# Simplifying the Surprise Exam*

Wesley H. Holliday
University of California, Berkeley

June 30, 2015

**Abstract**

In this paper, I argue for a solution to the surprise exam paradox, designated student paradox, and variations theoreof, based on an analysis of the paradoxes using modal logic. The solution to the paradoxes involves distinguishing between two setups, the Inevitable Event and the Promised Event, and between the two-day and $n$-day cases of the paradoxes. For the Inevitable Event, the problem in the two-day case is the assumption that the student knows the teacher's announcement; for more days, the student can know the announcement, and the base case of the student's backward induction is correct, but there is a mistake in the induction step. For the Promised Event, even the base case is questionable. After defending this analysis, I argue that it also leads to a solution to a modified version of the surprise exam paradox, due to Ayer and Williamson, based on the idea of a conditionally expected exam.

**Keywords:** surprise exam paradox, designated student paradox, conditionally expected exam, modal logic, epistemic logic

## 1 Introduction

According to a standard presentation of the *surprise exam paradox*, a teacher announces to her student that she will give him a surprise exam during a term of $n \geq 2$ days. An exam on day $i$ is a surprise iff the student does not know on the morning of $i$ that an exam will be given later on $i$. The student, a perfect logician, reasons as follows: "Since the exam will be a *surprise*, the teacher cannot wait until day $n$ to give the exam; because if she does, then on the morning of day $n$, my future self, remembering that an exam has not yet occurred, will know that the exam has to be later on day $n$—so it will not be a surprise. Moreover, the teacher cannot wait until day $n-1$ to give the exam; because if she does, then on the morning of day $n-1$, my future self, still knowing that the exam will be a *surprise*, will also know that the teacher cannot wait until day $n$ (on the basis of the reasoning I just used to eliminate $n$), and thus, remembering that an exam has not yet occurred, will know that the exam has to be later on day $n-1$—so it will not be a surprise." Repeating this backward elimination argument, the student concludes that the teacher cannot give a surprise exam. Having done so, he may be especially surprised when the exam comes on, say, day $n-1$.

*Note added May 25, 2016: For helpful feedback, I wish to thank Harvey Lederman, Jack Spencer, and the students in my Fall 2012 graduate seminar at Berkeley on Epistemic Logic and Epistemology, during which most of this paper was written. The paper was the basis for the discussion of the surprise exam paradox in the course on *Ten Puzzles and Paradoxes of Knowledge and Belief* that I co-taught with Eric Pacuit at ESSLLI 2013 in Dusseldorf, and since then the manuscript has circulated by email to students and colleagues at other universities (see, e.g., Rustenburg 2014). The paper has been under review at a journal since June 30, 2015, but as I continue to receive requests for the manuscript, I am now making it publicly available as a UC Berkeley Working Paper in Philosophy. The bibliography is now out of date, as the paper Holliday 2014b has since been published in *Thought*, and the paper Holliday 2014a is forthcoming in a volume in honor of Jaakko Hintikka.

Another version of the paradox, Sorensen's [1984] *designated student paradox*, removes some of the temporal aspects of the original version and thereby, according to Sorensen, undermines several analyses of the surprise exam paradox [Harrison, 1969, McLelland and Chihara, 1975, Wright and Sudbury, 1977]. A teacher displays to her class of $n \geq 2$ perfect 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.[1] 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. That is, he will not know before pulling his star off his back that it is the gold star. Student 1, at the front of the line, reasons as follows: "Since the gold star is a *surprise*, the teacher did not give the gold star to student $n$; because if she did, then student $n$, seeing all silver stars in front of him, would know he has the gold star—so it would not be a surprise. Moreover, the teacher did not give the gold star to student $n-1$; because if she did, then student $n-1$, knowing the gold star is a *surprise*, would also know that the teacher did not give the gold star to student $n$ (on the basis of the reasoning I just used to eliminate $n$), and thus, seeing all silver stars in front of him, would know he has the gold star—so it would not be a surprise." Repeating this backward elimination argument, student 1 concludes that the teacher's announcement is false. But then when the students pull the stars off their backs, it is, say, student $n-1$ who has the gold star, and he did not know beforehand that he had the gold star, so the teacher's announcement was true after all.

In this paper, I argue that the solution to these paradoxes, and variations thereof, involves distinguishing between two setups, the Inevitable Event and the Promised Event, and between the $n = 2$ and $n > 2$ cases. For the Inevitable Event setup, in the $n = 2$ case the student cannot know the teacher's announcement[2] (at least when he is clever enough to do the backward elimination reasoning), since such knowledge would lead to impossible Moorean knowledge of the form *I have the gold star but I don't know I have it*; for $n > 2$, the student can know the teacher's announcement, and the base case of the student's backward induction (eliminating $n$) is correct, but there is a mistake in the induction step. In particular, in the designated student paradox, when trying to eliminate student $n-1$, student 1 assumes that student $n-1$ knows the teacher's announcement;[3] but if student $n-1$ sees all silver stars in front of him, a possibility that student 1 has not initially eliminated (before he starts reasoning about $n-1$), then student $n-1$ is in essentially the same epistemic position as student 1 in the $n = 2$ case, in which case student $n-1$ cannot know the teacher's announcement (for the same reason about Moorean knowledge); it is because student 1 has not initially eliminated this possibility (before he starts reasoning about $n-1$) that he does not know that student $n-1$ knows the teacher's announcement, so his mistake is to assume it. The same analysis, for the Inevitable Event setup, applies to the surprise exam reasoning: the base case is correct, but the induction step is mistaken.[4] For the Promised Event setup, even the base case of the students' reasoning is questionable.

To make the case for this analysis precisely, I follow the tradition of others who have formalized the

[1] Unlike Sorensen's [1984] original presentation of the designated student paradox, in Sorensen's [1988, 317] later presentation of the paradox, the teacher only announces that she *will* put the stars on the students' backs, without actually doing so before the clever student starts reasoning. See the discussion of the Inevitable Event vs. the Promised Event in §4.

[2] Whenever I say that a student can(not) "know the teacher's announcement," I mean that he can(not) know that the teacher's announcement *is true*—or better, that he can(not) know the proposition expressed by the teacher's announcement. Of course, the student can know that the teacher *made* the announcement.

[3] Recall what the clever students said: "the teacher cannot wait until day $n-1$ to give the exam; because if she does, then on the morning of day $n-1$, my future self, still knowing the exam will be a *surprise*…" and "the teacher cannot give the gold star to student $n-1$; because if she does, then student $n-1$, knowing the gold star is a *surprise*…."

[4] This is not to say that the student's only mistake for $n > 2$ is to assume that student $n-1$ knows the teacher's announcement. In the standard presentation of the surprise exam paradox, a later mistake—or at least an oversight—is that the student fails to realize that when he (mistakenly) reaches the conclusion that the teacher cannot give a surprise exam, he is then susceptible to being surprised by an exam the next day, undermining his conclusion that there cannot be a surprise exam. See §8.

paradox in modal logic (e.g., Binkley 1968, Harrison 1969, McLelland and Chihara 1975, Sorensen 1988). However, I formalize the paradox using a weaker proof system and set of premises (and a simpler proof) than in many of the previous modal formalizations, thus yielding a stronger inconsistency result.

After arguing for the above analysis in more detail, comparing it to other analyses, and considering some of its broader lessons, I discuss a modified version of the surprise exam paradox, due to Ayer [1973] and Williamson [1992, 2000], based on the idea of a *conditionally expected exam*. Williamson suggests that this modification makes for a stronger paradox. I argue that the solution to the standard version of the paradox also leads to a solution to the modified version.

## 2 Modal Logic

To formalize the paradoxes, I will use a propositional modal language with a sentential operator $\Box_i$ and an atomic sentence $p_i$ for each $i \in \mathbb{N}$. For the surprise exam paradox, we can adopt the following readings:

$\Box_i\varphi$ "the student knows on the *morning* of day $i$ that $\varphi$";

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

For the designated student paradox, we can adopt the following readings, where $t$ is a time after the teacher has distributed the stars on the students' backs and made her announcement, but before the students remove the stars:

$\Box_i\varphi$ "the $i$-th student in line knows at $t$ that $\varphi$";

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

We can now express that there is a surprise exam on day $i$ (or a surprise gold star on student $i$): $p_i \wedge \neg\Box_i p_i$.

Although I will mostly write about $\Box_i$ in terms of knowledge, much of what follows can also be applied to rational belief. In particular, many of the results will not depend on the T axiom, $\Box_i\varphi \rightarrow \varphi$.

For a proof system, I use the polymodal version of the minimal normal modal logic **K**, the smallest system extending propositional logic with the following rule for each $i \in \mathbb{N}$ (Chellas 1980, §4.1):

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

which states that if the premise is a theorem, so is the conclusion. In the $m = 0$ case, $\mathrm{RK}_i$ is the standard rule of Necessitation ($\mathrm{Nec}_i$). Intuitively, $\mathrm{RK}_i$ says that the student on day $i$ (or the $i$-th student) knows all the logical consequences of what he knows, reflecting the assumption that he is a perfect logician.[5] As Sorensen [1988] notes about the surprise exam paradox, "members of the audience are standardly assumed to be perfect logicians who never forget and who do not overlook important facts" (254).[6]

There are a number of possible philosophical objections to $\mathrm{RK}_i$, even for ideally rational agents. From $\mathrm{RK}_i$ one can derive the principle $(\Box_i\alpha \wedge \Box_i\beta) \rightarrow \Box_i(\alpha \wedge \beta)$, which has been questioned as a principle of belief for rational agents [Kyburg, 1961, Makinson, 1965, Skyrms, 1967, 388], and $(\Box_i\alpha \wedge \Box_i(\alpha \rightarrow \beta)) \rightarrow \Box_i\beta$, which has been questioned as a principle of knowledge for rational agents, even if these agents competently deduce all the consequences of what they know [Dretske, 1970, Nozick, 1981, Dretske, 2005].[7] However, the relevant

[5] Moreover, $\mathrm{RK}_i$ reflects the assumption that for every $i, j \in \mathbb{N}$, the student on day $i$ (or the $i$-th student) knows that his self on day $j$ (or the $j$-th student) is also a perfect logician. For the relevance of this to the paradoxes, see §4.

[6] This is not to deny that interesting aspects of the paradox (or related paradoxes) can be brought out by considering boundedly rational agents or agents whose reasoning takes time. But for reasons of space, I will not consider such agents here.

[7] Note that if our agent knows $\alpha \rightarrow (\beta \rightarrow (\alpha \wedge \beta))$, then the first principle is derivable from the second.

question in this paper will be whether the use I make of $\mathrm{RK}_i$ (e.g., in the proof of Proposition 1) involves the kind of moves that worry epistemologists about those principles. More generally, the question about the logic $\mathbf{K}$ is not whether it involves false idealizations, which it surely does for real agents; the question is whether its false idealizations interfere with the analysis of this paper. If they do, then one should be able to pinpoint where they do. The situation is analogous to modeling in economics or physics; false idealizations abound, but what matters is whether they distort the analysis of the target phenomenon. In this paper I claim that formalizing the surprise exam paradox in a normal modal logic illuminates more than it distorts.

Another reason that objections to the use of $\mathbf{K}$ can be avoided here is that we can adopt a different reading of the $\Box_i$ operator, so that $\Box_i\varphi$ means "$\varphi$ is entailed by what the student knows on the morning of day $i$" (or "$\varphi$ is entailed by what the $i$-th student in line knows at $t$").[8] Then $\mathrm{RK}_i$ is completely uncontroversial. Moreover, we can take a "surprise exam" on day $i$ to be one such that it is *not entailed by* what the student knows on the morning of day $i$ that there will be an exam on the afternoon of day $i$ (cf. Sorensen 1988, 258), which is still expressed by the same formal sentence: $p_i \wedge \neg\Box_i p_i$. Then the clever student can carry out the backward elimination argument using the modified definition of surprise, which I leave to the reader.

Later we will consider extensions of $\mathbf{K}$ with axiom schemas such as those in Figure 1 below. Given schemas $\Sigma_1, \ldots, \Sigma_n$, $\mathbf{K\Sigma_1 \ldots \Sigma_n}$ is the smallest extension of $\mathbf{K}$ that includes all instances of $\Sigma_1, \ldots, \Sigma_n$. A sentence $\beta$ is *provable* in $\mathbf{K\Sigma_1 \ldots \Sigma_n}$ from a set of sentences (premises) $\Gamma$, written '$\Gamma \vdash_{\mathbf{K\Sigma_1 \ldots \Sigma_n}} \beta$', iff there is a sequence $\langle \chi_1, \ldots, \chi_l \rangle$ of sentences with $\beta = \chi_l$ such that for all $1 \leq k \leq l$, one of the following holds:

(i) $\chi_k$ is an instance of a propositional tautology;

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

(iii) $\chi_k$ is one of the sentences in $\Gamma$;

(iv) (RK) $\chi_k$ is of the form $(\Box_i\varphi_1 \wedge \cdots \wedge \Box_i\varphi_m) \rightarrow \Box_i\psi$ for some $i \in \mathbb{N}$, and for some $j < k$, $\chi_j$ 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 $\chi_i$ is $\chi_j \rightarrow \chi_k$.

If there is no such proof, I write '$\Gamma \nvdash_{\mathbf{K\Sigma_1 \ldots \Sigma_n}} \beta$'. As usual, $\beta$ is a *theorem* of $\mathbf{K\Sigma_1 \ldots \Sigma_n}$ iff $\beta$ is provable from no premises, i.e., $\vdash_{\mathbf{K\Sigma_1 \ldots \Sigma_n}} \beta$. It is important to observe the requirement in (iv) that the sentence $\chi_j$ to which the $\mathrm{RK}_i$ rule is applied must be a theorem of the logic. On this point, a word of warning: in order to shorten proofs, I will often say that a sentence $\beta$ is provable from a sentence $\alpha$ "using $\mathrm{RK}_i$," even though $\nvdash_{\mathbf{K}\Sigma_1 \ldots \Sigma_n} \alpha$, so $\mathrm{RK}_i$ cannot be applied to $\alpha$. What I mean in such cases is that there is some obvious theorem (usually a propositional tautology) $\varphi$ to which $\mathrm{RK}_i$ can be applied to obtain another theorem $\varphi'$ such that $\beta$ is obviously provable (as a matter of propositional logic) from $\alpha$ and $\varphi'$. By filling in the omitted steps, the reader can check that none of the proofs involve mistaken applications of the $\mathrm{RK}_i$ rule to non-theorems. Finally, when I apply the $m = 0$ case of the $\mathrm{RK}_i$ rule in a proof, I write '$\mathrm{Nec}_i$' instead of '$\mathrm{RK}_i$'.

| | | | |
|---|---|---|---|
| $\mathrm{T}_i$ | $\Box_i\varphi \rightarrow \varphi$ | $4_i$ | $\Box_i\varphi \rightarrow \Box_i\Box_i\varphi$ |
| $\mathrm{D}_i$ | $\Box_i\varphi \rightarrow \neg\Box_i\neg\varphi$ | $4_i^<$ | $\Box_i\varphi \rightarrow \Box_i\Box_j\varphi \ \ (i < j)$ |
| $\mathrm{R}_i$ | $\Box_i\varphi \rightarrow \Box_j\varphi \ \ (i < j)$ | $4_i^\leq$ | $\Box_i\varphi \rightarrow \Box_i\Box_j\varphi \ \ (i \leq j)$ |

Figure 1: Axiom Schemas

[8] See Cross 2001, §3 for a similar move in a discussion of the Knower Paradox. Also see Williamson's [1992, 220] "know$^{\#}$".

Returning to the axiom schemas in Figure 1, a schema like $\mathrm{R}_i$ includes all instances of $\Box_i\varphi \rightarrow \Box_j\varphi$ for the fixed $i$ and for *any* $j > i$. We could let $\mathrm{R}_{i,j}$ be the schema that includes all and only instances of $\Box_i\varphi \rightarrow \Box_j\varphi$ for the fixed $i$ and fixed $j$, but we will not have a need to do so here. The same points apply to the other schemas. Finally, while one might wish to adopt an axiom schema for some times/agents $i$ but not others, one might also wish to adopt it uniformly. To indicate the latter, let T be the schemata that includes $\mathrm{T}_i$ for all $i \in \mathbb{N}$; let D the schemata that includes $\mathrm{D}_i$ for all $i \in \mathbb{N}$; and so on for the other axioms.

As a final preliminary, recall that the *modal depth* of a modal formula is defined recursively as follows: $d(p_i) = 0$ for any atomic $p_i$; $d(\neg\varphi) = d(\varphi)$; $d(\varphi\#\psi) = max(d(\varphi), d(\psi))$ for any binary boolean connective $\#$; and $d(\Box_i\varphi) = d(\varphi) + 1$. It is a significant fact that all of the analysis in this paper will be done using the fragment of the modal language consisting of formulas of modal depth 3 or less.

## 3 The $n = 2$ Case

To formalize the $n = 2$ case of the paradox, I use the following premises, which are as weak as or weaker than any in the literature for the standard surprise exam or designated student paradox (see Appendix B):

$(A)$ $\Box_1((p_1 \wedge \neg\Box_1 p_1) \vee (p_2 \wedge \neg\Box_2 p_2))$;

$(B)$ $\Box_1(p_2 \rightarrow \Box_2\neg p_1)$;

$(C)$ $\Box_1\Box_2(p_1 \vee p_2)$.

For the surprise exam paradox, $(A)$ states that the student knows on the morning of day 1 the teacher's announcement that there will be a *surprise* exam. $(B)$ 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*.

For the designated student paradox, $(A)$ states that the first student in line, student 1, knows that the gold star is a *surprise*. $(B)$ 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). Finally, $(C)$ states that student 1 knows that student 2 knows that one of them has the gold star.

Previous formalizations (see Appendix B) of the paradox use extensions of **K** with axiom schemas such as those in Figure 1. However, it is easy to show that $(A)$, $(B)$, and $(C)$ generate a problem with just **K**:

**Proposition 1.** $\{(A), (B), (C)\} \vdash_{\mathbf{K}} \Box_1(p_1 \wedge \neg\Box_1 p_1)$.

*Proof.* Here is a simple proof, skipping purely propositional steps and some applications of RK to theorems:

$(A)$ $\Box_1((p_1 \wedge \neg\Box_1 p_1) \vee (p_2 \wedge \neg\Box_2 p_2))$ premise

$(B)$ $\Box_1(p_2 \rightarrow \Box_2\neg p_1)$ premise

$(C)$ $\Box_1\Box_2(p_1 \vee p_2)$ premise

(1) $\big(\Box_2(p_1 \vee p_2) \wedge \Box_2\neg p_1\big) \rightarrow \Box_2 p_2$ by propositional logic and $\mathrm{RK}_2$

(2) $\Box_1\big(\big(\Box_2(p_1 \vee p_2) \wedge \Box_2\neg p_1\big) \rightarrow \Box_2 p_2\big)$ from (1) by $\mathrm{Nec}_1$

(3) $\Box_1(\Box_2\neg p_1 \rightarrow \Box_2 p_2)$ from $(C)$ and (2) using propositional logic and $\mathrm{RK}_1$

(4) $\Box_1\neg(p_2 \wedge \neg\Box_2 p_2)$ from $(B)$ and (3) using propositional logic and $\mathrm{RK}_1$

(5) $\Box_1(p_1 \wedge \neg\Box_1 p_1)$ from $(A)$ and (4) using propositional logic and $\mathrm{RK}_1$. □

From (5) we can derive a contradiction by adding to **K** the "weak factivity" (cf. Fara 2010, 58) axiom schema

$\mathrm{J}_i$ $\Box_i\neg\Box_i\varphi \rightarrow \neg\Box_i\varphi$,

so we have the following inconsistency result, stronger than many previous results (see Appendix B):

**Corollary 1.** $\{(A),(B),(C)\} \vdash_{\mathbf{KJ_1}} \bot$.

But (5) is already paradoxical.[9] If we want to add another axiom to derive a contradiction from (5), this reflects the sense that (5) is paradoxical by itself, for both knowledge and rational belief. Of course, the J axiom is a special case of the T axiom, $\Box_i\varphi \rightarrow \varphi$, and Corollary 1 also holds for the minimal normal epistemic logic, **KT**; but it is noteworthy that it holds even for a weaker doxastic logic such as **KJ**.

Two ideas in standard presentations of the surprise exam paradox are that (i) the student concludes there can be no exam and (ii) when the exam actually occurs, the student is surprised. However, (i) and (ii) are unnecessary, since $(A)$, $(B)$, and $(C)$ generate the paradoxical (5) in **K** without implying (i) or (ii):

**Proposition 2.** $\{(A),(B),(C)\} \nvdash_{\mathbf{KD4}^<\mathbf{R}} \Box_1\neg(p_1 \vee p_2) \vee (p_1 \wedge \neg\Box_1 p_1) \vee (p_2 \wedge \neg\Box_2 p_2)$.[10]

*Proof.* See Appendix A. □

Thus, analyses of the paradox for $n = 2$ that depend on (i) and (ii) do not get to the heart of the matter.

The proof above does not even involve the student ruling out an *exam* on the last day, but rather a *surprise* exam on the last day. However, if we add the assumption, as many authors do (see Appendix B), that the student knows/believes on day 1 that there will be an exam on a given day only if it is a surprise, i.e., $\Box_1((p_1 \rightarrow \neg\Box_1 p_1) \wedge (p_2 \rightarrow \neg\Box_2 p_2))$, then the student can rule out an exam on day 2, and we can derive $\Box_1\neg p_2$ from (4) using the assumption. In **KT**, the assumption already follows from the conjunction of $(A)$, $(B)$, and $(C)$, for the trivial reason that the conjunction is inconsistent in **KT**.

It is also noteworthy that Proposition 1 uses $(C)$ instead of

[9] It is paradoxical given the meaning of $p_1$ stated in §2. Note that it is not relevant here to claim that there are other examples in which $\Box_i(\varphi \wedge \neg\Box_i\varphi)$ makes sense (for belief) and therefore that the J axiom is not valid in general (for belief). Perhaps it could be truly said of a non-ideally rational agent that he believes a particular stereotype (as evidenced by his behavior) and that he also believes that he does not believe the same stereotype (as evidenced by what he says). It is less clear whether $\Box_i(\varphi \wedge \neg\Box_i\varphi)$ can hold for the beliefs of an ideally rational agent. Adapting a famous example of Quine's [1956], one might claim that an ideally rational Ralph could believe that Ortcutt (*the man he is looking at*) is a spy while also believing that he (Ralph) does not believe that Ortcutt (*his nextdoor neighbor*) is a spy. Or adapting a famous example of Perry's [1977], one might claim that Heimson, who believes he is Hume (an error in matters of fact), might also believe that *it's raining but Heimson doesn't believe it is*. I do not think that such examples should be formalized as $\Box_i(\varphi \wedge \neg\Box_i\varphi)$ (see Holliday and Perry 2014), but in any case, our (5) obviously does not involve such "double vision" or confusions of identity. (Thanks to Jack Spencer and Harvey Lederman for discussion here.) Another kind of case, suggested to me by Harvey Lederman, involves a sort of "propositional double vision." If a trustworthy oracle were to testify to the truth of a sentence in a language that Ralph doesn't understand, or a sentence too complex for Ralph to understand, which in fact expresses a proposition that is Moorean for Ralph, then it seems Ralph could at least believe that the sentence is true. Whether Ralph could also believe the Moorean proposition it expresses is less clear. In any case, as Lederman points out, this example seems more related to the surprise exam paradox, and it may help explain Sorensen's [1988, 312] reaction that it must be rational to believe the teacher's announcement: the trustworthy teacher announces something that does not look like a Moorean sentence, so it seems rational to accept it—but then a logical deduction shows that together with other things the student believes, the announcement entails an explicitly Moorean conclusion. Since I am assuming the student is an ideal logician, he would make the relevant deduction of an explicitly Moorean conclusion. But then it does not seem rational for an agent to believe something that he recognizes as Moorean.

Whatever one makes of the J axiom for rational belief, remember that $\Box_i(\varphi \wedge \neg\Box_i\varphi)$ does not make sense for knowledge. It also does not make sense for the alternative reading of $\Box_i\varphi$ mentioned in §2: "$\varphi$ is entailed by what the student knows on the morning of day $i$." For this reading, the logic should be at least as strong as **KT**, in which $\Box_i(\varphi \wedge \neg\Box_i\varphi)$ is inconsistent.

[10] Of course, **KT** derives the formula from $\{(A),(B),(C)\}$, because it derives $\bot$ and everything else from $\{(A),(B),(C)\}$.

$(E)$ $\Box_1\Box_2((p_1 \wedge \neg\Box_1 p_1) \vee (p_2 \wedge \neg\Box_2 p_2))$.

Thus, analyses for $n = 2$ that blame $(E)$ do not get to the heart of the matter.[11] According to Williamson [2000, 138], "In ruling out a last-day examination, the pupils assume that they will still know on the last morning that there will be a surprise examination, defined as an examination on a day when the pupils do not know in the morning that there will be an examination that day." However, as the previous observations show, this need not be. It suffices for them to assume they will still know on the last morning that $p_1 \vee p_2$.

In response to the derivation of the paradoxical (5) for Proposition 1, we must reject either one of the premises $(A)$, $(B)$, $(C)$, or the rule $\mathrm{RK}_i$.[12] But which? I suspect that for each of $(A)$, $(B)$, and $(C)$, there is a way of filling in the surprise exam and designated student stories so that that premise should be rejected. The surprise exam and designated student stories are usually underdescribed, preventing a unique solution. Nonetheless, some ways of filling in the details are more natural than others. For example, it seems natural to assume that the student in the surprise exam story knows that he has good memory, and that student 1 in the designated student story knows that the students behind him in line have good eyesight. With these assumptions, $(B)$ is unproblematic. The teacher's announcement does not destroy the students' knowledge of good memory or good eyesight. So we should blame one of the other principles, but which?

# 4 The Inevitable Event vs. The Promised Event

The key to deciding which principle to blame is to distinguish two setups of the paradox: the Inevitable Event vs. the Promised Event. This distinction applies not only to the surprise exam scenario, but to other variations of the scenario. Perhaps the clearest examples of an Inevitable Event setup involve cards:

> A group of card players (the subjects) verify that the dealer has a standard deck of ordinary playing cards. The dealer shuffles the cards in view of the subjects. He then announces that he will turn up, one at a time, each card in the deck. This gives us a sequence $e_1, e_2, \ldots, e_{52}$ of situation events. The test event is the turning up of the jack of spades. The dealer announces that the turning up of this card will 'surprise' the players. [McLelland and Chihara, 1975, 72]

> [A] set of seven cards, known to include the Ace of Spades, is put in the man's cell, in a place where it cannot be tampered with, and every morning the prison chaplain comes in and draws a card. On the day when the chaplain draws the Ace of Spades, the man is to die, provided that he does not know that the Ace of Spades will be drawn on that day. [Ayer, 1973, 125]

When I presented the designated student paradox in §1, I essentially gave an Inevitable Event setup. Before the students line up, the teacher clearly displays the gold star and silver stars to them; each student sees that the others see the stars; they can communicate this to each other, etc. Then the teacher lines them up, walks behind them, and sticks the stars to their backs. We can even suppose that the students feel the stars on their backs by reaching around with their hands; they communicate this to each other, etc.

Here is an Inevitable Event setup of the surprise exam scenario: there is a commonly known and sacrosanct school policy stating that whatever happens, an exam will be administered every week.

In each of these cases, it would be radical skepticism about knowledge, not a solution to the paradox, to deny $(C)$. In the card scenario, for example, it is a skeptical stretch to claim that although the subjects

[11] Though see the discussion of the $4^<$ axiom for $n > 2$ in §7.
[12] I assume we will not need to reject propositional logic.

"*verify that the dealer has a standard deck of ordinary playing cards*," they cannot know (or rationally believe) that later they will know (or rationally believe) that it is a standard deck of ordinary playing cards, which contains a jack of spades.[13] Some philosophers might deny the possibility of any knowledge of the future, let alone knowledge of future knowledge. But the Inevitable Event setup of the designated student paradox does not involve knowledge of the future, but rather knowledge of what another person knows. It would be radical skepticism about social knowledge, not a solution to the paradox, to claim that given the preparation described above, student 1 does not know that student 2 knows that one of them has the gold star.

I conclude that in the Inevitable Event setup, $(C)$ is not to blame. Thus, having acquitted $(B)$ and $(C)$, we have narrowed down the suspects in the Inevitable Event setup to $(A)$ and $\mathrm{RK}_i$.

Observe that $\mathrm{RK}_i$ is used in each of steps (1)-(5). If $\mathrm{RK}_i$ is the problem, then in what step does it first reveal itself as problematic? Consider the designated student paradox, continuing with the Inevitable Event setup. In this case, step (1) reflects the idea that student 2 is a perfect logician who would not at the same time believe $p_1 \vee p_2$, believe $\neg p_1$, and yet fail to believe $p_2$. So far, so good it seems. Step (2) reflects the idea that student 1 knows that student 2 is such a perfect logician, which also seems okay.

With step (3), one might raise the following objection in terms of "justification": suppose that student 1's justification for believing $\Box_2(p_1 \vee p_2)$ rests on some justification he has for believing $\Box_2 p_1$. Then it looks suspect, the objection goes, for student 1 to engage in the following conditional proof: suppose $\Box_2 \neg p_1$; then given the fact that $\Box_2(p_1 \vee p_2)$ (justified by $\Box_2 p_1$, which is incompatible with the supposition) and the fact that 2 is a perfect logician, we conclude $\Box_2 p_2$; hence $\Box_2 \neg p_1 \rightarrow \Box_2 p_2$.[14] The first reply to this objection is that in the Inevitable Event setup we are considering, student 1's justification for believing $\Box_2(p_1 \vee p_2)$ does *not* rest on some justification for believing $\Box_2 p_1$. Instead, student 1's justification for believing that student 2 knows that one of them has the gold star rests on his justification for believing that student 2 *saw* both the gold star and the silver star, *saw* and *heard* the teacher walk behind each of them and stick the stars on their backs, etc., none of which is incompatible with the supposition for conditional proof that $\Box_2 \neg p_1$. Second, if student 1 believes $\Box_2 p_1$ and hence (I assume) $\neg \Box_2 \neg p_1$, then what is wrong with his concluding $\neg \Box_2 \neg p_1 \vee \Box_2 p_2$, i.e., $\Box_2 \neg p_1 \rightarrow \Box_2 p_2$? Nothing, I would say. Third, if we actually had $\Box_1 \Box_2 p_1$, then with just the addition of $(A)$ and $\Box_1 \neg (p_1 \wedge p_2)$, we would have an even shorter proof of a contradiction in **KT**. What these three points show is that trying to block $\mathrm{RK}_i$ at step (3) is not the solution to the paradox.

Neither, it seems to me, is trying to block $\mathrm{RK}_i$ at steps (4) or (5). Step (4) reflects the idea that student 1 uses his knowledge of $\Box_2 \neg p_1 \rightarrow \Box_2 p_2$ and his knowledge of $p_2 \rightarrow \Box_2 \neg p_1$ to deduce and come to know $p_2 \rightarrow \Box_2 p_2$, i.e., $\neg (p_2 \wedge \neg \Box_2 p_2)$, by the transitivity of implication, which seems okay. In the final step, student 1 uses his (supposed) knowledge of the disjunction in $(A)$ and his knowledge from (4) of the negation of the right disjunct to deduce and come to know the left disjunct by disjunctive syllogism, landing us in the paradoxical (5). Might one claim here that the rational agent should instead retain his belief in the disjunction in $(A)$ and retain his belief in the negation of the right disjunct, but still refrain from believing the left disjunct, in violation of $\mathrm{RK}_i$? That does not seem rational to me; instead, the rational response would be to give up one of the beliefs (like his belief in the disjunction in $(A)$) that got him into this mess. No doubt there is more to say, but I conclude that denying $\mathrm{RK}_i$ is not the solution to the paradox.

[13] But what if the dealer is an expert magician who can make cards disappear without anyone noticing? Let us stipulate that he is known not to be—and similarly for other far-fetched objections.

[14] Thanks to Jack Spencer for raising this issue and referring me to Hall's [1999, 694] claim about the surprise exam: "For consider *why* the student is justified in believing that, come Friday, he will justifiably believe that an exam is scheduled: his reason for this is simply that he justifiably believes that, come Friday, *the exam will already have taken place* (and that he will remember this, etc.)."

Having acquitted $(B)$, $(C)$, and $\mathrm{RK}_i$, we now see that the solution for the $n = 2$ Inevitable Event scenario is to reject $(A)$. The (first) student cannot know, or even rationally believe, the teacher's announcement, given that he has the knowledge or beliefs represented in $(B)$ and $(C)$, which he does in this case, and given that he is clever enough to engage in the backward elimination argument. The reason is that knowledge of the teacher's announcement would lead to impossible Moorean knowledge of the form *I have the gold star and I do not know that I have the gold star*, and similarly for rational belief, as shown by Proposition 1.[15]

As noted, this conclusion assumes that we are considering a student clever enough to engage in the backward elimination argument. If we were to consider a different, simpleminded student who would never think of such an argument, then such a student could know the teacher's announcement—he could know that he will be surprised—so we could reject $\mathrm{RK}_1$ and accept $(A)$. (But also note that since the teacher's announcement refers to the future ignorance of the student to whom it is addressed, the content of the announcement would be different if it were addressed to a different student instead of the clever one.)

As suggested at the end of §3, this conclusion also assumes that epistemic conditions (memory, eyesight, reasoning) for the other days/students are known to be good. If instead the clever student thinks he might receive a damaging blow to the head at midnight between day 1 and day 2, or that the student behind him in line might have bad eyesight or lousy reasoning skills, then we could reject $(B)$, $(C)$, or $\mathrm{RK}_2$ and accept $(A)$. But in the "good" $n = 2$ Inevitable Event setup, there seems to be no option but to deny $(A)$.

This shows the danger of relying too much on intuitions about knowability. Against Quine [1953], Sorensen [1988, 312] insists that the student can know the announcement when $n > 1$, since "If you cannot trust your teachers, who can you trust?" But the teacher's announcement in the "good" $n = 2$ Inevitable Event setup is just as unknowable/unbelievable by the student as her announcement in the $n = 1$ case.

Turning to Promised Event setups, suppose the teacher in the designated student scenario never shows the students any stars. Instead, she only claims that she will put a silver star on one of them and a gold star on the other, who will be surprised. In this case, a defender of $(A)$ might argue that denying $(C)$ is not radically skeptical; for now the only source of evidence for $p_1 \vee p_2$ is the teacher's claim, rather than the students' own perception. If student 2 sees a silver star in front of him—call this case $(*)$—then from his perspective, the teacher's claim that $(p_1 \wedge \neg\Box_1 p_1) \vee (p_2 \wedge \neg\Box_2 p_2)$ reduces to $p_2 \wedge \neg\Box_2 p_2$, which is unknowable/unbelievable by student 2. Wright and Sudbury [1977] make the analogous point about the surprise exam scenario. But nothing they say suggests that $p_1 \vee p_2$ is unknowable/unbelievable by student 2. So more must be said against $(C)$. The defender of $(A)$ might argue that since in case $(*)$ student 2 won't know/believe the *whole* of the teacher's announcement, she might not know/believe the *part* of the announcement according to which there is a gold star on one of the students; and then since student 1 cannot rule out $(*)$, $(C)$ is false.

This argument raises questions about the possibility of knowing part of the teacher's announcement without knowing all of it. In many presentations of the surprise exam paradox, the teacher's announcement comes in parts.[16] In other presentations, the teacher only announces "next week there will be a surprise

[15] In fact, on certain coarse-grained views of propositions, knowledge of the teacher's announcement, by someone who knows $(B)$ and $(C)$, would not just *lead to* but would *constitute* impossible Moorean knowledge.

[16] For example: "$K$ knows at time $t$ and thereafter that it is decreed that an event of a given kind will occur uniquely and within $K$'s ken at time $t + i$ for some integer $i$ less than or equal to a specified number $n$, and that it is decreed further that $K$ will not know the value of '$i$' until after (say) time $t + 1 - \frac{1}{2}$" [Quine, 1953, 65]; "A teacher announces to his class that there will be an examination in the afternoon of exactly one of the following $n$ days, where $n$ is some positive integer, and that the examination will take the students by surprise" [Binkley, 1968, 127]; "$X$ is told by his instructor that there will be an examination on either the second or the fourth of the month (but not both), but that he will not anticipate that the test will be given on the day before it is given" [Harrison, 1969, 75]; "A judge decrees on Sunday that a prisoner shall be hanged on noon of the following Monday, Tuesday, or Wednesday, that he shall not be hanged more than once, and that he shall not know until the morning of the hanging the day on which it will occur" [Kaplan and Montague, 1960, 70]; "A teacher announces in class that an examination will be held on some day during the following week, and moreover that the examination will be a

exam." Either way, it seems that one might know there will be an *exam*—since this depends only on the teacher carrying out her intention to give one, which she sincerely expressed—even if one does not know it will be a *surprise*—since this depends on the cooperation of the students' ignorance. Yet I think one can fill in the details of a Promised Event scenario so that $(C)$ is plausibly false. In this sense, the question of whether to reject $(A)$ or $(C)$ does not have a unique answer for the $n = 2$ Promised Event setup.

However, for those who accept $(A)$ and reject $(C)$ in some Promised Event $n = 2$ cases, note that it follows from $(A)$ using PL and $\mathrm{RK}_1$ that $\Box_1(p_1 \vee p_2)$. So rejecting $(C)$ means rejecting $\Box_1(p_1 \vee p_2) \to \Box_1\Box_2(p_1 \vee p_2)$. This is an instance of $4^<$ in Figure 1, which is the principle that I will argue we should reject in the Inevitable Event $n > 2$ case. Thus, the Promised Event and Inevitable Event analyses lead to a common verdict. In §7, I will focus on making the argument against $4^<$ in the Inevitable Event $n > 2$ case. This argument will make crucial use of the result from this section that $(A)$ is the culprit in the Inevitable Event $n = 2$ case.

# 5 Self-Refuting, Anti-Performatory, and Unassimilable Announcements

Before proceeding to the $n > 2$ case, I want to clarify the sense in which the teacher's announcement is *unknowable* in the $n = 2$ Inevitable Event case discussed in §4. To do so, I need to make a distinction between *self-refuting*, *anti-performatory*, and *unassimilable* announcements.

Following some of the earliest commentaries on the surprise exam paradox (O'Connor 1948, Cohen 1950), I will call an announcement of the sentence

(N) I am not speaking now

a *self-refuting* announcement. As Cohen [1950, 86] notes, it could be true that I am not speaking now, and I could silently think about this truth, but any *announcement* of (N) will be a *false* announcement, and it will be false *because of* the announcement. In that sense, any announcement of (N) is *self*-refuting.

In recent years, the term 'self-refuting announcement' has been used for what I will now call an *anti-performatory* announcement (see, e.g., van Benthem 2004, Holliday and Icard 2010). Suppose Ann and Bob attend a party at which the birthdays of all attendees will be revealed at the end of the party. If Ann announces to Bob before the end of the party

(M) You don't know it, but my birthday is in March.

this may be an anti-performatory announcement. As Hintikka [1962, 68-69] explains:

> If you know that I am well informed and if I address the words . . . to you, these words have a curious effect which may perhaps be called anti-performatory. You may come to know that what I say *was* true, but saying it in so many words has the effect of making what is being said false.

Or to put the last point a bit more carefully: one result of Ann's true announcement of (M) is that a *subsequent announcement* of (M) would be *false*, because as a result of Ann's first announcement, Bob *knows* that her birthday is in March. Indeed, as Hintikka points out, Bob can know that all of Ann's first announcement was true, including the part about his ignorance at the time of the announcement. To put the point another way, if we consider an eternalized, non-indexical version of (M),

surprise" [Chow, 1998, 41]; "A teacher announces that he will given an examination within a month. Examinations are always given at noon. He also announces that the exam will be a surprise exam" [Kripke, 2011, 27]

(M$^e$) Bob does not know it at $t$, but Ann's birthday is in March,

where $t$ is the time just before Ann's first announcement of (M), then Bob can come to know (M$^e$) as a result of Ann's announcement of (M).

As I am using the terms, anti-performatory announcements are not self-refuting. Roughly, an *anti-performatory* announcement is a *true* announcement of a sentence with the result that subsequent announcements of the sentence would be false; by contrast, a *self-refuting* announcement is a *false* announcement of a sentence, where the announcement is false *because* the sentence is announced—although the proposition expressed could have been true. Self-refuting announcement do not give listeners knowledge of the proposition expressed, since they are false announcements. By contrast, anti-performatory announcements can give listeners knowledge of the proposition that *was* expressed; but they cannot give listeners knowledge of the proposition that *would be* expressed by a subsequent announcement, since it would be false.[17]

Finally, let us turn to unassimilable announcements. Contrast (M) with

(M′) You won't know it until the end of the party, but my birthday is in March,

announced hours before the end of the party. While Bob can come to know, before the end of the party, that Ann's announcement of (M) was true, can he come to know, before the end of the party, that Ann's announcement of (M′) was true? To put the question more directly: can Bob come to know, before the end of the party, that *he won't know it until the end of the party, but Ann's birthday is in March*? No, he cannot, by the standard argument from Fitch [1963]. In this sense, Ann's announcement of (M′), unlike her announcement of (M), is an *unassimilable* announcement for Bob when she makes it.

Even though Bob cannot come to know the whole content of (M′), can he come to know *that Ann's birthday is in March* from Ann's announcing (M′)? Perhaps if Bob knew that Ann were the kind of person whose announcements always start out with a joke, followed by the word 'but', followed by a real assertion, then he could. But Ann is not that kind of person. At least in many cases, given that Bob will not come to know all of (M′), he will not come to know the part about Ann's birthday either. When this happens, Ann's announcement of (M′) will be unassimilable for Bob but *true*. Thus, the announcement will be unassimilable but *not self-refuting*. Moreover, if Ann's first announcement of (M′) is true, it seems that she can truly announce (M′) again, in which case her initial announcement of (M′) was *not anti-performatory*.[18]

Some early and recent commentators on the surprise exam paradox have tried to analyze the teacher's announcement as a self-refuting announcement [O'Connor, 1948, Cohen, 1950] or an anti-performatory announcement [Gerbrandy, 2007]. According to my analysis, this is a mistake. To see this, consider the $n = 2$ Inevitable Event setup of the designated student paradox.

Suppose that after showing the two students the gold and silver stars, lining the students up, and putting the gold star on the back of student 1, the teacher announces: "One of you has the gold star on his back but does not know it." This would (or could) be an *anti-performatory* announcement. Student 1 would (or could) come to know that he has the gold star—by reasoning that if he did not have it, then student 2 would have known before the teacher's latest announcement that *he* (student 2) had it, contradicting the content of that announcement. But if student 1 comes to know that he has the gold star, then the teacher's announcement is anti-performatory in the sense described above. She cannot truly utter the same sentence

[17] I have been assuming the *eternalist* view that announcements at different times of a present-tense sentence like (M) express different eternal propositions (see, e.g., Richard 1981); but my points could be made using *temporalist* language as well.

[18] In Holliday 2014a, I show how to formalize the dynamic distinction between unassimilable and anti-performatory announcements in a framework I call Sequential Epistemic Logic.

again; but the students can know that when she originally made the announcement, her utterance was true. So the announcement was not self-refuting in the sense described above.

Notice, however, that in the intended version of the paradox, the teacher does *not* announce, "One of you has the gold star on his back but *does not know it*." Instead, she announces, "One of you has the gold star on his back but *will not know it until he pulls the star off his back*."[19] One could even insert a parenthetical addition, as a nod toward analyses of the paradox that appeal to self-reference: "One of you has the gold star on his back but *will not know it* (even after this announcement!) *until he pulls the star off his back*." What the analysis of the previous sections shows is that with the $n = 2$ Inevitable Event setup, this announcement is an *unassimilable* announcement for student 1, assuming he is clever enough to do the backward elimination reasoning. For if he accepts the teacher's announcement, then he will derive that the gold star is not on the back of student 2, in which case the teacher's announcement reduces to

(U1) Student 1 has a gold star on his back but won't know it until he pulls the star off his back[20]

which is, of course, like the unassimilable (M′) above. Because the teacher's announcement is unassimilable for the student, it may well be true and hence not self-refuting. Moreover, it should be possible to make the announcement truly a second time, so it is not anti-performatory.

Similarly, in the surprise exam paradox, if the student accepts the teacher's announcement, then he will derive that the exam will not be on day 2, in which case the teacher's announcement reduces to

(U2) The exam will be on day 1 but you won't know it until the morning of day 1

which is, again, like the unassimilable (M′) above. Because the teacher's announcement is unassimilable for the student, it may well be true and hence not self-refuting. Moreover, it should be possible to make the announcement truly a second time, so it is not anti-performatory.

## 6 Assuming a Surprise vs. Assuming an Intention to Surprise

It is worth stressing one more point before proceeding to the $n > 2$ case. In the classic presentation of the surprise exam paradox, the student reasons using his supposed knowledge/rational belief, acquired from the teacher's announcement, that the exam *will be* a surprise, not just that the teacher *intends* or *will try* to make the exam a surprise. This distinction comes out especially clearly in an *eavesdropping* version of the surprise exam scenario.[21] Suppose that the teacher confidentially tells a colleague in the teacher's lounge that she will give her student a surprise exam sometime in the next term of $n$ days, but unbeknownst to the teacher, her student overhears the conversation through the door to the lounge. Further suppose that the student then starts reasoning about when the exam will be, on the assumption that the teacher intends or will try to make it a surprise, not assuming at the outset that she will necessarily be successful.

Case 1: no one at the school had previously mentioned anything about exams. Then on the basis of overhearing the teacher's comment to her colleague, the student cannot rule out an exam on the last day of the term. For as the student should allow, if the teacher thought that he (the student) had never heard anything about exams at the school, then she could reasonably plan to surprise him by giving an exam on the last day of the term. Thus, in this case, the student cannot even begin the backward elimination argument.

[19] She could add: "So, in particular, you will not know it at time $t$," where $t$ was the time picked out in our reading of $\Box_i\varphi$ for the designated student paradox in §2.

[20] And so won't know it at the time $t$ picked out in our reading of $\Box_i\varphi$ for the designated student paradox in §2.

[21] Thanks to Barteld Kooi for raising the issue of eavesdropping examples in conversation.

Case 2: the school had previously announced a policy of holding an exam each term; in the past, teachers gave students advance notice of the date of each exam, but today the teacher tells her colleague that she plans to spring the exam on her student with no advance notice. Now the student might reason that the exam cannot be on the last day, day $n$, because the teacher should know that if she waits until day $n$ to give the exam, then he (the student) will be expecting an exam on that day, given the school policy. However, if the student thinks the teacher might be unaware that he overheard her conversation, then it seems he should *not* rule out an exam on day $n-1$. For as the student should allow, if the teacher thought that he (the student) had not overheard her conversation, then she could reasonably plan to surprise him by giving him an exam on day $n-1$, since she would have no reason to think that he had ruled out the possibility of an exam on day $n$. So given what the student has overheard, it should seem consistent to him that the teacher intends to make the exam a surprise and that she has scheduled the exam on day $n-1$.

Thus, in both cases, the student should admit that he cannot run the backward elimination argument. By contrast, if he reasons using his supposed knowledge that the exam *will be* a surprise, then he might reason as before: "Since from what I overhead, the exam will be a *surprise*, the teacher cannot wait until day $n$ to give the exam; because if she does, then on the morning of day $n$, my future self, remembering that an exam has not yet occurred, will know that the exam has to be later on day $n$—so it will not be a surprise. Moreover, the teacher cannot wait until day $n-1$ to give the exam; because if she does, then on the morning of day $n-1$, my future self, still knowing the exam will be a *surprise*, will also know that the teacher cannot wait until day $n$ (on the basis of the reasoning I just used to eliminate $n$), and thus, remembering that an exam has not yet occurred, will know that the exam has to be later on day $n-1$—so it will not be a surprise. . . ." Assuming the student knows there will be a surprise exam, it seems that whether he came to know this fact from eavesdropping or from being told directly by the teacher does not matter.[22]

Thus, if the student reasons using his supposed knowledge of a surprise, then my diagnosis of the eavesdropping version of the surprise exam paradox is the same as my diagnosis of the standard version, in which the teacher tells the student directly. Recall that for the Inevitable Event setup of the standard surprise exam paradox with $n = 2$, I argued in §4 that the solution (under certain natural assumptions) is to reject $(A)$, the assumption that the student knows the teacher's announcement, i.e., knows that there will be a surprise exam. I would not deny, however, that the student might come to know from the teacher's announcement that the teacher *intends* or *will try* to make the exam a surprise. Analyzing how the student might reason from *this* assumption is interesting, since it involves analyzing how the student might predict what the teacher might predict about what he (the student) might predict about what she might predict, etc. The surprise exam paradox can be used as a springboard for thinking about this kind of complicated higher-order strategic reasoning, which leads naturally to game-theoretic and probabilistic concepts. But my goal here is more modest: it is to solve the classic version of the surprise exam paradox and related paradoxes in which the student reasons using his supposed knowledge that the relevant event *will be* a surprise.

## 7 The $n > 2$ Case

Where $S_k := (p_k \wedge \neg\Box_k p_k)$, the generalizations of $(A)$, $(B)$, and $(C)$ for arbitrary $n$ are:

$(A^n)$ $\Box_1(S_1 \vee \cdots \vee S_n)$;

[22] But one might reasonably ask: in a case where the student overhears the teacher's comment, and for all he knows, she was not aware of the eavesdropping, does the student really know that the exam will be a surprise, since that depends on the teacher carrying out a successful plan of surprise, which might be compromised by a case of unknown eavesdropping?

$(B^n)$ $\bigwedge_{1<k\leq n} \Box_1((p_k \vee \cdots \vee p_n) \to \Box_k \neg(p_1 \vee \cdots \vee p_{k-1}))$;

$(C^n)$ $\bigwedge_{1<k\leq n} \Box_1 \Box_k (p_1 \vee \cdots \vee p_n)$.

Interestingly, while these are inconsistent in **KJ** for $n = 2$, they are *consistent* even in **S5** for $n > 2$:

**Proposition 3** ($n = 2$ vs. $n > 2$)**.**

1. $\{(A^2), (B^2), (C^2)\} \vdash_{\mathbf{KJ_1}} \bot$.
2. for $n > 2$, $\{(A^n), (B^n), (C^n)\} \nvdash_{\mathbf{S5}} \bot$.

*Proof.* Part 1 is the same as Corollary 1. For part 2, see the end of this section. □

Proposition 3 shows that the $n = 2$ and arbitrary $n$ cases are importantly different, contrary to common opinion in the literature.[23] Yet we shall see that the solution for $n = 2$ leads to the solution for arbitrary $n$.

To formalize the paradox in the $n > 2$ case, we use $\mathbf{K4_1^<}$ (recall Figure 1) instead of **K**. McLelland and Chihara [1975] first identified the role of $4^<$ in formalizing the student's reasoning.[24] The following proposition strengthens their result by using a weaker proof system and set of premises.

**Proposition 4.** For all $n \in \mathbb{N}$, $\{(A^n), (B^n)\} \vdash_{\mathbf{K4_1^<}} \Box_1(p_1 \wedge \neg\Box_1 p_1)$.

*Proof.* Once again, I skip purely propositional steps ('PL' stands for propositional logic) and some applications of RK to theorems:

$(A^n)$ $\Box_1(S_1 \vee \cdots \vee S_n)$ premise

$(B^n)$ $\bigwedge_{1<k\leq n} \Box_1((p_k \vee \cdots \vee p_n) \to \Box_k \neg(p_1 \vee \cdots \vee p_{k-1}))$ premise

$(E^n)$ $\bigwedge_{1<k\leq n} \Box_1 \Box_k (S_1 \vee \cdots \vee S_n)$ from $(A^n)$ by $4_1^<$ and PL

Now we show that the student can rule out a surprise exam on the last day $n$, and if the student has ruled out day $k+1$ and all later days, then he can rule out day $k$ and all later days (for $k \geq 2$):

$(k+1,\, 5)$ $\Box_1 \neg(S_{k+1} \vee \cdots \vee S_n)$ (if $k = n$, let $\neg(S_{k+1} \vee \cdots \vee S_n) := \top$)

$(k,\, 0)$ $\Box_1 \Box_k \neg(S_{k+1} \vee \cdots \vee S_n)$ from $(k+1,\, 5)$ by $4_1^<$ and PL

$(k,\, 1)$ $\big(\Box_k(S_1 \vee \cdots \vee S_n) \wedge \Box_k \neg(p_1 \vee \cdots \vee p_{k-1}) \wedge \Box_k \neg(S_{k+1} \vee \cdots \vee S_n)\big) \to \Box_k p_k$ by PL and $\mathrm{RK}_k$

$(k,\, 2)$ $\Box_1\big[\big(\Box_k(S_1 \vee \cdots \vee S_n) \wedge \Box_k \neg(p_1 \vee \cdots \vee p_{k-1}) \wedge \Box_k \neg(S_{k+1} \vee \cdots \vee S_n)\big) \to \Box_k p_k\big]$ from $(k,\, 1)$ by $\mathrm{Nec}_1$

$(k,\, 3)$ $\Box_1(\Box_k \neg(p_1 \vee \cdots \vee p_{k-1}) \to \Box_k p_k)$ from $(E^n)$, $(k,\, 0)$, and $(k,\, 2)$ using $\mathrm{RK}_1$ and PL

$(k,\, 4)$ $\Box_1 \neg(p_k \wedge \neg\Box_k p_k)$ from $(B^n)$ and $(k, 3)$ using $\mathrm{RK}_1$ and PL

$(k,\, 5)$ $\Box_1 \neg(S_k \vee \cdots \vee S_n)$ from $(k+1,\, 5)$ and $(k,\, 4)$ using $\mathrm{RK}_1$ and PL.

Repeating the above reasoning, we eventually obtain:

[23] For example, Harrison [1969, 74] writes: "The form of the paradox which I shall consider is the two-day case. No loss of generality will result . . . ."

[24] Later authors have also concentrated on $4^<$, including Hall [1999], who analyzes a version of the surprise exam paradox involving justified belief (and degrees of belief) rather than knowledge.

(2, 5) $\Box_1\neg(S_2 \vee \cdots \vee S_n)$

(2, 6) $\Box_1(p_1 \wedge \neg\Box_1 p_1)$ from $(A^n)$ and (2, 5) using $\mathrm{RK}_1$ and PL. □

Thus, by the same argument used for Corollary 1, we have

**Corollary 2.** $\{(A^n), (B^n)\} \vdash_{\mathbf{KJ_1 4^<_1}} \bot$.

We can now apply the solution for $n = 2$ to $n > 2$. Consider the Inevitable Event setup of the designated student paradox. If student $n-1$ sees all silver stars in front of him—call this possibility $(\star)$—then he is in essentially the same epistemic position as the first student in the $n = 2$ case of the story. But we have already seen that the first student in the $n = 2$ case cannot know/rationally believe the teacher's announcement; so in $(\star)$, student $n-1$ cannot know/rationally believe the teacher's announcement either. Therefore, for $n > 2$, since student 1 does not initially know whether $(\star)$ obtains (before he starts reasoning about $n-1$), he cannot know/rationally believe that student $n-1$ knows/rationally believes the teacher's announcement. Thus, we cannot accept $(E^n)$. This has two important upshots. First, if he does not know/rationally believe that student $n-1$ knows/rationally believes the teacher's announcement, then student 1 cannot eliminate $n-1$, so his backward elimination argument is blocked. Second, for all we have seen (including Proposition 3.2 above), student 1 can know/rationally believe the teacher's announcement (unlike in the $n = 2$ case). Thus, we cannot accept axiom $4^<$, an instance of which is $\Box_1(S_1 \vee \cdots \vee S_n) \rightarrow \Box_1\Box_{n-1}(S_1 \vee \cdots \vee S_n)$.

The reason for rejecting $4^<$ is not just worries about "temporal retention" [McLelland and Chihara, 1975, 80], because in the designated student case we are interpreting $4^<$ as an atemporal multi-agent principle: if student 1 knows $\varphi$, then she knows that the other perfect logicians—who also heard the teacher's announcement but may have additional information from perception—know $\varphi$. In §1, this assumption was behind student 1's claim that "student $n-1$, knowing the gold star is a *surprise*, would also know that the teacher did not give the gold star to student $n$ (on the basis of the reasoning I just used to eliminate $n$)." The problem with $4^<$ is that it ignores the possibility of what Sorensen [1988] calls "contingent blindspots," propositions that are knowable by some agents, but not by others in a different epistemic situation, e.g., if that situation includes more information from perception, as in $(\star)$ above. The difference between the present analysis and that of Sorensen [1988, Ch. 8] is that according to the latter, student 1 cannot rule out that student $n$ has the gold star, because the teacher's announcement will be a contingent blindspot for student $n$ if student $n$ sees all silver stars in front of her. But as shown below, student 1 need not assume that student $n$ will know the teacher's announcement in order to rule out that student $n$ has the gold star. Hence the present analysis, unlike that of Sorensen [1988, Ch. 8], accepts the *base case* of student 1's backward elimination argument, at least for the Inevitable Event setup. This analysis is closer to Kripke's [2011, 36-38], except that Kripke's analysis turns on rejecting axiom R, understood as a temporal retention principle, whereas the present analysis turns on rejecting $4^<$ on the basis of contingent blindspots (but see §9 on R). It is easy to show that R is not derivable in $\mathbf{K4}^<$, so by Proposition 4, R is unnecessary to generate the paradox. Our analysis is also similar in spirit to Williamson's [2000], except that Williamson thinks the argument is "a reductio ad absurdum (relative to background assumptions) of the supposition that the pupils know on the first morning that they know on the second morning that ... they know on the last morning that there will be an unexpected examination" (144). We have shown something stronger: we should reject the supposition (strictly weaker than the one mentioned by Williamson) that the pupils know on the first morning that they will know on the *penultimate* morning that there will be an unexpected examination. Note that no formulas of modal depth greater than 3 are used in the proof of Proposition 4, so the higher iterations of knowledge that Williamson mentions are not needed to derive the problematic conclusion.

Let us see how to model the epistemic situation of the three students without $4^<$. Figure 2 displays a model with three "possible worlds," distinguished by which student has the gold star: student 1 ($w_1$), student 2 ($w_2$), or student 3 ($w_3$). The labeled arrows represent the epistemic accessibility relations of students 1, 2, and 3, respectively. (Ignore the outgoing 1-arrows from $w_3$ for the moment.) So for any student $i$ and formula $\varphi$, $\Box_i\varphi$ is true at a world $w$ iff $\varphi$ is true at all worlds $w'$ for which there is an arrow labeled by $i$ pointing from $w$ to $w'$. Since all of the relations are equivalence relations, **S5** is sound with respect to the model, and one can check that $(A^3)$, $(B^3)$, and $(C^3)$ are true at $w_1$. This establishes Proposition 3.2 for $n = 3$, and the construction of models for $n > 3$ is a straightforward generalization.

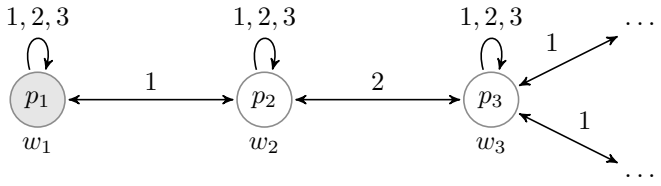


Figure 2: model for Proposition 3.2

World $w_1$ in Figure 2 seems to give a natural representation of a designated student scenario. The gold star is on the back of student 1; it is common knowledge that someone has the gold star; and everyone knows that the student with the gold star, whoever he is, does not know he has it, so he will be surprised to learn that he does—that is, everyone knows the teacher's announcement. What is interesting is that although student 1 knows that student 3 does not have the gold star, student 1 cannot rule out the following ($w_2$): that student 2 has the gold star and, seeing a silver star on the back of student 1, cannot rule out ($w_3$) that student 3 has the gold star and, seeing a silver star on the backs of students 1 and 2, knows that she (student 3) has the gold star. Hence the model falsifies $\Box_1\neg p_3 \to \Box_1\Box_2\neg p_3$ and $(E^3)$ at $w_1$. Incidentally, note that in the world ($w_3$) where student 3 knows that she has the gold star, she should not know (or even believe) that student 1 knows that she (3) has the gold star. There are a number of possibilities for what student 3 could know in this world about what student 1 knows, and I have included the outgoing arrows from $w_3$ as a reminder of this fact. Yet whatever worlds one connects to these arrows to represent what student 1 knows about what 2 knows about what 3 knows about what 1 knows, the model will prove Proposition 3.2.[25]

What this analysis shows is that the mistake in student 1's argument for $n > 2$ is in the induction step, where he assumes that student $n-1$ knows $S_1 \vee \cdots \vee S_n$. However, the *base case* of student 1's argument, where he rules out the possibility that student $n$ has the gold star, does not depend on assuming that students $n-1$ or $n$ know $S_1 \vee \cdots \vee S_n$.[26] As shown by Proposition 5.1 below, assuming the **S5**-consistent $(A^n)$, $(B^n)$, and $(C^n)$ for $n > 2$ (Proposition 3.2), it follows that in the designated student scenario, student 1 can rule out that student $n$ *unknowingly* has the gold star; and in the surprise exam scenario, on day 1 the student can rule out a *surprise* exam on day $n$. If we also assume, as most authors do (see Appendix B), that the student knows/believes that an exam (gold star) will be given on day $n$ (to student $n$) *only if* it is a surprise, then the student can rule out an *exam* (gold star) on day (student) $n$ (Proposition 5.2). In fact, this assumption already follows from $(A^n)$, $(B^n)$, and $(C^n)$ for knowledge in **KT** (Proposition 5.3). Since

[25] Since the modal depth of $(A^n)$, $(B^n)$, and $(C^n)$ is 2, the addition of worlds accessible only from $w_3$ will not affect the truth values of $(A^n)$, $(B^n)$, and $(C^n)$ at $w_1$, since the new worlds will be more than 2 steps away from $w_1$.

[26] As in the $n = 2$ case, compare this to Williamson [2000, 138]: "In ruling out a last-day examination, the pupils assume that they will still know on the last morning that there will be a surprise examination, defined as an examination on a day when the pupils do not know in the morning that there will be an examination that day."

all of these assumptions are consistent even in **S5**, I see no reason to reject the base case. Rejecting $(C^n)$ would motivate rejecting the base case; but as noted for $(C)$ in §3, to reject $(C^n)$ in the Inevitable Event setup would amount to radical skepticism. I would rather accept the base case of the student's argument.

**Proposition 5.**

1. For all $n \in \mathbb{N}$, $\{(A^n), (B^n), (C^n)\} \vdash_{\mathbf{K}} \Box_1 \neg S_n$.
2. For all $n \in \mathbb{N}$, $\{(A^n), (B^n), (C^n), \Box_1(p_n \to \neg\Box_n p_n)\} \vdash_{\mathbf{K}} \Box_1 \neg p_n$.[27]
3. For all $n \in \mathbb{N}$, $\{(A^n), (B^n), (C^n)\} \vdash_{\mathbf{KT}} \Box_1(p_n \to \neg\Box_n p_n) \wedge \Box_1 \neg p_n$.

*Proof.* The proof for part 1 is a straightforward generalization of the first four steps of the proof for Proposition 1. The proof for part 2 simply uses $\Box_1 \neg S_n$, $\Box_1(p_n \to \neg\Box_n p_n)$, PL, and $\mathrm{RK}_1$. The proof for part 3 uses part 2 and the fact that $\{(A^n), (B^n)\} \vdash_{\mathbf{KT}} \Box_1(p_n \to \neg\Box_n p_n)$, which is easy to show with PL and $\mathrm{RK}_1$. □

# 8 Self-undermining Reasoning

I have argued that for the $n > 2$ Inevitable Event setup of the designated student paradox, the clever student 1 makes a subtle but serious mistake: he fails to take into account the possibility, which he cannot eliminate at the beginning of his reasoning, that student $n-1$ sees all silver stars in front of him, in which case he (student $n-1$) would be like student 1 in the $n = 2$ case, blocked (on pain of having impossible Moorean knowledge) from knowing that the teacher's announcement is true—contrary to what student 1 assumes about him (student $n-1$). But is this the clever student's *only* mistake? Perhaps there is more than one.

Another mistake of the clever student in the surprise exam paradox seems to be that he ignores the *self-undermining* nature of his reasoning. His reasoning leads to the conclusion that a surprise exam cannot be given; but if he leaves the matter there and does not reassess, then he is especially susceptible to being surprised by an exam (e.g., on the first day). Something has clearly gone wrong. But to fix the problem, the answer is *not* for the student to restart his reasoning as follows: "having concluded that there cannot be a surprise exam, I am now in a new state of being especially susceptible to a surprise exam. Let's see, which day could this surprise exam come on? It couldn't come on day $n$, because .... And it couldn't come on day $n-1$, because ...." This would lead him right back where he started, with the conclusion that there cannot be a surprise exam, making him especially susceptible to a surprise exam again. His real mistake was not that he stopped reasoning, failing to recognize the self-undermining nature of his reasoning so far. The problem is that he engaged in bad reasoning to his conclusion *in the first place*, as explained above. Repeating that reasoning only makes matters worse, as two (or more) wrongs don't make a right.

There is a second sense in which the clever student's reasoning may seem to be self-undermining: undermining not only the conclusion, but also the premises or intermediate steps of the reasoning. If after drawing the conclusion that there cannot be a surprise exam, the clever student leaves the matter there and does not reassess later, then his *future selves* will not believe the teacher's announcement—but the assumption that his future selves will know the teacher's announcement is a crucial step in his reasoning to that very conclusion.[28] This may indeed be a mistake, but it is *derivative* of the original mistake, described above.

[27] Instead of adding the assumption $\Box_1(p_n \to \neg\Box_n p_n)$, it suffices to add the assumption that student 1 knows there can be at most one exam: $\Box_1 \bigwedge_{1 \le i < j \le n} \neg(p_i \wedge p_j)$.

[28] This problematic aspect of the student's reasoning is emphasized by Shear [2014].

Compare the following objections to the clever student's reasoning when he gets to $n-1$ in the Inevitable Event setup. Objection I: if the teacher waits until day $n-1$ to give the exam, a possibility you haven't yet ruled out, then on the morning of day $n-1$ you will be in the position of student 1 in the $n=2$ case, where you can know that the teacher's announcement is true only if you can have Moorean knowledge, which you cannot; thus, it is a mistake to assume that you will know on day $n-1$ that the teacher's announcement is true. Objection II: you are assuming that on day $n-1$ you will know—and hence believe—that the teacher's announcement is true. But if you continue on this path of reasoning, you will conclude that there can be no surprise exam; and if you leave the matter there and do not reassess later, then you will not believe on day $n-1$ that the teacher's announcement is true, contrary to what you assumed. Note that unlike Objection II, Objection I shows that the clever student has already made a mistake in the $n-1$ step, no matter where his reasoning goes from there. Objection I aims to stop the clever student's reasoning dead in its tracks, so that Objection II need not apply. Also note that unlike Objection I, Objection II is not so clearly applicable to the designated student paradox. Consider how it would go: you (clever student 1) are assuming that student $n-1$ knows—and hence believes—that the teacher's announcement is true. But if you continue on this path of reasoning, you will conclude that there can be no exam; and if you leave the matter there and do not reassess later, then student $n-1$ (?) will not believe that the teacher's announcement is true, contrary to what you assumed. Of course, there is a gap in the reasoning where the question mark appears, since the part about student $n-1$'s beliefs does not immediately follow from the part about clever student 1. Luckily, we need not try to fill the gap, because we can already point out the clever student's mistake in the designated student paradox with the analogue of Objection I, given at the beginning of this section.

# 9 Nonmonotonic Learning

The analysis of §7 not only explains why we should reject $4^<$, but also why we should reject R in Figure 1. To see why, consider a *two-stage* version of the $n=3$ Inevitable Event designated student paradox, proposed to me by Jack Spencer. After the teacher displays the two silver stars and one gold star, she asks student 2 to close his eyes (students 1 and 3 leave their eyes open). She then tells the students that the gold star is a surprise—whoever gets it will not know he has it until the students break out of line—and she sticks the stars on their backs. Let $t_1$ be the time at which all the stars are on the students' backs, but student 2 has his eyes closed. At $t_2$, the teacher tells student 2 to open his eyes so that he can see what is on student 1's back (we can suppose that at $t_1$, student 2 did not know whether he would be allowed to open his eyes).

From the perspective of student 1, this scenario is the same as the standard designated student scenario (we can even suppose that student 1 was unaware of the teacher's instruction that 2 should close his eyes). So if student 1 can rule out that student 3 has the gold star in the standard version, as argued in §7, then student 1 can also rule out that student 3 has the gold star in the two-stage version. Thus, we should have $\Box_1^{t_1}\neg p_3$, i.e., at $t_1$, student 1 knows $\neg p_3$. Moreover, at $t_1$, student 2's epistemic position seems to be essentially the same as student 1's,[29] since both have seen the stars and heard the teacher's announcement but neither has seen what is on anyone's back. Thus, if student 1 can rule out that student 3 has the gold star, so $\Box_1^{t_1}\neg p_3$, then student 2 should also be able to rule out that student 3 has the gold star, so $\Box_2^{t_1}\neg p_3$.

Indeed, for $i,j\in\{1,2\}$, given

$(A^3)_i^{t_j}\ \ \Box_i^{t_j}((p_1\wedge\neg\Box_1^{t_2}p_1)\vee(p_2\wedge\neg\Box_2^{t_2}p_2)\vee(p_3\wedge\neg\Box_3^{t_2}p_3))$

[29] Except that student 2 knows that he is student 2, whereas student 1 knows that he is student 1; but as Jack Spencer also pointed out, this asymmetry can be removed by a further modification of the scenario, which I leave to the reader.

$(B^3_3)^{t_j}_i$ $\Box^{t_j}_i(p_3 \to \Box^{t_2}_3 \neg(p_1 \vee p_2))$

$(C^3_3)^{t_j}_i$ $\Box^{t_j}_i \Box^{t_2}_3 (p_1 \vee p_2 \vee p_3)$,

we have the following simplified analogue of Proposition 5.

**Proposition 6.** For all $i, j \in \{1, 2\}$:

1. $\{(A^3)^{t_j}_i, (B^3_3)^{t_j}_i, (C^3_3)^{t_j}_i\} \vdash_{\mathbf{K}} \Box^{t_j}_i \neg(p_3 \wedge \neg\Box^{t_2}_3 p_3)$.
2. $\{(A^3)^{t_j}_i, (B^3_3)^{t_j}_i, (C^3_3)^{t_j}_i\} \vdash_{\mathbf{KT}} \Box^{t_j}_i \neg p_3$.

*Proof.* Here is a proof, which begins like the proof of Proposition 1:

(1) $\big(\Box^{t_2}_3(p_1 \vee p_2 \vee p_3) \wedge \Box^{t_2}_3 \neg(p_1 \vee p_2)\big) \to \Box^{t_2}_3 p_3$ by PL and $\mathrm{RK}^{t_2}_3$

(2) $\Box^{t_j}_i \big(\Box^{t_2}_3(p_1 \vee p_2 \vee p_3) \wedge \Box^{t_2}_3 \neg(p_1 \vee p_2)\big) \to \Box^{t_2}_3 p_3\big)$ from (1) by $\mathrm{Nec}^{t_j}_i$

(3) $\Box^{t_j}_i \big(\Box^{t_2}_3 \neg(p_1 \vee p_2) \to \Box^{t_2}_3 p_3\big)$ from $(C^3_3)^{t_j}_i$ and (2) using PL and $\mathrm{RK}^{t_j}_i$

(4) $\Box^{t_j}_i \neg(p_3 \wedge \neg\Box^{t_2}_3 p_3)$ from $(B^3_3)^{t_j}_i$ and (3) using PL and $\mathrm{RK}^{t_j}_i$

(5) $\Box^{t_j}_i (p_3 \to \neg(p_1 \vee p_2))$ from $(B^3_3)^{t_j}_i$ using PL, $\mathrm{T}^{t_2}_3$, and $\mathrm{RK}^{t_j}_i$

(6) $\Box^{t_j}_i (p_3 \to \neg\Box^{t_2}_3 p_3)$ from $(A^3)^{t_j}_i$ and (5) using PL and $\mathrm{RK}^{t_j}_i$

(7) $\Box^{t_j}_i \neg p_3$ from (4) and (6) using PL and $\mathrm{RK}^{t_j}_i$. □

Importantly, Proposition 6 does not result from an inconsistency in the premises, as shown by the following result, similar to Proposition 3.2.

**Proposition 7.** For all $i, j \in \{1, 2\}$, $\{(A^3)^{t_j}_i, (B^3_3)^{t_j}_i, (C^3_3)^{t_j}_i\} \nvdash_{\mathbf{S5}} \bot$.[30]

*Proof.* It is an easy exercise to construct a model that makes $(A^3)^{t_j}_i$, $(B^3_3)^{t_j}_i$, and $(C^3_3)^{t_j}_i$ true and in which for each $k \in \{1, 2, 3\}$ and $j \in \{1, 2\}$, the accessibility relation for $\Box^{t_j}_k$ is an equivalence relation. □

The upshot of all this is that at $t_1$, *when student 2 has his eyes closed*, student 2 can (like student 1) know the teacher's announcement that the gold star is a surprise, and he can know that student 3 does not have the gold star. Yet as argued in §7, if student 2 sees a silver star on student 1's back—whether he sees this as soon as the teacher puts the silver star on student 1's back or only later after opening his eyes—then he is like the first student in the $n = 2$ case: if he has the knowledge represented by $(A^3)^{t_2}_2$, $(B^3_3)^{t_2}_2$, $(C^3_3)^{t_2}_2$, then he can have Moorean knowledge of the form *I have the gold star but I don't know that I have it*, as shown by the following result, similar to Proposition 1.

**Proposition 8.**

1. $\{(A^3)^{t_2}_2, (B^3_3)^{t_2}_2, (C^3_3)^{t_2}_2, \Box^{t_2}_2 \neg p_1\} \vdash_{\mathbf{K}} \Box^{t_2}_2 (p_2 \wedge \neg\Box^{t_2}_2 p_2)$.
2. $\{(A^3)^{t_2}_2, (B^3_3)^{t_2}_2, (C^3_3)^{t_2}_2, \Box^{t_2}_2 \neg p_1\} \vdash_{\mathbf{KJ_2}} \bot$.

*Proof.* Continuing the proof of Proposition 6.1 for $i = j = 2$, we have

(8) $\Box^{t_2}_2 (p_2 \wedge \neg\Box^{t_2}_2 p_2)$ from (4), $(A^3)^{t_2}_2$, and $\Box^{t_2}_2 \neg p_1$ using PL and $\mathrm{RK}^{t_2}_2$. □

[30] This also holds with stronger principles $(B^3)^{t_j}_i$ and $(C^3)^{t_j}_i$, based on $(B^3)$ and $(C^3)$, in place of $(B^3_3)^{t_j}_i$ and $(C^3_3)^{t_j}_i$.

Like Proposition 1, Proposition 8 shows that we must reject one of the premises. In a case where student 2 has the gold star, we cannot reject $\Box_2^{t_2}\neg p_1$, since that is part of the very description of $t_2$. So we must reject one of $(A^3)_2^{t_2}$, $(B_3^3)_2^{t_2}$, or $(C_3^3)_2^{t_2}$. Whichever one we reject, we have a significant result: assuming we accept $(A^3)_2^{t_1}$, $(B_3^3)_2^{t_1}$, and $(C_3^3)_2^{t_1}$ (and I see no reason not to accept them), rejecting any one of $(A^3)_2^{t_2}$, $(B_3^3)_2^{t_2}$, or $(C_3^3)_2^{t_2}$ means rejecting the temporal retention principle

R′ $\Box_i^{t_j}\varphi \rightarrow \Box_i^{t_k}\varphi$ $(j < k)$,

which is the same as R in Figure 1 but with the changing indices in the superscripts rather than the subscripts.

Before further discussion of the failure of R′, let us ask: which premise should we reject? As in §4, assuming a normal setup of the story, $(B_3^3)_2^{t_2}$ is unproblematic; and although in a Promised Event setup, $(C_3^3)_2^{t_2}$ may be dubitable, in an Inevitable Event setup, $(A^3)_2^{t_2}$ is the most plausible to blame. Thus, in a case where at $t_2$ student 2 is looking at a silver star on the back of student 1, at $t_2$ student 2 cannot know that the gold star is a surprise. Moreover, if he does not know that the gold star is a surprise, then we cannot conclude that he knows that student 3 does not have the gold star. In particular, replacing $(A^3)_2^{t_2}$ with

$(a^3)_2^{t_2}$ $\Box_2^{t_2}(p_1 \vee p_2 \vee p_3)$,

it is easy to check the following.

**Proposition 9.** $\{(a^3)_2^{t_2}, (B^3)_2^{t_2}, (C^3)_2^{t_2}\} \nvdash_{\mathbf{S5RR'}} \Box_2^{t_2}\neg p_3$.

The argument so far suggests that the following is a true description of the two-stage Inevitable Event designated student scenario:

$$\Box_2^{t_1}\neg p_3 \wedge (\Box_2^{t_2}\neg p_1 \rightarrow \neg\Box_2^{t_2}\neg p_3),$$

which provides another counterexample to R′, since $\Box_2^{t_2}\neg p_1$ can certainly be true.

In the version of the two-stage designated student paradox where student 2 has the gold star (and students 1 and 3 know that (only) student 2 has his eyes closed at $t_1$), we can model the epistemic situation of the three students at $t_1$ as in Figure 3, ignoring some possible higher-order uncertainty (and associated arrows to worlds) that does not affect our analysis. Observe that the teacher's announcement, $S_1 \vee S_2 \vee S_3$, and the fact that student 3 does not have the gold star, $\neg p_3$, are common knowledge. At $w_2$, student 3 knows that the gold star is on student 2's back, but students 1 and 2 do not know whether 1 or 2 has the gold star.

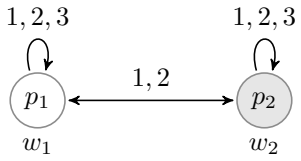


Figure 3: model for time $t_1$ of the two-stage designated student scenario

When student 2 open his eyes (and students 1 and 2 know that he has opened his eyes), the model for $t_2$ should look like the model in Figure 2, except that the shaded actual world should be $w_2$. Note what has happened in the transition between these models: student 2 has gained knowledge of $\neg p_1$, reflected by the fact that 2's uncertainty between $w_1$ and $w_2$ has disappeared; but student 2 has also lost knowledge of the teacher's announcement, $S_1 \vee S_2 \vee S_3$, and of $\neg p_3$, reflected by the fact that 2's uncertainty between $w_2$ and $w_3$ has appeared (note that if this new uncertainty had not appeared, then student 2 would know $p_2$ in the

model for $t_2$). To repeat, the reason student 2 can no longer know the teacher's announcement (if he retains the kind of cleverness that allowed him to know $\neg p_3$ at $t_1$) is that if he did, this knowledge would lead to impossible Moorean knowledge of the form *I have the gold star but I don't know that I have it.*

One of the important lessons of the surprise exam and designated student paradoxes is that the phenomenon of contingent blindspots leads to the phenomenon of nonmonotonic learning: one can lose some knowledge by gaining other knowledge. Of course, there are many other examples of nonmonotonic learning, e.g., in which an agent gains knowledge of some misleading evidence against a proposition, which undermines his previous knowledge of that proposition. This naturally leads to the question: can an agent know a proposition even though (he realizes that) if one of the possibilities compatible with his knowledge obtains, then he will no longer know that proposition? This is closely related to Kripke's [2011, 45] puzzle about *dogmatism*, about whether an agent who knows a proposition $P$ knows that all future evidence against $P$ is misleading (which is, after all, implied by the truth of $P$), a puzzle that Kripke says he was led to by considering the surprise exam paradox (45n17). Kripke's conclusion is this: "The commensense view is, for example, that you *do* know that I have written certain papers on modal logic but that future evidence could lead you to change your mind about this. So, you should rationally leave yourself open to such changes-of-mind, even though it is the case that you *know* that I wrote these papers. The question is why?" (45).

Returning to the two-stage designated student paradox, note that if at $t_1$ the teacher were to tell student 2 that he will open his eyes at $t_2$, and if student 2 were convinced by the analysis of this paper, then he would realize at $t_1$ that if one of the possibilities compatible with his knowledge obtains at $t_2$—namely, the possibility in which he opens his eyes and sees a silver star on the the back of student 1—then at $t_2$ he will no longer know that the gold star is a surprise.[31] Is this compatible with student 2 knowing at $t_1$ that the gold star is a surprise? Kripke's commonsense view suggests that the answer is *yes*. Similarly, in the original surprise exam paradox for $n = 3$, on the morning of day 1 the student can know the teacher's announcement and yet, being convinced by the analysis of this paper, realize that if there is no exam later on day 1, then on the morning of day 2 he will no longer know the teacher's announcement.[32]

## 10 Other Recalcitrant Variations?

The analysis developed so far of the surprise exam and designated student paradoxes also handles Sorensen's [1984] two other "recalcitrant variations": the *sacrificial virgin paradox* and the *paradox of the undiscoverable position*. I treat the paradox of the undiscoverable position elsewhere [Holliday, 2014b], so here I will go over only the sacrificial virgin paradox. The following is Sorensen's [1984, 361] description of the paradox:

> Every fifty years the inhabitants of a tropical paradise sacrifice a virgin to the local volcano in an elaborate ceremony. Virgins from all around are blindfolded and brought before the volcano. They all hold hands in a line and can only communicate one sentence: 'No one to your right is the sacrificial virgin'. This sentence can only be signalled by squeezing the hand of the virgin to one's left. The virgins are reliable and dutibound to so signal if and only if it is known to be true. Besides all this, the virgins also know that a necessary condition for being the sacrificial virgin is that one remain ignorant of the honour until one is tossed in. The chief must take the leftmost virgin up to the mouth of the volcano, and if the offering is acceptable, push her in and tell the

[31] Note that in this case, student 2 would not be in essentially the same epistemic position as student 1 at time $t_1$.

[32] As the reader may have been thinking all along in this section, since it reintroduces temporal aspects, the multi-stage designated student scenario is very much like the original surprise exam scenario.

> rest of the virgins to go home. If the offering is unacceptable, he sends that virgin home and repeats the procedure with the new leftmost virgin. This procedure continues until one virgin is sacrificed, so it is known that one will be sacrificed.[33] After hearing the announcement that one virgin will be sacrificed, someone objects that the ceremony cannot take place: The rightmost virgin knows she is rightmost since her righthand is free. She knows that if she is offered, then none of the virgins to her left have been sacrificed. So if she is the sacrificial virgin then she will have to be offered knowing that she is the only alternative remaining, and thus would know she is the sacrificial virgin. Since the sacrificial virgin must not know, the rightmost virgin knows that she is not the sacrificial virgin. This knowledge obliges her to squeeze the hand of the virgin to her immediate left signaling the sentence 'None of the virgins to your right is the sacrificial virgin'. This virgin is either the leftmost virgin or a middle virgin (a middle virgin is any virgin between the leftmost and rightmost virgins). If she is a middle virgin, she will reason that if she is offered, she will know that none of the virgins to her left have been sacrificed. *By the signal she knows that none to her right are sacrificial virgins, and thus she will be able to deduce that she will be sacrificed. But since the sacrificial virgin cannot know she will be sacrificed, this middle virgin knows she will not be sacrificed.* Therefore, she will squeeze the hand of the virgin to her left, triggering the same deduction if this third virgin is a middle virgin .... [emphasis added]

The mistake in the reasoning can be seen by first considering the two virgin case. In this case, the leftmost virgin cannot know, at the time that her right hand is squeezed, both that the rules of the ceremony will hold and that none to her right are sacrificial virgins. For if she did, then given her knowledge that she is leftmost (since she can feel her left hand free), she would have Moorean knowledge of the form *I am the sacrificial virgin but I don't know it*, which is impossible. Now consider the case of more than two virgins. Here the mistake in the reasoning is to assume (what is cleverly omitted from Sorensen's explicit description, although it must be assumed) that if the second-to-rightmost virgin knows, *at the time that her right hand is squeezed*, both that the rules of the ceremony will hold and that none to her right are sacrificial virgins, then she also knows that she *will* know, *at the time that she is brought before the volcano*, both that the rules of the ceremony will hold and that none to her right are sacrificial virgins. The problem is that if she is brought before the volcano, a possibility that she has not eliminated at the beginning of her reasoning, then she will be like the leftmost virgin in the two virgin case, who cannot know both that the rules of the ceremony will hold and that none to her right are sacrificial virgins. Thus, the second-to-rightmost virgin does not know, just after her right hand is squeezed, that she will retain the relevant knowledge; thus, she does not know that she is not the sacrificial virgin; and thus, she should not, according to the rules of the ceremony, squeeze the hand of the virgin to her left. All of this can be formalized for precision, but I trust that the reader can now do so on his or her own, having seen the templates already provided.

## 11 Conditionally Expected Exams

In this section, I consider another version of the surprise exam paradox, due to Ayer [1973] and Williamson [1992, 2000, Ch. 6], that removes from the teacher's announcement the "existential assumption" that *there will be* a surprise exam. Referring to the unexpected execution scenario, Ayer writes: "Suppose that the condition were that the man should not know, on the day set for his execution, that he was to be executed on that day if he was to be executed at all" (125). Ayer notes that this still leads to a paradox, but he does

[33] My footnote: this suggests an Inevitable Event setup.

not discuss it further. By contrast, Williamson discusses this version in detail. First, for a definition, he says that an exam is *conditionally expected* iff on the morning of the exam, the student knows that *if* there is an exam at all during the term, it will be that day. Second, he supposes that the student knows that at most one exam can be given during the term,[34] and that the teacher announces only that *there will be no conditionally expected exam.* Now the student reasons that there cannot be an exam on the last day of the term, for then it would be conditionally expected. Hence there cannot be an exam on the penultimate day either, for then it would also be conditionally expected, since the possibility of an exam on the last day has already been eliminated. Repeating this backward elimination argument, the student concludes that there will be no exam. But according to Williamson [2000], "That is absurd, for we may stipulate that in fact quite some time before the end of term there will be an examination that is not conditionally expected" (144).

To formalize Williamson's argument in the $n = 2$ case, I use the following premises:

$(F)$ $\Box_1(\neg(p_1 \wedge \Box_1((p_1 \vee p_2) \to p_1)) \wedge \neg(p_2 \wedge \Box_2((p_1 \vee p_2) \to p_2)))$;

$(G)$ $\Box_1(\neg p_1 \to \Box_2 \neg p_1)$;

$(H)$ $\Box_1 \neg(p_1 \wedge p_2)$.

$(F)$ states that the student knows the teacher's announcement that there will be no conditionally expected exam; $(G)$ states that he knows that he will remember on day 2 if there was no exam on day 1; and $(H)$ states that he knows there will be at most one exam. All of these premises can be reinterpreted for a designated student version of Williamson's scenario. Whatever the interpretation, we have the following result:

**Proposition 10.**

1. $\{(F), (G), (H)\} \nvdash_{\mathbf{S5R}} \bot$.
2. $\{(F), (G), (H)\} \vdash_{\mathbf{K}} \Box_1 \neg p_2$.
3. $\{(F), (G), (H)\} \vdash_{\mathbf{KT}} \neg(p_1 \vee p_2)$ and hence $\{\Box_1((F) \wedge (G) \wedge (H))\} \vdash_{\mathbf{KT}} \Box_1 \neg(p_1 \vee p_2)$.

*Proof.* It is easy to check that a model with a single world where $p_1$ and $p_2$ are false, with reflexive accessibility relations for $\Box_1$ and $\Box_2$, demonstrates the first part of the proposition. For the second and third parts, here is a proof, skipping purely propositional steps and some applications of RK to theorems:

(1) $\Box_2 \neg p_1 \to \Box_2((p_1 \vee p_2) \to p_2)$ by PL and $\mathrm{RK}_2$

(2) $\Box_1\big(\Box_2 \neg p_1 \to \Box_2((p_1 \vee p_2) \to p_2)\big)$ from (1) by $\mathrm{Nec}_1$

(3) $\Box_1\big(\neg p_1 \to \Box_2((p_1 \vee p_2) \to p_2)\big)$ from $(G)$ and (2) using $\mathrm{RK}_1$ and PL

(4) $\Box_1(\neg p_1 \to \neg p_2)$ from $(F)$ and (3) using $\mathrm{RK}_1$ and PL

(5) $\Box_1 \neg p_2$ from $(H)$ and (4) using $\mathrm{RK}_1$ and PL

(6) $\Box_1((p_1 \vee p_2) \to p_1)$ from (5) using $\mathrm{RK}_1$ and PL

(7) $\neg(p_1 \wedge \Box_1((p_1 \vee p_2) \to p_1))$ from $(F)$ by T and PL

(8) $\neg p_1$ from (6) and (7) by PL

[34] Williamson [2000, 144] remarks that this assumption can be removed. For simplicitly, here I follow Williamson's actual presentation that includes the assumption.

(9) $\neg p_2$ from (5) by T and PL

(10) $\neg(p_1 \vee p_2)$ from (8) and (9) by PL

(11) $((F) \wedge (G) \wedge (H)) \rightarrow \neg(p_1 \vee p_2)$ from previous steps by PL

(12) $\Box_1((F) \wedge (G) \wedge (H)) \rightarrow \Box_1\neg(p_1 \vee p_2)$ from (11) by $\mathrm{RK}_1$. □

Note that if we had the 4 axiom, the "KK principle," then the antecedent of line (12) would be derivable from the antecedent of line (11), so $\Box_1\neg(p_1 \vee p_2)$ would be derivable from $(F)$, $(G)$, and $(H)$. But we encounter a problem even earlier. Since Williamson says "we may stipulate that in fact quite some time before the end of term there will be an examination that is not conditionally expected," we already have a problem at line (10). The assumption of $p_1 \vee p_2$ is inconsistent with $(F)$, $(G)$, and $(H)$ in **KT**, so if we make Williamson's stipulation, then we must reject one of those premises. If the student knows that she has a decent memory and that there is a school rule according to which at most one exam may be given within $n$ days (here $n = 2$), then $(G)$ and $(H)$ are unproblematic. The unique solution in this case is to reject $(F)$. Once again, the student does not know the teacher's announcement for $n = 2$ (recall §4).

Let us now consider $n > 2$. The generalizations of $(F)$, $(G)$, and $(H)$ for arbitrary $n$ are:

$(F^n)$ $\Box_1\big(\bigwedge\limits_{1\leq k\leq n} \neg(p_k \wedge \Box_k((p_1 \vee \cdots \vee p_n) \rightarrow p_k))\big)$;

$(G^n)$ $\bigwedge\limits_{1<k\leq n} \Box_1(\neg(p_1 \vee \cdots \vee p_{k-1}) \rightarrow \Box_k\neg(p_1 \vee \cdots \vee p_{k-1}))$;

$(H^n)$ $\Box_1\big(\bigwedge\limits_{1\leq i<j\leq n} \neg(p_i \wedge p_j)\big)$.

As before, we find that there is an important difference between the $n = 2$ and $n > 2$ cases. Compare the following result to Proposition 3.

**Proposition 11** ($n = 2$ vs. $n > 2$ again)**.**

1. $\{(F^2), (G^2), (H^2)\} \vdash_{\mathbf{KT}} \neg(p_1 \vee p_2)$.
2. For $n > 2$, $\{(F^n), (G^n), (H^n)\} \nvdash_{\mathbf{S5}} \neg(p_1 \vee \cdots \vee p_n)$.
   Hence $\{(F^n), (G^n), (H^n)\} \nvdash_{\mathbf{S5}} \Box_1\neg(p_1 \vee \cdots \vee p_n)$.

*Proof.* Part 1 is the same as Proposition 10.3. For part 2, the **S5** model in Figure 2 establishes the $n = 3$ case, since $(F^3)$, $(G^3)$, and $(H^3)$ are all true at $w_1$, whereas $\neg(p_1 \vee p_2 \vee p_3)$ is false, and the models for higher $n$ are straightforward generalizations. □

As before, we also find that for $n > 2$, adding the $4^<$ axiom generates a problem. Compare the following result with Proposition 4.

**Proposition 12.**

1. For all $n \in \mathbb{N}$, $\{(F^n), (G^n), (H^n)\} \vdash_{\mathbf{K4_1^<}} \Box_1\neg(p_2 \vee \cdots \vee p_n)$.
2. For all $n \in \mathbb{N}$, $\{(F^n), (G^n), (H^n)\} \vdash_{\mathbf{KT_1 4_1^<}} \neg(p_1 \vee \cdots \vee p_n)$.
   Hence $\{\Box_1((F^n) \wedge (G^n) \wedge (H^n))\} \vdash_{\mathbf{KT_1 4_1^<}} \Box_1\neg(p_1 \vee \cdots \vee p_n)$.

*Proof.* Here is a proof, skipping purely propositional steps and some applications of RK to theorems:

$(F^n)$ $\Box_1\big(\bigwedge\limits_{1\leq k\leq n} \neg(p_k \wedge \Box_k((p_1 \vee \cdots \vee p_n) \to p_k))\big)$ premise

$(G^n)$ $\bigwedge\limits_{1<k\leq n} \Box_1(\neg(p_1 \vee \cdots \vee p_{k-1}) \to \Box_k\neg(p_1 \vee \cdots \vee p_{k-1}))$ premise

$(H^n)$ $\Box_1\big(\bigwedge\limits_{1\leq i,j\leq n\ i\neq j} \neg(p_i \wedge p_j)\big)$ premise.

Now we show that the student can rule out an exam on the last day $n$, and if the student has ruled out day $k+1$ and all later days, then he can rule out day $k$ and all later days (for $k \geq 2$):

$(k+1,5)$ $\Box_1\neg(p_{k+1} \vee \cdots \vee p_n)$ (if $k = n$, let $\neg(p_{k+1} \vee \cdots \vee p_n) := \top$)

$(k,0)$ $\Box_1\Box_k\neg(p_{k+1} \vee \cdots \vee p_n)$ from $(k+1,5)$ by $4_1^<$ and PL

$(k,1)$ $\big(\Box_k\neg(p_1 \vee \cdots \vee p_{k-1}) \wedge \Box_k\neg(p_{k+1} \vee \cdots \vee p_n)\big) \to \Box_k((p_1 \vee \cdots \vee p_n) \to p_k)$ by $\mathrm{RK}_k$ and PL

$(k,2)$ $\Box_1\big(\big(\Box_k\neg(p_1 \vee \cdots \vee p_{k-1}) \wedge \Box_k\neg(p_{k+1} \vee \cdots \vee p_n)\big) \to \Box_k((p_1 \vee \cdots \vee p_n) \to p_k)\big)$ from $(k,1)$ by $\mathrm{Nec}_1$

$(k,3)$ $\Box_1\big(\neg(p_1 \vee \cdots \vee p_{k-1}) \to \Box_k((p_1 \vee \cdots \vee p_n) \to p_k)\big)$ from $(G^n)$, $(k,0)$, and $(k,2)$ using $\mathrm{RK}_1$ and PL

$(k,4)$ $\Box_1\neg p_k$ from $(H^n)$, $(F^n)$, and $(k,3)$ using $\mathrm{RK}_1$ and PL

$(k,5)$ $\Box_1\neg(p_k \vee \cdots \vee p_n)$ from $(k,0)$ and $(k,4)$ using $\mathrm{RK}_1$ and PL

Repeating this reasoning, we eventually obtain the following, which proves part 1:

$(2,5)$ $\Box_1\neg(p_2 \vee \cdots \vee p_n)$.

$(2,6)$ $\Box_1((p_1 \vee \cdots \vee p_n) \to p_1)$ from $(2,5)$ using $\mathrm{RK}_1$ and PL

$(2,7)$ $\neg(p_1 \wedge \Box_1((p_1 \vee \cdots \vee p_n) \to p_1))$ from $(F^n)$ by $\mathrm{T}_1$ and PL

$(2,8)$ $\neg p_1$ from $(2,6)$ and $(2,7)$ by PL

$(2,9)$ $\neg(p_2 \vee \cdots \vee p_n)$ from $(2,5)$ by $\mathrm{T}_1$ and PL

$(2,10)$ $\neg(p_1 \vee \cdots \vee p_n)$ from $(2,8)$ and $(2,9)$ by PL

This completes the proof of part 2. □

We have derived $(2,10)$, which is inconsistent with Williamson's stipulation that an exam occurs, from $(F^n)$, $(G^n)$, and $(H^n)$ with the help of $4^<$. Which of these assumptions is to blame? For reasons noted above, $(G^n)$ and $(H^n)$ seem unproblematic. So we can narrow the suspects down to $(F^n)$ and $4^<$.

As in §7, we can apply the solution for the $n = 2$ case to the $n > 2$ case. For ease of exposition (and to suggest connections with the designated student scenario), let us refer to the student on day $i$ as 'student $i$'. The mistake in the clever student's reasoning can be seen as follows. If there *is* an exam and it falls on day $n$—call this possibility (#)—then student $n-1$ is essentially in the epistemic position of student 1 in the $n = 2$ case of the story in which there is an exam. But we have already seen with our analysis of Proposition 10.3 that student 1 in the $n = 2$ case cannot know the teacher's announcement if there is indeed an exam; so in (#), student $n-1$ cannot know the teacher's announcement either. Therefore, for $n > 2$, since student 1 does not initially know whether (#) obtains, he cannot know that student $n-1$ knows the teacher's announcement. Yet for all we have seen (including Proposition 11.2), student 1 can know the teacher's announcement (unlike in the $n = 2$ case), as in $(F^n)$. Hence we should not accept axiom $4^<$.

However, unlike the proof for Proposition 4, the proof for Proposition 12 does not explicitly use the assumption that student 1 knows that student $n-1$ knows the teacher's announcement. For the only application of $4^<$ is in the steps from $(k+1,5)$ to $(k,0)$. Note, though, that the derivation of $(k+1,5)$ uses the fact that student 1 knows the teacher's announcement in $(k+1,4)$. Now if student $n-1$ does not know the teacher's announcement, it would seem to be illegitimate to "project" student 1's knowledge in $(k+1,5)$ to student $n-1$ in $(k,0)$; for if student 1 does not know that student $n-1$ knows the teacher's announcement, why assume that student 1 knows that student $n-1$ knows something *that student 1 came to know from knowing the teacher's announcement*? We should not, so again we should not accept $4^<$.

This diagnosis, blaming the iteration principle $4^<$, is similar in spirit to that of Williamson [2000, §6.2], though it differs in the details. According to Williamson, the argument involving conditionally expected exams "uses the supposition that the pupils know on the first morning that they know on the second morning that ... they know on the last morning that there will be no conditionally expected examination to show that there will be no examination at all" (144). However, as our proof of Proposition 12 shows, for any $n > 2$, no formulas of modal depth greater than 2 are required to derive the conclusion, $\neg(p_1 \vee \cdots \vee p_n)$, which is inconsistent with Williamson's stipulation that an exam occurs; and no formulas of modal depth greater than 3 are required to derive $\Box_1(\neg p_1 \vee \cdots \vee p_n)$. Thus, the higher iterations of knowledge that Williamson mentions need not be assumed in order to derive the problematic conclusions.

Finally, note that although we should reject the induction step from $(k+1,5)$ to $(k,0)$, as in §7 there is no need to reject the base case of the students reasoning. Compare the following result to Proposition 5.2-3.

**Proposition 13.** For all $n \in \mathbb{N}$, $\{(F^n), (G^n), (H^n)\} \vdash_{\mathbf{K}} \Box_1 \neg p_n$.

*Proof.* Observe that in the proof of Proposition 12, the proof of $(n,5)$ only requires $\mathbf{K}$, not $4^<$. □

## 12 Conclusion

Several versions of the surprise exam paradox admit of a simple and uniform solution: contrary to Sorensen [1988], in the $n = 2$ case the teacher's announcement is unknowable/unbelievable by the (first) student; and *because* of that fact, $4^<$ should be rejected in the $n > 2$ case; thus, contrary to Quine [1953], the solution in the $n > 2$ case does not involve denying that the (first) student can know the teacher's announcement, and contrary to Sorensen [1988], it does not involve denying the base case of the student's reasoning.

## References


A.J. Ayer. On a Supposed Antinomy. *Mind*, 82(325):125–126, 1973.

Johan van Benthem. What One May Come to Know. *Analysis*, 64(2):95–105, 2004.

Robert Binkley. The Surprise Examination in Modal Logic. *The Journal of Philosophy*, 65(5):127–136, 1968.

Timothy Y. Chow. The Surprise Examination or Unexpected Hanging Paradox. *American Mathematical Monthly*, 105:41–51, 1998.

L.J. Cohen. Mr. O'Connor's "Pragmatic Paradoxes". *Mind*, 59(223):85–87, 1950.

Charles B. Cross. The Paradox of the Knower without Epistemic Closure. *Mind*, 110(438):319–333, 2001.

Fred Dretske. Epistemic Operators. *The Journal of Philosophy*, 67(24):1007–1023, 1970.

Fred Dretske. *Contemporary Debates in Epistemology*, chapter The Case against Closure, pages 13–26. Blackwell Publishing, Malden, MA, 2005.

Michael Fara. Knowability and the capacity to know. *Synthese*, 173:53–73, 2010.

Frederic B. Fitch. A Logical Analysis of Some Value Concepts. *The Journal of Symbolic Logic*, 28(2):135–142, 1963.

Jelle Gerbrandy. The Surprise Examination in Dynamic Epistemic Logic. *Synthese*, 155:21–33, 2007.

Ned Hall. How to Set a Surprise Exam. *Mind*, 108(432):647–703, 1999.

Craig Harrison. The Unanticipated Examination in View of Kripke's Semantics for Modal Logic. In J.W. Davis, D.J. Hockney, and W.K. Wilson, editors, *Philosophical Logic*, pages 74–88. D. Reidel Publishing Company, 1969.

Jaakko Hintikka. *Knowledge and Belief: An Introduction to the Logic of the Two Notions*. Cornell: Cornell University Press, 1962.

Wesley H. Holliday. Knowledge, Time, and Paradox: Introducing Sequential Epistemic Logic. Manuscript, 2014a.

Wesley H. Holliday. On Being in an Undiscoverable Position. Manuscript, 2014b.

Wesley H. Holliday and Thomas F. Icard. Moorean Phenomena in Epistemic Logic. In Lev Beklemishev, Valentin Goranko, and Valentin Shehtman, editors, *Advances in Modal Logic*, volume 8, pages 178–199. College Publications, 2010.

Wesley H. Holliday and John Perry. Roles, Rigidity, and Quantification in Epistemic Logic. In Alexandru Baltag and Sonja Smets, editors, *Johan van Benthem on Logic and Information Dynamics*. Springer, Dordrecht, 2014. forthcoming.

David Kaplan and Richard Montague. A Paradox Regained. *Notre Dame Journal of Formal Logic*, 1:79–90, 1960.

Saul A. Kripke. *Philosophical Troubles*, chapter On Two Paradoxes of Knowledge, pages 27–51. Oxford University Press, 2011.

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

D.H. Makinson. The Paradox of the Preface. *Analysis*, 25(6):205–207, 1965.

James McLelland and Charles Chihara. The Surprise Examination Paradox. *Journal of Philosophical Logic*, 4:71–89, 1975.

Robert Nozick. *Philosophical Explanations*. Harvard University Press, Cambridge, MA, 1981.

D.J. O'Connor. Pragmatic Paradoxes. *Mind*, 57(227):358–359, 1948.

John Perry. Frege on Demonstratives. *Philosophical Review*, 86:474–497, 1977.

W.V. Quine. On a So-Called Paradox. *Mind*, 65-67(245):217–242, 1953.

W.V. Quine. Quantifiers and Propositional Attitudes. *The Journal of Philosophy*, 53(5):177–187, 1956.

Mark Richard. Temporalism and Eternalism. *Philosophical Studies*, 39(1):1–13, 1981.

Niels Rustenburg. The Surprise Examination Paradox Examined. BS Thesis, Utrecht University, 2014.

Ted Shear. Stability and the Prediction Paradox. Manuscript, 2014.

Brian Skyrms. The Explication of "X knows that p". *The Journal of Philosophy*, 64(12):373–389, 1967.

Roy Sorensen. Recalcitrant Variations of the Prediction Paradox. *Australasian Journal of Philosophy*, 69 (4):355–362, 1984.

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

Timothy Williamson. Inexact Knowledge. *Mind*, 101(402):217–242, 1992.

Timothy Williamson. *Knowledge and its Limits*. Oxford University Press, 2000.

Crispin Wright and Aidan Sudbury. The Paradox of the Unexpected Examination. *Australasian Journal of Philosophy*, 55(1):41–58, 1977.

## Appendix A: Unnecessary Assumptions

In this appendix, I return to the points about the unnecessary assumptions (i) and (ii) in §3. First, Figure 4 displays a model that proves Proposition 2 when we interpret $\Box_1$ and $\Box_2$ as explained in §7. It is easy to see that $\mathbf{KD4}^{<}\mathbf{R}$ is sound with respect to this model: $\mathbf{D}$ is valid on any frame satisfying the seriality condition $\forall x \exists y R_i xy$; $\mathbf{4}^{<}$ is valid on any frame satisfying $\forall x, y, z((R_1 xy \wedge R_2 yz) \rightarrow R_1 xz)$; $\mathbf{R}$ is valid on any frame satisfying $\forall x, y(R_2 xy \rightarrow R_1 xy)$; and the model in Figure 4 satisfies these conditions. The model in Figure 4 is not supposed to represent a natural scenario, in the way that the model in Figure 2 does. Indeed, since the model in Figure 4 satisfies $\Box_1(p_1 \wedge \neg\Box_1 p_1)$ at the left world, it is unnatural. But the point is that $\Box_1\neg(p_1 \vee p_2) \vee (p_1 \wedge \neg\Box_1 p_1) \vee (p_2 \wedge \neg\Box_2 p_2)$ is false at the left world; so the model establishes the fact that $\{(A), (B), (C)\} \nvdash_{\mathbf{KD4}^{<}\mathbf{R}} \Box_1\neg(p_1 \vee p_2) \vee (p_1 \wedge \neg\Box_1 p_1) \vee (p_2 \wedge \neg\Box_2 p_2)$, showing that assumptions (i) and (ii) in §3 are not necessary for generating the paradoxical result that $\Box_1(p_1 \wedge \neg\Box_1 p_1)$ from $(A)$, $(B)$, and $(C)$.

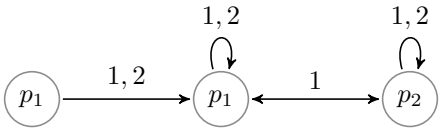


Figure 4: model for Proposition 2

# Appendix B: Previous Results for $n = 2$

In this appendix, I compare Proposition 1 of §3 with previous results in the literature for $n = 2$, concluding that Proposition 1 gives a stronger result in terms of the weakness of its proof system and premises.

As noted in §3, previous formalizations of the paradox use extensions of **K** with axiom schemas such as:

| | | | |
|---|---|---|---|
| T | $\Box_i\varphi \to \varphi$ | 4 | $\Box_i\varphi \to \Box_i\Box_i\varphi$ |
| D | $\Box_i\varphi \to \neg\Box_i\neg\varphi$ | $4^<$ | $\Box_i\varphi \to \Box_i\Box_j\varphi \ \ (i < j)$ |
| R | $\Box_i\varphi \to \Box_j\varphi \ \ (i < j)$ | $4^\leq$ | $\Box_i\varphi \to \Box_i\Box_j\varphi \ \ (i \leq j)$ |

Using the notation for extensions of **K** from §3, we can switch between thinking of $\mathbf{K\Sigma_1}\ldots\mathbf{\Sigma_n}$ as a proof system and as the set of theorems of that proof system. Taking the latter perspective, the following fact compares the strength of all the systems we will consider. Recall that the J axiom schema is $\Box_i\neg\Box_i\varphi \to \neg\Box_i\varphi$.

**Fact 1** (Comparing Systems)**.**

1. $\mathbf{KD} \subsetneq \mathbf{KJ} \subsetneq \mathbf{KT}$;
2. $\mathbf{KJ} \subsetneq \mathbf{KJ4} = \mathbf{KD4}$;
3. $\mathbf{KJ4^<} \subsetneq \mathbf{KJ4^\leq} = \mathbf{KD4^\leq} \subsetneq \mathbf{KD4R}$;
4. $\mathbf{KJ4^<} \subsetneq \mathbf{KT4^<} \subsetneq \mathbf{KT4^<R} \subsetneq \mathbf{KT4R}$.

For each part, we leave it to the reader to verify that $\mathbf{L} \subseteq \mathbf{L'}$, by deriving the axioms of **L** in $\mathbf{L'}$, or to verify $\mathbf{L} \not\subseteq \mathbf{L'}$ semantically, by showing that an axiom of **L** can be falsified in a modal model for which $\mathbf{L'}$ is sound.[35] The Facts below are also left to the reader as elementary exercises in finding proofs and countermodels.

In addition to using different extensions of **K**, other authors use different premises in formalizing the paradox. Harrison [1969] and McLelland and Chihara [1975] use the following premises (where $\veebar$ is exclusive disjunction) and prove the results in Proposition 14:

$P_1$ $\Box_1(p_1 \veebar p_2)$;

$P_2$ $\Box_1(\neg\Box_1 p_1 \wedge \neg\Box_2 p_2)$;

$P_3$ $\Box_1(\neg p_1 \to \Box_2\neg p_1)$.

**Proposition 14.**

1. (Harrison 1969) $\{P_1, P_2, P_3\} \vdash_{\mathbf{KT4R}} \bot$;
2. (McLelland and Chihara 1975) $\{P_1, P_2, P_3\} \vdash_{\mathbf{KT4^<R}} \bot$.

Corollary 1 is a stronger inconsistency result than Proposition 14 in the following sense:

**Fact 2** (Comparison with Harrison, McLelland and Chihara)**.**

1. $\mathbf{KJ} \subsetneq \mathbf{KT4^<R} \subsetneq \mathbf{KT4R}$;
2. $\{P_1, P_2\} \vdash_{\mathbf{K}} (A)$, $\{P_1, P_3\} \vdash_{\mathbf{K}} (B)$, and $\{P_1\} \vdash_{\mathbf{K4^<}} (C)$;

[35] The semantic proofs are straightforward, since all of the axioms in Fact 1 correspond to first-order conditions on the accessibility relations $R_i$. The only unfamiliar case is the J axiom, which corresponds to $\forall x\exists y(R_ixy \wedge \forall z(R_iyz \to R_ixz))$.

3. $\{(A),(B)\} \nvdash_{\mathbf{KD4R}} P_1 \vee P_2$ and $\{(A),(B),(C)\} \nvdash_{\mathbf{KD4}^{<}\mathbf{R}} P_1 \vee P_2$.

In place of $P_2$, Binkley [1968] uses

$P_2'$ $\Box_1((p_1 \to \neg\Box_1 p_1) \wedge (p_2 \to \neg\Box_2 p_2))$,

which is a weakening of Harrison's in the following sense:

**Fact 3.** $\vdash_{\mathbf{K}} P_2 \to P_2'$ and $\vdash_{\mathbf{KT}} P_2' \to P_2$, but $\nvdash_{\mathbf{KDR}} (P_1 \wedge P_2') \to P_2$.

Thinking of $\Box_1$ and $\Box_2$ as operators for belief rather than knowledge, Binkley shows:

**Proposition 15** (Binkley 1968)**.** $\{P_1, P_2', P_3\} \vdash_{\mathbf{KD4}^{\leq}} \bot$.

Corollary 1 is a stronger inconsistency result than Proposition 15 in the following sense:

**Fact 4** (Comparison to Binkley)**.**

1. $\mathbf{KJ} \subsetneq \mathbf{KD4}^{\leq}$;
2. $\{P_1, P_2'\} \vdash_{\mathbf{K}} (A)$, $\{P_1, P_3\} \vdash_{\mathbf{K}} (B)$, and $\{P_1\} \vdash_{\mathbf{K4}^{<}} (C)$;
3. $\{(A),(B)\} \nvdash_{\mathbf{KD4R}} P_1 \vee P_2'$ and $\{(A),(B),(C)\} \nvdash_{\mathbf{KD4}^{<}\mathbf{R}} P_1 \vee P_2'$.

Sorensen [1988] replaces $P_1$ and $P_3$ with

$P_1'$ $\Box_1(p_1 \vee p_2)$ and

$P_3'$ $\Box_1(p_2 \to \Box_2 \neg p_1)$,

which is a weakening of the previous authors' in the following sense:

**Fact 5.** $\vdash_{\mathbf{K}} (P_1 \wedge P_3) \to (P_1' \wedge P_3')$, $\vdash_{\mathbf{K}} (P_1' \wedge P_3') \to P_3$, and $\vdash_{\mathbf{KT}} (P_1' \wedge P_3') \to P_1$, but $\nvdash_{\mathbf{KD}} (P_1' \wedge P_3') \to P_1$.

Modifying Harrison's result, Sorensen proves the following:

**Proposition 16** (Sorensen 1988)**.** $\{P_1', P_2', P_3'\} \vdash_{\mathbf{KT4R}} \bot$.

Corollary 1 is a stronger inconsistency result than Proposition 16 in the following sense:

**Fact 6** (Comparison to Sorensen)**.**

1. $\mathbf{KJ} \subsetneq \mathbf{KT4R}$;
2. $\{P_1', P_2'\} \vdash_{\mathbf{K}} (A)$, $P_3'$ is $(B)$, and $\{P_1'\} \vdash_{\mathbf{K4}^{<}} (C)$;
3. $\{(A),(B)\} \nvdash_{\mathbf{KD4R}} P_2'$ and $\{(A),(B),(C)\} \nvdash_{\mathbf{KD4}^{<}\mathbf{R}} P_2'$.

For the $n = 2$ case of the designated student paradox, Sorensen replaces all of the premises with the following and proves the result in Proposition 17:

$P_1^\star$ $\Box_1\Box_2(p_1 \vee p_2)$;

$P_2^\star$ $\Box_1\Box_2((p_1 \to \neg\Box_1 p_1) \wedge (p_2 \to \neg\Box_2 p_2))$;

$P_3^\star$ $\Box_1\Box_2(p_2 \to \Box_2 \neg p_1)$.

**Proposition 17** (Sorensen 1988)**.** $\{P_1^\star, P_2^\star, P_3^\star\} \vdash_{\mathbf{KT}} \bot$.

Corollary 1 is a stronger inconsistency result than Proposition 17 in the following sense:

**Fact 7** (Comparison to Sorensen)**.**

1. $\mathbf{KJ} \subsetneq \mathbf{KT}$;
2. $\{P_1^\star, P_2^\star\} \vdash_{\mathbf{KT}} (A)$, $\{P_3^\star\} \vdash_{\mathbf{KT}} (B)$, and $P_1^\star$ is $(C)$;
3. $\{(A), (B)\} \nvdash_{\mathbf{KD4R}} P_2^\star$ and $\{(A), (B), (C)\} \nvdash_{\mathbf{KD4}^<\mathbf{R}} P_2^\star$.