- Modal logic formalization and results
- Considerations and conclusions

# Background

- Epistemic logics are modal logics for reasoning about knowledge, belief, and related notions.
- Knowledge is represented through the operator $K$.
- $K(\phi)$ reads as $\phi$ *is known*.
- Sometimes, $K_i$ is intended as a temporal characterization, expressing what is known at the time of a given event $e_i$.

# Background

What are the epistemic principles that we may assume as valid in the logic, and what are the consequences of these choices?

# The principles of knowledge

- *Factivity* ($T$) is a widely accepted principle:

$$K(\phi) \rightarrow \phi$$

- Other studied principles:

  $K$ $K(\phi \rightarrow \psi) \rightarrow K(\phi) \rightarrow K(\psi)$

  4 $K(\phi) \rightarrow KK(\phi)$

  - Necessitation: $K(\phi)$ whenever $\phi$ is a tautology
  - Closure: $K(\phi)$ for each *derivable* $\phi$

# The principles of knowledge

Questioning the principles of knowledge

- it is a way to deal with many paradoxes (factivity, introspection, . . . )
- it is way to trade expressiveness for decidability (theory of knowing machines, . . . )

Here, we concentrate on the knowledge of factivity in the setting of a well-known paradox . . .

# The Surprise Examination paradox

A teacher announces that there will be a surprise exam next week. The students reason that the exam cannot occur on Friday (the final day of the school week), because if it did, they would already know by then (by process of elimination) that it must be Friday, and thus it would not be surprising. Having ruled out Friday, Thursday is then the last day on which the exam can possibly occur. By the exact same reasoning, then, the exam cannot be on Thursday, because if it were, they would already know by then (by process of elimination) that it must be Thursday (since they have ruled out Friday already). In similar manner, the examination cannot occur on Wednesday, Tuesday, or Monday. The students conclude that the exam cannot occur at all. They are therefore quite surprised when the teacher gives them the exam anyway.

# Formal machinery

## Notations and assumptions

- We use propositional variables $D_1 \ldots D_n$ for the $n$ school days, where each $D_i$ reads for *the exam takes place on day i* and expresses the teacher decision.
- We use the family of operators $K_i$, where each $K_i(\phi)$ reads for $\phi$ *is known by the students at midnight just before day i*.
- We assume that propositions are *stable*.
- We assume a single-agent system, where each $K_i$ refers to the knowledge of the classroom.
- We assume the classrooom to be *perfect recall*.

# Formalizing the paradox (1)

For each $n \geq 1$, let $T_n$ be the theory consisting of:

- $(A_1^n)$ $\bigvee_{i=1}^{n} D_i$.
- $(A_2^n)$ $\bigwedge_{1 \leq i < j \leq n} \neg(D_i \wedge D_j)$.
- $(A_3^n)$ $\bigvee_{i=1}^{n}(D_i \wedge \neg K_i(D_i))$.
- $(A_4^n)$ $\bigwedge_{i=1}^{n-1}((\neg D_i) \to K_{i+1}(\neg D_i))$.
- $(A_5^n)$ $K_i(\phi)$ for all $1 \leq i \leq n$ and tautologies $\phi$.
- $(A_6^n)$ $K_i(\phi \to \psi) \to K_i(\phi) \to K_i(\psi)$
  for all $1 \leq i \leq n$ and all $\phi, \psi$.
- $(A_7^n)$ $K_i(\phi) \to K_j(\phi)$ for all $1 \leq i < j \leq n$.
- $(A_\infty^n)$ $K_i(\phi)$ for all $1 \leq i \leq n$ and all $\phi$ such that $T_n \models \phi$.

# Formalizing the paradox (1)

For each $n \geq 1$, let $T_n$ be the theory consisting of:

- $(A_1^n)$ $\bigvee_{i=1}^{n} D_i$.
- $(A_2^n)$ $\bigwedge_{1 \leq i < j \leq n} \neg(D_i \wedge D_j)$.
- $(A_3^n)$ $\bigvee_{i=1}^{n}(D_i \wedge \neg K_i(D_i))$.
- $(A_4^n)$ $\bigwedge_{i=1}^{n-1}((\neg D_i) \to K_{i+1}(\neg D_i))$.
- $(A_5^n)$ $K_i(\phi)$ for all $1 \leq i \leq n$ and tautologies $\phi$.
- $(A_6^n)$ $K_i(\phi \to \psi) \to K_i(\phi) \to K_i(\psi)$ $(K)$
  for all $1 \leq i \leq n$ and all $\phi, \psi$.
- $(A_7^n)$ $K_i(\phi) \to K_j(\phi)$ for all $1 \leq i < j \leq n$.
- $(A_\infty^n)$ $K_i(\phi)$ for all $1 \leq i \leq n$ and all $\phi$ such that $T_n \models \phi$.

# Formalizing the paradox (1)

For each $n \geq 1$, let $T_n$ be the theory consisting of:

- $(A_1^n)$ $\bigvee_{i=1}^{n} D_i$. truthfulness of the
- $(A_2^n)$ $\bigwedge_{1 \leq i < j \leq n} \neg(D_i \wedge D_j)$. teacher announcement
- $(A_3^n)$ $\bigvee_{i=1}^{n} (D_i \wedge \neg K_i(D_i))$.
- $(A_4^n)$ $\bigwedge_{i=1}^{n-1} ((\neg D_i) \to K_{i+1}(\neg D_i))$.
- $(A_5^n)$ $K_i(\phi)$ for all $1 \leq i \leq n$ and tautologies $\phi$.
- $(A_6^n)$ $K_i(\phi \to \psi) \to K_i(\phi) \to K_i(\psi)$ for all $1 \leq i \leq n$ and all $\phi, \psi$.
- $(A_7^n)$ $K_i(\phi) \to K_j(\phi)$ for all $1 \leq i < j \leq n$.
- $(A_\infty^n)$ $K_i(\phi)$ for all $1 \leq i \leq n$ and all $\phi$ such that $T_n \models \phi$.

# Formalizing the paradox (1)

For each $n \geq 1$, let $T_n$ be the theory consisting of:

- $(A_1^n)$ $\bigvee_{i=1}^n D_i$.
- $(A_2^n)$ $\bigwedge_{1 \leq i < j \leq n} \neg(D_i \wedge D_j)$.
- $(A_3^n)$ $\bigvee_{i=1}^n (D_i \wedge \neg K_i(D_i))$. surprise!
- $(A_4^n)$ $\bigwedge_{i=1}^{n-1} ((\neg D_i) \rightarrow K_{i+1}(\neg D_i))$.
- $(A_5^n)$ $K_i(\phi)$ for all $1 \leq i \leq n$ and tautologies $\phi$.
- $(A_6^n)$ $K_i(\phi \rightarrow \psi) \rightarrow K_i(\phi) \rightarrow K_i(\psi)$
  for all $1 \leq i \leq n$ and all $\phi, \psi$.
- $(A_7^n)$ $K_i(\phi) \rightarrow K_j(\phi)$ for all $1 \leq i < j \leq n$.
- $(A_\infty^n)$ $K_i(\phi)$ for all $1 \leq i \leq n$ and all $\phi$ such that $T_n \models \phi$.

# Formalizing the paradox (1)

For each $n \geq 1$, let $T_n$ be the theory consisting of:

- $(A_1^n)$ $\bigvee_{i=1}^n D_i$.
- $(A_2^n)$ $\bigwedge_{1 \leq i < j \leq n} \neg(D_i \wedge D_j)$.
- $(A_3^n)$ $\bigvee_{i=1}^n (D_i \wedge \neg K_i(D_i))$.
- $(A_4^n)$ $\bigwedge_{i=1}^{n-1}((\neg D_i) \rightarrow K_{i+1}(\neg D_i))$. stability and perfect recall . . .
- $(A_5^n)$ $K_i(\phi)$ for all $1 \leq i \leq n$ and tautologies $\phi$.
- $(A_6^n)$ $K_i(\phi \rightarrow \psi) \rightarrow K_i(\phi) \rightarrow K_i(\psi)$ for all $1 \leq i \leq n$ and all $\phi, \psi$.
- $(A_7^n)$ $K_i(\phi) \rightarrow K_j(\phi)$ for all $1 \leq i < j \leq n$. . . . assumptions
- $(A_\infty^n)$ $K_i(\phi)$ for all $1 \leq i \leq n$ and all $\phi$ such that $T_n \models \phi$.

# Formalizing the paradox (1)

For each $n \geq 1$, let $T_n$ be the theory consisting of:

- $(A_1^n)$ $\bigvee_{i=1}^{n} D_i$.
- $(A_2^n)$ $\bigwedge_{1 \leq i < j \leq n} \neg(D_i \wedge D_j)$.
- $(A_3^n)$ $\bigvee_{i=1}^{n}(D_i \wedge \neg K_i(D_i))$.
- $(A_4^n)$ $\bigwedge_{i=1}^{n-1}((\neg D_i) \rightarrow K_{i+1}(\neg D_i))$.
- $(A_5^n)$ $K_i(\phi)$ for all $1 \leq i \leq n$ and tautologies $\phi$. necessitation and
- $(A_6^n)$ $K_i(\phi \rightarrow \psi) \rightarrow K_i(\phi) \rightarrow K_i(\psi)$
  for all $1 \leq i \leq n$ and all $\phi, \psi$.
- $(A_7^n)$ $K_i(\phi) \rightarrow K_j(\phi)$ for all $1 \leq i < j \leq n$.
- $(A_\infty^n)$ $K_i(\phi)$ for all $1 \leq i \leq n$ and all $\phi$ such that $T_n \models \phi$. closure

# Formalizing the paradox (1)

Theorem For any $n \geq 1$, $T_n$ is inconsistent.

Idea By induction on $n$, and starting by showing that $T_n \models \neg D_n$.

Remark No factivity was assumed, or other debatable principles (like $KK$), to derive the contradiction.

# Formalizing the paradox (2)

Surprise Let us extend the notion of surprise, to include the scenario in which the students *know* that the exam is on day $m > n$, while it actually occurs on day $n$.

Remark Intuitively, such an extension is meaningless if we assume factivity and knowledge thereof. Why?

# Formalizing the paradox (2)

For each $n \geq 1$, let $U_n$ be the theory consisting of:

- $A_1^n$, $A_2^n$, $A_4^n$, $A_5^n$, $A_6^n$, $A_7^n$.
- $(A_3^{n\prime})$ $\bigvee_{i=1}^{n}(D_i \wedge \neg K_i(D_i)) \vee \bigvee_{1 \leq i < j \leq n}(D_i \wedge K_i(D_j))$.
- $(A_T^n)$ $K_i(\phi) \rightarrow \phi$ for all $1 \leq i \leq n$ and all $\phi$.
- $(A_\infty^n{}')$ $K_i(\phi)$ for all $1 \leq i \leq n$ and all $\phi$ such that $U_n \models \phi$.

# Formalizing the paradox (2)

For each $n \geq 1$, let $U_n$ be the theory consisting of:

- $A_1^n$, $A_2^n$, $A_4^n$, $A_5^n$, $A_6^n$, $A_7^n$.
- $(A_3^{n\prime})$ $\bigvee_{i=1}^{n}(D_i \wedge \neg K_i(D_i)) \vee \bigvee_{1 \leq i < j \leq n}(D_i \wedge K_i(D_j))$. surprise!
- $(A_T^n)$ $K_i(\phi) \rightarrow \phi$ for all $1 \leq i \leq n$ and all $\phi$.
- $(A_\infty^n{}')$ $K_i(\phi)$ for all $1 \leq i \leq n$ and all $\phi$ such that $U_n \models \phi$.

# Formalizing the paradox (2)

For each $n \geq 1$, let $U_n$ be the theory consisting of:

- $A_1^n$, $A_2^n$, $A_4^n$, $A_5^n$, $A_6^n$, $A_7^n$.
- $(A_3^{n\prime})$ $\bigvee_{i=1}^{n}(D_i \land \neg K_i(D_i)) \lor \bigvee_{1 \leq i < j \leq n}(D_i \land K_i(D_j))$.
- $(A_T^n)$ $K_i(\phi) \rightarrow \phi$ for all $1 \leq i \leq n$ and all $\phi$. factivity
- $(A_\infty^n{}')$ $K_i(\phi)$ for all $1 \leq i \leq n$ and all $\phi$ such that $U_n \models \phi$.

# Formalizing the paradox (2)

Theorem For any $n \geq 1$, $U_n$ is inconsistent.

# Formalizing the paradox (3)

### Relaxing knowledge of factivity

We save factivity, but we assume a weaker form of knowledge-of-factivity: on each day, the students know that they were factive on all earlier days.

# Formalizing the paradox (3)

For each $n \geq 1$, let $(V_n)_0$ be the theory containing:

- $A^n_1$, $A^n_2$, $A^{n\prime}_3$, $A^n_4$, $A^n_5$, $A^n_6$, $A^n_7$.
- $(A^n_T{}')$ $K_j(K_i(\phi) \to \phi)$ for all $1 \leq i < j \leq n$ and all $\phi$.
- $(A^n_\infty{}'')$ $K_i(\phi)$ for all $1 \leq i \leq n$ and all $\phi$ such that $(V_n)_0 \models \phi$.

For each $1 \leq i \leq n$, let $(V_n)^i_0$ be the theory containing:

- $(V_n)_0$.
- $(A^n_{T,i})$ $K_j(\phi) \to \phi$ for all $1 \leq j < i$ and all $\phi$.
- $(A^n_{i,\infty})$ $K_j(\phi)$ for any $1 \leq j \leq i$ and all $\phi$ such that $(V_n)^j_0 \models \phi$.

For each $n$, let $V_n$ be the theory containing:

- $(V_n)^1_0, \ldots, (V_n)^n_0$.
- $(A^n_T)$ $K_i(\phi) \to \phi$ for all $1 \leq i \leq n$ and all $\phi$.

# Formalizing the paradox (3)

For each $n \geq 1$, let $(V_n)_0$ be the theory containing:

- $A_1^n$, $A_2^n$, $A_3^{n}{}'$, $A_4^n$, $A_5^n$, $A_6^n$, $A_7^n$. weak knowledge
- $(A_T^n{}')$ $K_j(K_i(\phi) \to \phi)$ for all $1 \leq i < j \leq n$ and all $\phi$. of factivity
- $(A_\infty^n{}'')$ $K_i(\phi)$ for all $1 \leq i \leq n$ and all $\phi$ such that $(V_n)_0 \models \phi$.

For each $1 \leq i \leq n$, let $(V_n)_0^i$ be the theory containing:

- $(V_n)_0$.
- $(A_{T,i}^n)$ $K_j(\phi) \to \phi$ for all $1 \leq j < i$ and all $\phi$.
- $(A_{i,\infty}^n)$ $K_j(\phi)$ for any $1 \leq j \leq i$ and all $\phi$ such that $(V_n)_0^j \models \phi$.

For each $n$, let $V_n$ be the theory containing:

- $(V_n)_0^1, \ldots, (V_n)_0^n$.
- $(A_T^n)$ $K_i(\phi) \to \phi$ for all $1 \leq i \leq n$ and all $\phi$.

# Formalizing the paradox (3)

Theorem For any $n > 2$, $V_n$ is consistent.

Idea Build a model in which on every day, the students' knowledge consists of the bare minimum required to satisfy $V_n$.

# Considerations about the assumptions

Surprise $\bigvee_{i=1}^{n}(D_i \wedge \neg K_i(D_i))$

Redefined $\bigvee_{i=1}^{n}(D_i \wedge \neg K_i(D_i)) \vee \bigvee_{1 \leq i < j \leq n}(D_i \wedge K_i(D_j))$

- Redefining surprise as we did makes sense if we reformulate the principle of knowledge-of-factivity.
- Our reformulation makes sense because, e.g., if the students know the exam will be on Friday, they will certainly be surprised by a Thursday exam.
- The above is impossible if knowledge is factive, so it's tempting to simplify the definition of surprise – But why should we assume the students themselves would simplify the definition that way?

# Conclusions

- Factivity and knowledge-of-factivity play an important role not deeply investigated in the literature.
- Previous attempts concentrate on other properties ($KK$), principles (retention), alternative formulations not related to knowledge (based on provability).
- We showed that by weakening knowledge-of-factivity and by not simplifying the definition of surprise accordingly, the surprise exam paradox vanishes – *We propose this as a solution to the paradox*.
- Future work: our solution may be generalized to work with other situations.

# Thank you!

Questions?