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

Alessandro Aldini[1][0000−0002−7250−5011]
Samuel Allen Alexander[2][0000−0002−7930−110X]
Pierluigi Graziani[3][0000−0002−8828−8920]

[1]University of Urbino, alessandro.aldini@uniurb.it
[2]The U.S. Securities and Exchange Commission, samuelallenalexander@gmail.com
[3]University of Urbino, pierluigi.graziani@uniurb.it

Abstract. Fitch's Paradox and the Paradox of the Knower both make use of the *Factivity Principle*. The latter also makes use of a second principle, namely *the Knowledge-of-Factivity Principle*. Both the principle of factivity and the knowledge thereof have been the subject of various discussions, often in conjunction with a third principle known as *Closure*. In this paper, we examine the well-known *Surprise Examination paradox* considering both the principles on which this paradox rests and some formal characterisations of the surprise notion, crucial in this paradox. Standard formalizations of the Surprise Examination paradox in modal logic do not seem, at first glance, to depend on either factivity or knowledge-of-factivity, but we will argue that both factivity and knowledge-of-factivity play a key implicit role in the paradox. Namely, they are implicitly, perhaps unintentionally, used in order to simplify the definition of surprise. We analyze modal logical formalizations of three versions of the paradox concluding that the Surprise Examination paradox is the result of two flaws: the assumption of knowledge-of-factivity, and the over-simplification of the definition of "surprise" accordingly. By fixing these two flaws, the Surprise Examination paradox vanishes.

1 Introduction

Many epistemic paradoxes are based on the *Factivity Principle*, which says that if p is known, then p is true. For example, Fitch's Paradox and the Paradox of the Knower both make use of this principle. The latter also makes use of a second principle, namely: it is known that if p is known then p is true. We call this second principle *Knowledge-of-factivity*. Both the principle of factivity and the knowledge thereof have been the subject of various discussions [21], often in conjunction with a third principle known as *Closure*, i.e., if C is provable from a set of premises, and those premises are known, then C is known [15]. In this paper, we examine the well-known *Surprise Examination paradox* (see for example [13, 16, 10, 6, 20, 8] for thorough surveys of the literature on this paradox). Standard formalizations of this paradox in modal logic do not seem, at first glance, to depend on either factivity or knowledge-of-factivity, however both principles play a key implicit role in the paradox.[1] Namely, they are implicitly, perhaps unintentionally, used in order to simplify the definition of the core notion of surprise. In standard modal logical formalizations of the paradox, students are said to be "surprised" if and only if, just prior to the occurrence of the weekly surprise exam, the students do not know that the exam will occur that day. Certainly this is a sufficient condition for the students to be surprised, but we argue it should not (at least by the students) be considered a necessary condition. We argue the students' definition of surprise should also include another disjunct: if the surprise exam occurs on day n and, just prior to its occurrence, the students know that the surprise exam will occur on day m (where $m > n$), this also should count as an instance of surprise. Such a situation is impossible assuming factivity, and thus, if we (consciously or unconsciously) assume the students know their own factivity, then the students know the additional disjunct is false; this seemingly justifies the simpler definition of surprise. We will analyze modal logical formalizations of three versions of the paradox. The first (standard) version uses the simplified definition of surprise, and a contradiction is achieved even without assuming factivity or knowledge-of-factivity. The second version uses the modified definition of surprise, and a contradiction is achieved assuming both factivity and knowledge-of-factivity. The third version uses the modified definition of surprise, and it assumes factivity, but it only assumes a weaker form of knowledge-of-factivity, namely, that on each day, the students know that they were factive on all earlier days. By constructing a model, we prove that (if the school week has at least 3 days) this third formalization does not lead to a contradiction. Thus in our opinion the Surprise Examination paradox is the result of two flaws: the assumption of knowledge-of-factivity, and the over-simplification of the definition of "surprise" accordingly. By fixing these two flaws, the Surprise Examination paradox vanishes.

The rest of the paper is organized as follows: in Section 2 we introduce the Surprise Examination paradox. In Section 3 we describe a standard modal logic formalization of the paradox in which a contradiction is achieved. In Section 4 we introduce a version of the paradox with surprise re-defined, in which case a contradiction is achieved as well if we assume factivity and knowledge of factivity. In Section 5, we discuss a new resolution to the paradox. We show that by re-defining surprise as in the previous section and weakening knowledge-of-factivity (while still requiring factivity), the Surprise Examination paradox disappears. In Section 6, we address whether knowledge-of-factivity should be assumed. We conclude with remarks on the obtained results and possible future work.

2 Surprise Examination paradox

The paradox discussed in this article has many names and many variants, namely the unexpected hanging, the unexpected tiger, the prediction paradox, etc., and

[1] See also [13] and [21].

although it has often been underestimated as a topic, many scholars have devoted attention to it by exploring its possible solutions and/or criticality. So correctly Michael Scriven [19] wrote “a new and powerful paradox has come to light”. The paradox was first circulated by word of mouth in the early 1940’s and today *PhilPapers* has more than 1000 articles on this topic. The more common version of the surprise examination paradox goes as follows:

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.

As John Earman remarked [8] there are three mutually reinforcing reasons for the longevity of the surprise exam paradox:

> One is that the paradox resonates with a number of other paradoxes including the liar, sorities, Moore’s paradox, the lottery paradox. Second, the surprise exam is a kind of Rorschach test for philosophy. Logicians see it is an opportunity to display their wares —including Gödel’s incompleteness theorems (e.g. Ardeshir and Ramezanian 2012, Chow 1998, Fitch 1964, Halpern and Moses 1968, and Kritchman and Raz 2010). Epistemologists see it as an opportunity to explore the concepts of knowledge and justified belief (e.g. Sorensen 1982, 1984, 1988, and 2017). Still others see it as a hybrid of logical and epistemological issues (e.g. Kaplan and Montague 1960). Third, the variety of reactions to this Rorschach test is fueled by the fact that the surprise exam announcement is a misnomer: there are multiple ways of reading the announcement, and the resulting paradoxes, if any, call for resolutions that may differ from reading to reading.

In the following, we will analyze three epistemic modal logic formalizations of the paradox and consider the role that the assumptions of factivity and knowledge of factivity play in these reformulations.

3 The Surprise Examination paradox in modal logic

In this paper we consider a simple propositional epistemic logic to formalize the Surprise Examination paradox, because its schematic characterization of knowledge and its logical machinery allow us to tackle clearly the issues raised by the paradox and, in particular, those related to the properties of (ideal) knowers. The paradox is set out in such a formal setting in several works, see for example [5, 17, 11, 18]. In particular, we introduce propositional variables $D_1, \ldots, D_n$ for the n days of the school week, each D_i being thought of as *the exam takes place on day i*. Moreover, in order to take into proper consideration the dynamic nature of the paradox, we use a notion of knowledge related explicitly to time and occurences of subsequent events.

Generalized and more expressive forms of knowledge are obtained by enriching the classical modal operator K with parameters expressing, e.g., the agent or the instant of time under consideration when evaluating knowledge of a given formula [4]. In our setting, it is particularly convenient to reason about the knowledge of students at specific days. Thus, we consider a formalization of knowledge in which the epistemic operator K is indexed. While in the tradition of epistemic logic K_i refers to what agent i knows, in this paper we let the modal operator K_i denote the knowledge of the subjects at the time of event e_i. In particular, in our setting $K_i(\phi)$ is going to be read as: "ϕ is known by the students at midnight just before day i". Such an interpretation is adopted similarly in [17, 18]. Thus, for example, the formula $D_2 \to K_2(\neg D_1)$ might be read: *On midnight just before day 2, if the exam is on such a day, the students know that the exam was not on day 1.*

When dealing with agents and time it is worth setting some assumptions that justify the properties of our family of epistemic operators (we refer to [9] for an overview of the following considerations).

The propositions (and formulas) that we use to describe our problem are *stable*, so that their truth values do not change over time.

Moreover, our formal system is not explicitly multi-agent. This means that we will not model the teacher and each student as separate entities. On one hand, the truth values of the propositions D_i will express the teacher's decision. To this aim, we assume that the teacher is not a liar when announcing that there will be a (unique) surprise exam next week. On the other hand, the classroom of students is considered as the unique entity to which the knowledge operators refer. In other words, $K_i(\phi)$ expresses that the whole classroom of students know ϕ at the time of event e_i. By the way, such an assumption implies that our system is synchronous: time (and hence the passage of time) is common knowledge, i.e., all the students have somehow access to a shared clock and their knowledge is aligned.

As another assumption, the students are *perfect recall* agents. Their knowledge might grow over time to reflect that new knowledge can be acquired, while still keeping track of old knowledge. To clarify, it is reasonable to assume that the students do not forget what happened in the previous days of the week.

Based on such assumptions, we are now ready to discuss intuitively the most important properties that we consider when formalizing the Surprise Examination paradox. This is typically done by stating a list of axioms, all of which seem quite plausible based on the scenario of a teacher announcing a surprise examination.

We start with the formalization of "surprise". The fact that the exam will be a surprise will be modeled by the disjunction $\bigvee_i (D_i \wedge \neg K_i(D_i))$. Therefore, for some i the following is true: the exam occurs on day i but the students do not know (at midnight before the exam) that the exam occurs on that day. Certainly, if this is true, then we should consider the students to be surprised. For example, if $D_3 \wedge \neg K_3(D_3)$ holds, then that means the exam is on day 3, but at midnight before day 3, the students do not know the exam is on day 3.

The fact that there will be an exam can be captured by the axiom $\bigvee_i D_i$, while the fact that the exam will only fall on one unique day is captured by $\bigwedge_{i<j} \neg(D_i \wedge D_j)$. These two axioms, together, express that the teacher is not a liar.

The stability and perfect recall assumptions allow us to state two more properties concerning knowledge. On the one hand, the fact that as days go without the exam taking place the students refine their knowledge, is captured by $\bigwedge_i ((\neg D_i) \rightarrow K_{i+1}(\neg D_i))$. On the other hand, the fact that the acquired knowledge is not forgotten is captured by $K_i(\phi) \rightarrow K_j(\phi)$ for $j > i$ (this is also called the *retention principle* [17, 18]).

In addition to these properties, which are specific to the given problem, we will also assume a minimum set of standard properties of knowledge that will be used in the proofs, like, e.g., the fact that knowledge is closed under implication. However, we will exclude a property that has always been controversial in the history of the paradox, that is the KK principle, stating that if the students know ϕ, then they know that they know ϕ. We will show that the paradox arises even in the absence of such a general condition of positive introspection, contrary to [17, 22] and other authors who argued that KK was the cause of the paradox (see also [18]).

But before introducing the formal system, let us consider the definition of surprise more closely. We argued that $\bigvee_i (D_i \wedge \neg K_i(D_i))$ seems to imply surprise. What about the converse? Are there any other ways the students could be surprised, not included in this disjunction? It seems to us that there is another way the students could be surprised. If the students know the exam will be on Friday, they will be very surprised indeed if the exam is on Thursday. Thus, a more inclusive definition for surprise would be as follows. The students are surprised if the disjunction

$$\bigvee_i (D_i \wedge \neg K_i(D_i)) \vee \bigvee_{i<j} (D_i \wedge K_i(D_j))$$

holds. But is the second disjunct above even possible? Knowledge is supposed to be factive, in other words, truthful. If the students know the exam is on Friday, then the exam cannot be on Thursday—that would violate the truthfulness of the students' knowledge. Since knowledge is factive, the two definitions of surprise are equivalent. However, there remains a much deeper question. We ourselves, as outside observers, know that knowledge is factive and therefore that the two definitions of surprise are equivalent. But do the students themselves know that? We should neither assume the students know the two definitions are equivalent, nor that they know themselves to be factive. We discuss this more in Section 6, but in short: the students cannot predict the teacher won't announce contradictions in future, thus *to them*, K_i should be thought of as a "provable-from-teacher" (not a "knowledge") operator. If *we* know the teacher is truthful, then *we* know that said provability is in fact knowledge (hence our choice of the letter K), but that doesn't imply the students know that.

We will show that by redefining surprise in the above way, and weakening knowledge-of-factivity (while still requiring factivity), the Surprise Examination paradox disappears.

In the next section we will specify the details of the semantics we use in this paper, but in short, we use propositional semantics, treating purely-modal formulas $K_i(\phi)$ like propositional atoms.

3.1 Formalizing the paradox

Based on the motivations and intuitions surveyed above, we will state a theory containing formalized versions of the assumptions of the Surprise Examination paradox. But first, we will define the logic we are using.

Definition 1. *We work in the language $\mathscr{L}$ consisting of propositional atoms $D_1, D_2, \ldots$ and modal operators $K_1, K_2, \ldots$, whose syntax and semantics are as follows.*

- *Formulas of $\mathscr{L}$ are defined by induction as follows:*
  1. *Every D_i $(i = 1, 2, \ldots)$ is a formula.*
  2. *For every formula ϕ and every $i = 1, 2, \ldots$, $K_i(\phi)$ is a formula.*
  3. *Whenever ϕ and ψ are formulas, so are $\neg\phi$, $\phi \wedge \psi$, $\phi \vee \psi$, and $\phi \rightarrow \psi$.*
- *By the* basic formulas of *$\mathscr{L}$ we mean formulas of the form D_i or $K_i(\phi)$.*
- *By a* model *we mean an assignment of truth-values to the basic formulas of $\mathscr{L}$.*
- *If $\mathscr{M}$ is a model and ϕ is a formula, we define the truth-value of ϕ in $\mathscr{M}$, writing $\mathscr{M} \models \phi$ if that truth-value is True or $\mathscr{M} \not\models \phi$ if that truth-value is False, as follows:*
  1. *If ϕ is a basic formula of $\mathscr{L}$ then $\mathscr{M} \models \phi$ iff $\mathscr{M}$ assigns truth value True to ϕ.*
  2. *$\mathscr{M} \models \neg\phi$ iff $\mathscr{M} \not\models \phi$.*
  3. *$\mathscr{M} \models \phi \wedge \psi$ iff $\mathscr{M} \models \phi$ and $\mathscr{M} \models \psi$.*
  4. *$\mathscr{M} \models \phi \vee \psi$ iff $\mathscr{M} \models \phi$ or $\mathscr{M} \models \psi$.*
  5. *$\mathscr{M} \models \phi \rightarrow \psi$ iff $\mathscr{M} \not\models \phi$ or $\mathscr{M} \models \psi$.*
- *A* theory *is a set of formulas.*
- *For any model $\mathscr{M}$ and theory T, $\mathscr{M} \models T$ means $\mathscr{M} \models \phi$ for all $\phi \in T$.*
- *A theory T is* consistent *if there is some model $\mathscr{M}$ such that $\mathscr{M} \models T$; otherwise T is* inconsistent.
- *For any theory T and formula ϕ, $T \models \phi$ means that for every model $\mathscr{M}$, if $\mathscr{M} \models T$ then $\mathscr{M} \models \phi$.*
- *For any theories T_1, T_2, $T_1 \models T_2$ means $T_1 \models \phi$ for all $\phi \in T_2$.*

- *A* tautology *is a formula ϕ such that $\emptyset \models \phi$.*

Thus, e.g., $K_1(D_1) \to K_1(D_1)$ is a tautology, but $K_1(D_1 \to D_1)$ is not.

Since our semantics are propositional, we have the usual completeness result for propositional logic, namely:

Lemma 1. *(Completeness) For any theory T and any formula ϕ, $T \models \phi$ if and only if there exist finitely many $\phi_1, \ldots, \phi_n \in T$ such that $\phi_1 \to \cdots \to \phi_n \to \phi$ is a tautology.*

We also make use of shorthands such as $\bigwedge_{i=1}^n D_i$ for $D_1 \wedge \cdots \wedge D_n$, $\bigvee_{i=1}^n \phi_i$ for $\phi_1 \vee \cdots \vee \phi_n$, and so on. These are not new symbols in $\mathscr{L}$, they are simply meta-symbols. In every case, it will be clear what the actual $\mathscr{L}$-formulas denoted by them are.

The following theory is intended to capture the standard assumptions in the Surprise Examination paradox (for a week with n school-days).

Definition 2. *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 ϕ.*
- *(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 ϕ, ψ.*
- *(A_7^n) $K_i(\phi) \to K_j(\phi)$ for all $1 \leq i < j \leq n$.*
- *(A_∞^n) $K_i(\phi)$ for all $1 \leq i \leq n$ and all ϕ such that $T_n \models \phi$.*

As previously stated informally, A_1^n and A_2^n express the truthfulness of the announcement made by the teacher, A_3^n formalizes the idea of surprise, A_4^n and A_7^n express properties of knowledge that hold by virtue of the stability and perfect recall assumptions. Moreover, we have three more axioms expressing standard properties of knowledge: A_5^n formalizes the necessitation principle (all tautologies are known), A_6^n states that knowledge is closed under logical consequence, and A_∞^n is the classical closure axiom expressing that what can be derived is also known [15], [7]. In the following, after a preliminary Lemma, we prove the contradiction underlying the paradox.

Lemma 2. *(Closure Lemma) Let $n \geq 1$, $1 \leq i \leq n$. Suppose T is any $\mathscr{L}$-theory such that:*

- *T includes A_5^n and A_6^n.*
- *T includes $K_i(\phi)$ whenever T includes ϕ.*

Then:

(1) For any ϕ, if $T \models \phi$, then $T \models K_i(\phi)$.
(2) For any $\phi_1, \ldots, \phi_\ell$ and ϕ, if $\phi_1 \to \cdots \to \phi_\ell \to \phi$ is a tautology, and if $T_n \models \phi_j$ for all $1 \leq j \leq \ell$, then $T_n \models K_i(\phi)$.

Proof. (1) Assume $T \models \phi$. Then there are finitely many $\phi_1, \dots, \phi_\ell \in T$ such that $\phi_1 \to \cdots \to \phi_\ell \to \phi$ is a tautology. By A_5^n,

$$T \models K_i(\phi_1 \to \cdots \to \phi_\ell \to \phi).$$

By repeated application of A_6^n,

$$T \models K_i(\phi_1) \to \cdots \to K_i(\phi_\ell) \to K_i(\phi).$$

By assumption, since T contains $\phi_1, \dots, \phi_\ell$, T contains $K_i(\phi_1), \dots, K_i(\phi_\ell)$. Thus $T \models K_i(\phi)$.

(2) Since $T_n \models \phi_j$ for all $1 \leq j \leq \ell$, by (1) we see $T_n \models K_i(\phi_j)$ for all $1 \leq j \leq \ell$. The rest of the proof is similar to the proof of (1). □

Theorem 1. *(The Surprise Examination paradox) For any $n \geq 1$, T_n is inconsistent.*

Proof. By induction on n. The base case $n = 1$ is trivial because $A_1^1 \equiv D_1$, thus A_∞^1 includes $K_1(D_1)$, and $A_3^1 \equiv D_1 \to \neg K_1(D_1)$.

Assume $n > 1$.

Preliminary claim: $T_n \models \neg D_n$. To see this, we reason within T_n as follows:

- Assume D_n.
- From A_2^n it follows that $(\neg D_1) \wedge \cdots \wedge (\neg D_{n-1})$.
- From A_4^n it follows that $K_2(\neg D_1) \wedge \cdots \wedge K_n(\neg D_{n-1})$.
- From A_7^n it follows that $K_n(\neg D_1) \wedge \cdots \wedge K_n(\neg D_{n-1})$.
- By A_5^n, $K_n((\bigvee_{i=1}^n D_i)) \to \neg D_1 \to \cdots \to \neg D_{n-1} \to D_n)$.
- By repeated usages of A_6^n, it follows from the previous bullet that $K_n(\bigvee_{i=1}^n D_i) \to K_n(\neg D_1) \to \cdots \to K_n(\neg D_{n-1}) \to K_n(D_n)$.
- By A_∞^n and A_1^n, it follows that $K_n(\bigvee_{i=1}^n D_i)$.
- From the previous four bullets it follows that $K_n(D_n)$.
- By A_3^n, $\bigvee_{i=1}^n (D_i \wedge \neg K_i(D_i))$.
- Since $(\neg D_1) \wedge \cdots \wedge (\neg D_{n-1})$, the previous bullet implies $\neg K_n(D_n)$.
- Contradiction. Discharge assumption and conclude $\neg D_n$.

This proves the preliminary claim.

To finish the proof, it will suffice to show $T_n \models T_{n-1}$, since T_{n-1} is inconsistent by induction. For this, it suffices to prove that whenever $T_{n-1} \models \phi$, then $T_n \models \phi$. We prove this by induction on the number of applications of A_∞^{n-1} needed to prove $T_{n-1} \models \phi$.

Case A_1^{n-1}: ϕ is $\bigvee_{i=1}^{n-1} D_i$. Then $T_n \models \phi$ by A_1^n plus the preliminary claim.

Case A_2^{n-1}: ϕ is $\bigwedge_{1 \leq i < j \leq n-1} \neg(D_i \wedge D_j)$. Then $T_n \models \phi$ by A_2^n.

Case A_3^{n-1}: ϕ is $\bigvee_{i=1}^{n-1} (D_i \wedge \neg K_i(D_i))$. By A_3^n, $T_n \models \bigvee_{i=1}^n (D_i \wedge \neg K_i(D_i))$. By the preliminary claim, $T_n \models \neg D_n$. It follows that $T_n \models \phi$.

Case A_4^{n-1}: ϕ is $\bigwedge_{i=1}^{n-2} (\neg D_i) \to K_{i+1}(\neg D_i)$. Then $T_n \models \phi$ by A_4^n.

Cases A_5^{n-1}, A_6^{n-1}, A_7^{n-1}: similar to the previous cases, the results follow by A_5^n, A_6^n, and A_7^n, respectively.

Case A_∞^{n-1}: ϕ is $K_i(\psi)$ for some $1 \leq i \leq n-1$ and some ψ such that $T_{n-1} \models \psi$. Since $T_{n-1} \models \psi$, there are $\psi_1, \ldots, \psi_\ell \in T_{n-1}$ such that $\psi_1 \to \cdots \to \psi_\ell \to \psi$ is a tautology and such that for all $1 \leq j \leq \ell$, $T_{n-1} \models \psi_j$ can be proven using fewer applications of A_∞^{n-1} than are needed to prove $T_{n-1} \models \phi$. Thus by induction, $T_n \models \psi_j$ for each $1 \leq j \leq \ell$. By Lemma 2 (part 2), $T_n \models K_i(\psi)$. □

4 Redefining the surprise axiom

As suggested in Section 2, we now consider a variant of our formal system in which the axiom modeling surprise is re-defined by adding a disjunct. The following definition states such a variant.

Definition 3. *For each $n \geq 1$, by U_n we mean the theory consisting of the following axioms:*

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

Note that A_3^n is replaced by the more inclusive property we discussed above, called $A_3^n{}'$. As we mentioned, such an extension is actually equivalent to the original axiom if we assume truthfulness of knowledge and knowledge of such a factivity principle. Hence, under this hypothesis, the paradox is not actually solved. To state this result formally, the system we consider combines the new version of surprise $A_3^n{}'$, the factivity axiom (see A_T^n, which states the truthfulness of knowledge), and, by virtue of $A_\infty^n{}'$, also the knowledge of such a factivity.

Theorem 2. *(A modified Surprise Examination paradox) For any $n \geq 1$, U_n is inconsistent.*

Proof. Since Theorem 1 says T_n is inconsistent, it will suffice to show $U_n \models T_n$. For this, it suffices to prove that whenever $T_n \models \phi$, then $U_n \models \phi$. We prove this by induction on the number of applications of T_∞^n needed to prove $T_n \models \phi$.

Trivial case: ϕ is an instance of A_1^n, A_2^n, A_4^n, A_5^n, A_6^n, or A_7^n. Then $U_n \models \phi$ since U_n includes these axioms too.

Case A_3^n: ϕ is $\bigvee_{i=1}^n (D_i \wedge \neg K_i(D_i))$. By $A_3^n{}'$, $U_n \models \bigvee_{i=1}^n (D_i \wedge \neg K_i(D_i)) \vee \bigvee_{1 \leq i < j \leq n} (D_i \wedge K_i(D_j))$. To show $U_n \models \phi$, it suffices to show that for all $1 \leq i < j \leq n$, $U_n \models \neg(D_i \wedge K_i(D_j))$. Fix $1 \leq i < j \leq n$. We reason within U_n as follows:

- Assume $D_i \wedge K_i(D_j)$.
- By A_2^n, it follows that $\neg D_j$.
- By A_T^n, $K_i(D_j) \to D_j$.
- Contradiction. Discharge assumption and conclude $\neg(D_i \wedge K_i(D_j))$.

Case A_∞^n: ϕ is $K_i(\psi)$ for some $1 \leq i \leq n$ and some ψ such that $T_n \models \psi$. Since $T_n \models \psi$, there are $\psi_1, \ldots, \psi_\ell \in T_n$ such that $\psi_1 \to \cdots \to \psi_\ell \to \psi$ is a tautology and such that for all $1 \leq j \leq \ell$, $T_n \models \psi_j$ can be proved using fewer applications of A_∞^n than are needed to prove $T_n \models \phi$. By induction, $U_n \models K_i(\psi_j)$ for all $1 \leq j \leq \ell$. By Lemma 2 (part 2), $U_n \models K_i(\psi)$. □

5 A resolution to the paradox

Apparently, extending the notion of surprise alone does not bring benefits. This is true if a general notion of knowledge of factivity is assumed. For instance, in the system of the previous section we can derive forms like, e.g., $K_1(K_2(\phi) \to \phi)$. Generally speaking, at any time the factivity of any knowledge – past, present, or future – is known. However, we argue that if we limit the knowledge of factivity, then we obtain a system in which the paradox disappears.

Formally, in order to limit the knowledge of factivity, we restrict axiom A^n_T by assuming that only the factivity of the knowledge of past events is known, thus obtaining a new axiom $A^n_T{}'$.

Definition 4. *For each $n \geq 1$, by $(V_n)_0$ we mean the theory containing the following axioms:*

- A^n_1, A^n_2, $A^n_3{}'$, 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 ϕ.*
- *($A^n_\infty{}''$) $K_i(\phi)$ for all $1 \leq i \leq n$ and all ϕ such that $(V_n)_0 \models \phi$.*

For each $1 \leq i \leq n$, by $(V_n)^i_0$ we mean the theory containing the following axioms:

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

For each n, by V_n we mean 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 ϕ.*

The inductive definition of the theory V_n preserves the condition stating that $K_j(K_i(\phi) \to \phi)$ holds for all $j > i$, i.e., simply put, the students become aware tomorrow of the factivity of what they know today. Such an inductive definition will allow us to prove by induction that V_n is consistent (if $n > 2$), thus making the paradox disappear. We will show V_n is consistent by constructing a model, i.e., an assignment of truth values to the basic formulas of $\mathscr{L}$ (see Definition 1), and showing that that model satisfies V_n.

Lemma 3. *For any n, for any $1 \leq i < j \leq n$, $(V_n)^i_0 \subseteq (V_n)^j_0$.*

Proof. By inspection. □

Theorem 3. *For every $n > 2$, V_n is consistent.*

Proof. The intuitive idea is that we will construct a model in which on each day, the students' knowledge consists of the bare minimum required to satisfy V_n, namely, exactly those facts which V_n requires them to know on that day, and nothing else. We will then verify that the resulting model satisfies all the required axioms, and this will be mostly straightforward, with the exception of $A^n_3{}'$, for which we will have to construct another model (see below).

Since $n > 2$, we may fix some $1 \leq m < n - 1$. For each $1 \leq i \leq n$, let W_i be the theory containing the following axioms:

- $(V_n)^i_0$.
- $\bigwedge_{1\leq j<i, j\neq m} \neg D_j$.

By Lemma 3 it follows that whenever $1 \leq i < j \leq n$, $W_j \models W_i$.

We define a model $\mathscr{M}$ as follows:

- $\mathscr{M} \models D_m$.
- For all $i \neq m$, $\mathscr{M} \not\models D_i$.
- For all $1 \leq i \leq n$, for all ϕ, $\mathscr{M} \models K_i(\phi)$ iff $W_i \models \phi$.

To show V_n is consistent, it suffices to show that $\mathscr{M} \models V_n$.

Claim 1: For each $1 \leq p \leq n$, $\mathscr{M} \models W_p$. We prove this by induction on p. We will show $\mathscr{M} \models \phi$ for all $\phi \in W_p$. Fix any such ϕ.

Case A^n_1: ϕ is $\bigvee_{i=1}^n D_i$. Then clearly $\mathscr{M} \models \phi$.

Case A^n_2: ϕ is $\bigwedge_{1\leq i<j\leq n} \neg(D_i \wedge D_j)$. Then clearly $\mathscr{M} \models \phi$.

Case $A^{n\prime}_3$: ϕ is $\bigvee_{i=1}^n (D_i \wedge \neg K_i(D_i)) \vee \bigvee_{1\leq i<j\leq n}(D_i \wedge K_i(D_j))$. To show $\mathscr{M} \models \phi$, it suffices to show $\mathscr{M}$ satisfies any one of the disjuncts. We will show $\mathscr{M} \models D_m \wedge \neg K_m(D_m)$. By construction $\mathscr{M} \models D_m$. It remains to show $\mathscr{M} \models \neg K_m(D_m)$. In other words, we must show $W_m \not\models D_m$. We will construct a model $\mathscr{N}$ such that $\mathscr{N} \models W_m$ and $\mathscr{N} \not\models D_m$.

The intuitive idea is that in $\mathscr{N}$, the exam will occur on day $n-1$, the students will initially (before day m) know the bare minimum that W_m requires them to know, but, starting on day m, the students' knowledge will become inconsistent: from that day on, they will know everything (including incorrectly knowing that the exam will occur on day m). The fact that the students are not factive in $\mathscr{N}$ (on days $\geq m$) is not problematic: the purpose of $\mathscr{N}$ is not to, itself, directly satisfy V_n (which requires factivity), but only to show that $W_m \not\models D_m$.

Let $\mathscr{N}$ be the model defined by:

- $\mathscr{N} \models D_{n-1}$.
- For all $i \neq n-1, \mathscr{N} \not\models D_i$.
- For all $1 \leq j < m$ and all ϕ, $\mathscr{N} \models K_j(\phi)$ iff $W_j \models \phi$.
- For all $m \leq j \leq n$ and all ϕ, $\mathscr{N} \models K_j(\phi)$.

Since $m < n-1$, $\mathscr{N} \not\models D_m$. We claim $\mathscr{N} \models W_m$. We will prove more, for the sake of a stronger induction hypothesis: we will prove by induction on q that $\mathscr{N} \models W_q$ for all $q \leq m$. Let $\psi \in W_q$.

Subcase $A^n_1(W_q)$: ψ is $\bigvee_{i=1}^n D_i$. Then clearly $\mathscr{N} \models \psi$.

Subcase $A^n_2(W_q)$: ψ is $\bigwedge_{1\leq i<j\leq n} \neg(D_i \wedge D_j)$. Then clearly $\mathscr{N} \models \psi$.

Subcase $A^{n\prime}_3(W_q)$: ψ is $\bigvee_{i=1}^n (D_i \wedge \neg K_i(D_i)) \vee \bigvee_{1\leq i<j\leq n}(D_i \wedge K_i(D_j))$. To show $\mathscr{N} \models \psi$, it suffices to show $\mathscr{N}$ satisfies any one of the disjuncts. We will show $\mathscr{N} \models D_{n-1} \wedge K_{n-1}(D_n)$. By construction $\mathscr{N} \models D_{n-1}$. And $\mathscr{N} \models K_{n-1}(D_n)$ because $m \leq n-1 \leq n$ (so K_{n-1} is defined in the 4th bullet of the definition of $\mathscr{N}$).

Subcase $A^n_4(W_q)$: ψ is $\bigwedge_{i=1}^{n-1}((\neg D_i) \rightarrow K_{i+1}(\neg D_i))$. Fix any $1 \leq i \leq n-1$, we will show $\mathscr{N} \models (\neg D_i) \rightarrow K_{i+1}(\neg D_i)$. If $i+1 < m$, then, since W_{i+1} contains $\bigwedge_{1\leq j<i+1, j\neq m} \neg D_j$, in particular (using $j = i$), $W_{i+1} \models \neg D_i$. Thus

$\mathscr{N} \models K_{i+1}(\neg D_i)$ by bullet 3 in the definition of $\mathscr{N}$. On the other hand, if $i+1 \geq m$, then $\mathscr{N} \models K_{i+1}(\neg D_i)$ by bullet 4 in the definition of $\mathscr{N}$.

Subcase $A_5^n(W_q)$: ψ is $K_i(\rho)$ for some $1 \leq i \leq n$ and some tautology ρ. If $i < m$ then $W_i \models \rho$ (because ρ is a tautology) and thus $\mathscr{N} \models K_i(\rho)$ (by bullet 3 in the definition of $\mathscr{N}$). On the other hand, if $i \geq m$, then $\mathscr{N} \models K_i(\rho)$ by bullet 4 in the definition of $\mathscr{N}$.

Subcase $A_6^n(W_q)$: ψ is $K_i(\rho \to \tau) \to K_i(\rho) \to K_i(\tau)$ for some $1 \leq i \leq n$ and some ρ, τ. If $i \geq m$ then $\mathscr{N} \models K_i(\tau)$ by bullet 4 in the definition of $\mathscr{N}$. But assume $i < m$. Assume $\mathscr{N} \models K_i(\rho \to \tau)$ and $\mathscr{N} \models K_i(\rho)$. By bullet 3 in the definition of $\mathscr{N}$ this means $W_i \models \rho \to \tau$ and $W_i \models \rho$. Thus $W_i \models \tau$, so $\mathscr{N} \models K_i(\tau)$.

Subcase $A_7^n(W_q)$: ψ is $K_i(\rho) \to K_j(\rho)$ for some $1 \leq i < j \leq n$. If $j \geq m$ then $\mathscr{N} \models K_j(\rho)$ by bullet 4 in the definition of $\mathscr{N}$. But assume $j < m$. Assume $\mathscr{N} \models K_i(\rho)$. By bullet 3 in the definition of $\mathscr{N}$, $W_i \models \rho$. Since $i < j$, we have $W_j \models W_i$, thus $W_j \models \rho$, so $\mathscr{N} \models K_j(\rho)$.

Subcase $A_T^n{}'(W_q)$: ψ is $K_j(K_i(\rho) \to \rho)$ for some $1 \leq i < j \leq n$ and some ρ. If $j \geq m$ then $\mathscr{N} \models K_j(K_i(\rho) \to \rho)$ by bullet 4 in the definition of $\mathscr{N}$. But assume $j < m$. By $A_{T,j}^n$, $(V_n)_0^j \models K_i(\rho) \to \rho$, thus $W_j \models K_i(\rho) \to \rho$ since W_j includes $(V_n)_0^j$. Thus $\mathscr{N} \models K_j(K_i(\rho) \to \rho)$ by bullet 3 in the definition of $\mathscr{N}$.

Subcase $A_\infty^n{}''(W_q)$: ψ is $K_i(\rho)$ for some $1 \leq i \leq n$ and some ρ such that $(V_n)_0 \models \rho$. If $i \geq m$ then $\mathscr{N} \models K_i(\rho)$ by bullet 4 in the definition of $\mathscr{N}$. But assume $i < m$. Since $(V_n)_0 \subseteq (V_n)_0^i \subseteq W_i$, we have $W_i \models \rho$ and thus $\mathscr{N} \models K_i(\rho)$ by bullet 3 in the definition of $\mathscr{N}$.

Subcase $A_{T,q}^n(W_q)$: ψ is $K_j(\rho) \to \rho$ for some $1 \leq j < q$. Assume $\mathscr{N} \models K_j(\rho)$. Since $j < q \leq m$, bullet 3 in the definition of $\mathscr{N}$ says $W_j \models \rho$. Since $j < q$, we can finally use our strong q-induction hypothesis: by induction, $\mathscr{N} \models W_j$. Thus $\mathscr{N} \models \rho$.

Subcase $A_{q,\infty}^n(W_q)$: ψ is $K_j(\rho)$ for some $1 \leq j \leq q$ and some ρ such that $(V_n)_0^j \models \rho$. If $j \geq m$ then $\mathscr{N} \models K_j(\rho)$ by bullet 4 in the definition of $\mathscr{N}$. But assume $j < m$. Since $(V_n)_0^j \models \rho$ and $(V_n)_0^j \subseteq W_j$, we see $W_j \models \rho$ and thus $\mathscr{N} \models K_j(\rho)$ by bullet 3 in the definition of $\mathscr{N}$.

Subcase $\bigwedge_{1 \leq j < q, j \neq m} \neg D_j$ (W_q): ψ is $\bigwedge_{1 \leq j < q, j \neq m} \neg D_j$. If $q = 1$ then the conjunction is empty, so $\mathscr{N} \models \psi$ vacuously. Assume $q > 1$. Let $1 \leq j < q$, $j \neq m$. Since $q \leq m < n-1$, it follows that $j < n-1$, so $\mathscr{N} \models \neg D_j$ by bullet 2 in the definition of $\mathscr{N}$.

This concludes the proof that $\mathscr{N} \models W_q$ for all $q \leq m$. In particular, $\mathscr{N} \models W_m$. Since $\mathscr{N} \not\models D_m$, this concludes the proof that $W_m \not\models D_m$. This concludes the proof that $\mathscr{M} \models \neg K_m(D_m)$. Since $\mathscr{M} \models D_m$, this concludes the proof that $\mathscr{M} \models D_m \wedge \neg K_m(D_m)$. This concludes Case $A_3^n{}'$.

Case A_4^n: ϕ is $\bigwedge_{i=1}^{n-1}((\neg D_i) \to K_{i+1}(\neg D_i))$. Fix $1 \leq i \leq n-1$ and assume $\mathscr{M} \models \neg D_i$. Since $\mathscr{M} \models D_m$, this implies $i \neq m$. Thus, since W_{i+1} includes $\bigwedge_{1 \leq j < i+1, j \neq m} \neg D_j$, we see $W_{i+1} \models \neg D_i$. Thus $\mathscr{M} \models K_{i+1}(\neg D_i)$.

Case A_5^n: ϕ is $K_i(\psi)$ for some $1 \leq i \leq n$ and some tautology ψ. Since ψ is a tautology, $W_i \models \psi$, thus $\mathscr{M} \models K_i(\psi)$.

Case A^n_6: ϕ is $K_i(\psi \to \rho) \to K_i(\psi) \to K_i(\rho)$ for some $1 \leq i \leq n$ and some ψ, ρ. Assume $\mathscr{M} \models K_i(\psi \to \rho)$ and $\mathscr{M} \models K_i(\psi)$. Then $W_i \models \psi \to \rho$ and $W_i \models \psi$, thus $W_i \models \rho$, thus $\mathscr{M} \models K_i(\rho)$.

Case A^n_7: ϕ is $K_i(\psi) \to K_j(\psi)$ for some $1 \leq i < j \leq n$. Assume $\mathscr{M} \models K_i(\psi)$, so $W_i \models \psi$. Since $W_j \models W_i$, we see $W_j \models \psi$. Thus $\mathscr{M} \models K_j(\psi)$.

Case $A^n_T{}'$: ϕ is $K_j(K_i(\psi) \to \psi)$ for some $1 \leq i < j \leq n$ and some ψ. By $A^n_{T,j}$, $(V_n)^j_0 \models K_i(\psi) \to \psi$. Since $(V_n)^j_0 \subseteq W_j$, we see $W_j \models K_i(\psi) \to \psi$, thus $\mathscr{M} \models K_j(K_i(\psi) \to \psi)$.

Case $A^n_\infty{}''$: ϕ is $K_i(\psi)$ for some $1 \leq i \leq n$ and some ψ such that $(V_n)_0 \models \psi$. Since $(V_n)_0 \subseteq (V_n)^i_0 \subseteq W_i$, we see $W_i \models \psi$, thus $\mathscr{M} \models K_i(\psi)$.

Case $A^n_{T,p}$: ϕ is $K_r(\psi) \to \psi$ for some $1 \leq r < p$. Assume $\mathscr{M} \models K_r(\psi)$, so $W_r \models \psi$. By our p-induction hypothesis, $\mathscr{M} \models W_r$. Thus $\mathscr{M} \models \psi$.

Case $A^n_{p,\infty}$: ϕ is $K_j(\psi)$ for some $1 \leq j \leq p$ and some ψ such that $(V_n)^j_0 \models \psi$. Since $(V_n)^j_0 \subseteq W_j$, we see $W_j \models \psi$. Thus $\mathscr{M} \models K_j(\psi)$.

Case $\bigwedge_{1 \leq j < i, j \neq m} \neg D_j$: ϕ is $\bigwedge_{1 \leq j < i, j \neq m} \neg D_j$. Then clearly $\mathscr{M} \models \phi$.

This concludes the proof of Claim 1.

Claim 2: $\mathscr{M} \models K_i(\phi) \to \phi$ for all $1 \leq i \leq n$ and all ϕ. Fix any such i and assume $\mathscr{M} \models K_i(\phi)$, which means $W_i \models \phi$. By Claim 1, $\mathscr{M} \models W_i$. Thus $\mathscr{M} \models \phi$.

Since each W_i includes $(V_n)^i_0$, Claims 1–2 together show $\mathscr{M} \models V_n$. □

It is a straightforward exercise to show that $(V_2)_0 \models \neg D_2$ (by similar reasoning as in the preliminary claim in the proof of Theorem 1), so by Lemma 2, $(V_2)_0 \models K_1(\neg D_2)$. From this it is an easy exercise to show $(V_2)_0 \models K_1(D_1)$. Together this rules out the first two disjuncts of

$$A^2_3{}' \equiv (D_1 \wedge \neg K_1(D_1)) \vee (D_2 \wedge \neg K_2(D_2)) \vee (D_1 \wedge K_1(D_2)),$$

so $(V_2)_0 \models D_1 \wedge K_1(D_2)$. By $A^2_T{}''$, $V_2 \models D_2$. But $V_2 \models \neg D_2$. So V_2 is inconsistent. And V_1 is clearly inconsistent. Thus the requirement $n > 2$ in Theorem 3 is sharp.

6 Whether knowledge-of-factivity should be assumed?

We will argue that knowledge-of-factivity should not be assumed. Due to anticipated controversy, we will address the question in the style of a metaphysical disputation, giving the top priority not to ourselves but to our opponents, as in Thomas Aquinas's *Summa Theologica*. For in philosophy, anyone can argue anything, thus the real test is not how one argues one's own position, as much as how one replies to objections. Thus we present, in order:

- Anticipated objections to our answer.
- Our answer.
- Replies to objections.

Objection 1: Factivity is part of the definition of knowledge, which definition knowers should know. Therefore, knowledge-of-factivity should be assumed.

Objection 2: Knowledge should conform to Kripke's possible-worlds semantics. And factivity itself should certainly be assumed, thus factivity should hold in all possible worlds. But in Kripke's semantics, anything that holds in all possible worlds is known.

Objection 3: The teacher's implicit trustworthiness is essential to the surprise examination paradox. Presumably the students are aware that they trust the teacher. And presumably they would have known their own factivity, had the teacher stayed silent. But an announcement from a trusted source should not cause the students to suddenly doubt their own factivity.

Objection 4: For any statement $S(X)$, for any A and B, if $A = B$, then $S(B)$ implies $S(A)$. Let $S(X)$ be "The students know the factivity of X". Let A be the consequences of the teacher's announcements, and let B be the true consequences of the teacher's announcements. Clearly we should assume $S(B)$. But the teacher is truthful, thus $A = B$, so $S(B)$ implies $S(A)$.

Objection 5: If the paradox is allegedly resolved by weakening knowledge-of-factivity, it re-emerges if the teacher announces "there will be a surprise exam next week, and you are factive". Thus, as far as paradoxes go, we gain little or nothing by refusing knowledge-of-factivity.

We answer that: There are different types of students. On one extreme, there are students who merely take the teacher's sayings as a guide to more quickly discover what they could have discovered on their own. Thus, the boy in Plato's *Meno* (82b-85d) could have discovered geometry on his own, without Socrates. For this type of student, knowledge-of-factivity might be a very reasonable assumption. On the opposite extreme, some students accept whatever the teacher says. For example, a robot might be programmed to accept everything its owner tells it. Students of the former extreme can never be taught contingent facts, just as Russell could not teach Wittgenstein that there is no rhinocerus in the room. Examinations are contingent, thus the surprise examination paradox only makes sense for students who accept what the teacher tells them. So we should here treat students as if they're bound to accept whatever their teacher says. Now, the students can remember things the teacher said in the past, but they cannot predict what the teacher will say in the future. On Monday, they cannot predict that "tomorrow (or even later today) the teacher will not declare anything contradictory," and so they cannot predict that what they can deduce from the teacher will, on Tuesday, be factive. Indeed, the students themselves might not even call their knowledge "knowledge", but rather "belief" or "teacher-provability". But if we outside observers do know that the teacher is truthful, then *we* can call the students' belief "knowledge" even if they themselves don't. So our answer is, we should not assume knowledge-of-factivity, because students cannot predict the teacher won't say something false.

Finally, we can reply to the previous objections.

Reply to Objection 1: When the students reason about what they can or cannot deduce from their teacher, they do not know that they are reasoning about their own knowledge. To them, the K_i operators of Definition 1 are provability operators, some of which might be un-factive if, in future, the teacher says something contradictory. If we outside observers know the teacher will never contradict herself, then we can think of K_i as knowledge, but that doesn't imply the students must. "For it is possible for us to think we do not know what in fact we do know" (Descartes).

Reply to Objection 2: Kripke semantics are only appropriate when the knower knows that the modalities in question conform to Kripke semantics. See also [3].

Reply to Objection 3: Prior to the teacher's saying anything, the students may have known the factivity of their past knowledge. But they could not then have known the factivity of what we call their future knowledge (and what they might call "things we will be able to deduce from the teacher in future"), because they cannot predict the future.

Reply to Objection 4: This fallacy is known as the morning star paradox [12]. For, let $S(X)$ be "Everyone knows the morning star is X", let A be the evening star, and let B be the morning star. Then $S(B)$ seems plausible, and $A = B$ (A and B equal planet Venus), yet $S(A)$ seems implausible. This shows replacement does not work this way in modal logic in general.

Reply to Objection 5: If the teacher announces factivity, then the set of all things the teacher has announced becomes a theory that proves its own consistency. If the resulting inconsistency is a "paradox", then by the same logic, so is Gödel's 2nd incompleteness theorem. But said theorem is not generally considered a paradox.

7 Conclusion

We have argued that factivity, and knowledge-of-factivity, play an implicit role in the surprise examination paradox, even though at first glance the paradox might not seem to assume them at all. If students know the surprise exam will take place on Friday, then it would be surprising to them if the surprise exam takes place on Thursday. That such situations are not included in the definition of surprise in standard formalizations of the paradox is apparently because such situations are impossible: the students *cannot* know the exam will take place on Friday if in fact the exam takes place on Thursday, because knowledge is factive. But for the students themselves to simplify the definition of surprise in this way, they themselves would need to know their own factivity. So, even though at first glance the paradox does not seem to hinge on factivity or the knowledge thereof, it seems that factivity and knowledge-of-factivity play an implicit role in the definition of surprise.

One might hope that the surprise examination paradox would vanish if we merely redefined surprise to include such impossible situations as the students knowing the exam will be Friday even though the exam is in fact Thursday. But we showed in Theorem 2 that the paradox evades this attempted resolution, provided that factivity and knowledge-of-factivity are assumed.

However, in Theorem 3, we showed that if surprise is thus redefined, and if the assumption of knowledge-of-factivity is dropped (even while still assuming factivity itself), then the paradox vanishes. In fact, we showed even more. We showed the paradox still vanishes even if knowledge-of-factivity is not totally dropped: if knowledge-of-factivity is weakened to the statement that, on each day, the students know the factivity of their own knowledge from prior days, then that weakening is already sufficient to remove the paradox.

Thus, in our opinion, the surprise examination paradox results from two flaws: the assumption of knowledge-of-factivity (that, in addition to the students' knowledge being factive, that the students themselves *know* as much), and the over-simplification of the definition of surprise accordingly. This is an important step forward with respect to conjectures proposed in the literature (as those, e.g., by McLelland and Chihara [17]) and focusing on the role of different forms of positive introspection (based on the KK rule) as the causal factors triggering the paradox. Along this line of reasoning, in [18] the authors show that the KK principle is unrelated to the causes of the paradox, and in fact even in our formalization this rule is not assumed to derive a contradiction[2]. The same authors of [18] focus instead on the retention principle, by showing that invalidating it suffices to restore consistency. In other cases, the paradox is resolved by assuming that the teacher's announcement is actually never known, or that the students do not trust it for the whole week, which is similar to stating that the students do not retain knowledge in general. With respect to these proposals, our work contributes with a new perspective allowing us to resolve the paradox by giving up a very specific form of knowledge.

We also point out that our solution relies purely on arguments about knowledge, thus differing from other approaches that deal with the paradox by replacing the notion of knowledge by the notion of provability [6], as done, e.g., in [14], where Gödel's second Incompleteness Theorem is used to offer a way out of the antinomy.

In future work, we would like to investigate the possibility of simultaneously resolving multiple epistemic paradoxes at once by the construction of a single model. For example, imagine that to the formulas of $\mathscr{L}$ (Definition 1) we added an additional clause saying that for every $i = 1, 2, \ldots$, there is a non-atomic formula L_i; and imagine that for semantics, we declare that for every model $\mathscr{M}$, $\mathscr{M} \models L_i$ iff $\mathscr{M} \models K_i(\neg L_i)$. Thus L_i is a variation of the liar sentence for the students' knowledge on midnight just before day i: intuitively, L_i could be thought of as the sentence: "On midnight just before day i, we know this sentence is false". In this expanded logic, it can be shown that, e.g., S4 (or even weaker systems including knowledge-of-factivity) are inconsistent—this is a temporal variation of the Paradox of the Knower. We conjecture that by modifying the construction in Section 5, it would be possible to construct a single model which simultaneously resolves the surprise examination paradox and this temporal version of the Paradox of the Knower. We conjecture it would even be possible to construct the model in such a way as to satisfy $K_j(\neg L_i)$ whenever $i < j$, i.e., so that the students know (on any day) the falsehood of earlier days' liar sentences. Resolving multiple paradoxes at once, with the same model, and by weakening the same assumption, would be, in our opinion, strong evidence in

[2] It can be shown that a temporal version of the KK axiom can be added to V_n, without disrupting its consistency, thus emphasizing once more that KK is *not* the cause of the paradox.

favor of the correctness of said resolution. Moreover, it would confirm the central role of own factivity and its knowledge/ignorance, as already emphasized, e.g., in computational contexts, see, e.g., [2, 1].

References

1. Alessandro Aldini, Vincenzo Fano, and Pierluigi Graziani. Theory of knowing machines: Revisiting Gödel and the mechanistic thesis. In F. Gadducci and M. Tavosanis, editors, *History and Philosophy of Computing*, pages 57–70. Springer, 2016.
2. Samuel A. Alexander. A machine that knows its own code. *Studia Logica*, 102:567–576, 2014.
3. Sergei Artemov. Knowing the model, 2016. doi: 10.48550/arXiv.1610.04955.
4. Ido Ben-Zvi and Yoram Moses. Agent-time epistemics and coordination. In Kamal Lodaya, editor, *Logic and Its Applications*, pages 97–108. Springer, 2013.
5. Robert Binkley. The surprise examination in modal logic. *Journal of Philosophy*, 65(5):127–136, 1968.
6. Timothy Y. Chow. The surprise examination or unexpected hanging paradox. *The American Mathematical Monthly*, 105(1):41–51, 1998.
7. John M. Collins. Epistemic closure principles. In *Internet Encyclopedia of Philosophy*. 2006.
8. John Earman. A user's guide to the surprise exam paradoxes. *PhilSci Archive*, July 2021.
9. Ronald Fagin, Joseph Y. Halpern, Yoram Moses, and Moshe Y. Vardi. *Reasoning About Knowledge*. MIT Press, 2003.
10. Martin Gardner. The paradox of the unexpected hanging. In *The unexpected hanging and other mathematical diversions*, pages 11–23. The University of Chicago Press, 1991.
11. Jelle Gerbrandy. The surprise examination in dynamic epistemic logic. *Synthese*, 155(1):21–33, 2007.
12. Stig Kanger. The morning star paradox. *Theoria*, 23(1):1–11, 1957.
13. David Kaplan and Richard Montague. A paradox regained. *Notre Dame J. Formal Logic*, 1(3):79–90, 1960.
14. Shira Kritchman and Ran Raz. The surprise examination paradox and the second incompleteness theorem. *Notices of the AMS*, 57(11):1454–1458, December 2010.
15. Steven Luper. Epistemic closure. *The Stanford Encyclopedia of Philosophy*, 2020.
16. Avishai Margalit and Maya Bar-Hillel. Expecting the unexpected. *Philosophia*, 13:263–288, 1984.
17. James McLelland and Charles Chihara. The surprise examination paradox. *Journal of Philosophical Logic*, 4(1):71–89, 1975.
18. Julien Murzi, Leonie Eichhorn, and Philipp Mayr. Surprise, surprise: KK is innocent. *Thought: A Journal of Philosophy*, 10(1):4–18, 2021.
19. Michael Scriven. Paradoxical announcements. *Mind*, 60:403–407, 1951.
20. Roy Sorensen. Epistemic paradoxes. *The Stanford Encyclopedia of Philosophy*, 2022.
21. Fredrik Stjernberg. Restricting factiveness. *Philosophical Studies*, 146(1):29–48, 2009.
22. Timothy Williamson. *Knowledge and its Limits*. Oxford University Press, 2000.