Utrecht University

The Surprise Examination Paradox Examined

Author:
Niels Rustenburg

Student number:
3539792

First Supervisor
Albert Visser

Second Supervisor
Gerard Vreeswijk

A thesis of 7.5 EC submitted in partial fulfillment for the degree of Bachelor of Science

September 3, 2014

CONTENTS

1. INTRODUCTION

An important topic in the field of Artificial Intelligence is the design and utilization of so called intelligent agents which make use of (formal) logic to interact with their environment. One feature which is very important to these intelligent agents, is the ability to reason about knowledge of other agents, and perhaps even their own future knowledge. Seeing as these agents use logic to represent everything they know and do, the strengths and shortcomings of the logical system they employ will have a large impact on such an agent's capacities. Because of this link between the two, the advancements in the field of Logic are of utmost importance for the field of Intelligent Agents to progress. Something which has led to large advancements in the field of Epistemic Logic is the analysis of epistemic paradoxes. These paradoxes test the limits of our logical thinking and force us to adjust. One such paradox is the Surprise Examination paradox, the treatment of which has already changed and will hopefully continue changing the way we think about knowledge. While the most known variation is called the Surprise Examination, the paradox goes by many different names, among which the Hangman and the Designated Student paradox, which I will come back to later on in this thesis.

For a clear view on what the Surprise Examination is, I will give the formulation thereof as presented by Holliday:

> *A teacher announces to her student that she will give him a surprise exam during a term of $n \geq 2$ days. The student, a perfect logician, reasons as follows: "Since (I know) the exam will be a surprise, (I know) 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 occured, will know that the exam has to be later on day n—so it will not be a surprise. Moreover, (I know) 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, 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 occured, 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 occurs on, say, day* $n - 1$. (Holliday, 2014b)

The aim of this thesis is to give a clear presentation of the approaches made by Kaplan et al. (1960) and Holliday (2014b), and then make a critical comparison between the two, assessing their differences and attempting to explain where these differences come from. In doing so I do not necessarily aim to decide which analysis is better, but rather to determine whether these approaches are capable of complementing each other, or whether they are incompatible.

I will start by giving a framework of the logic through which the two analyses will be presented. Followed by the presentations of these analyses and I will end with a critical comparison and some concluding remarks on the matter.

2. FRAMEWORK MODAL LOGIC

In order to analyse the paradox in the coming sections, we first need to formalize the paradox, and so a decision has to be made on how that will be done. I have decided to use the same propositional modal language as used in Holliday (2014b). This language contains a sentential operator $\Box_i$ and an atomic sentence p_i for each $i \in \mathbb{N}$. For the surprise exam paradox we shall use the following reading:

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

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

Now we can write $p_i \wedge \neg\Box_i p_i$ to express that there is a surprise exam on day i. The proof system we use is 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)

$$\mathsf{RK}_i \ \frac{(\varphi_1 \wedge ... \wedge \varphi_m) \rightarrow \psi}{(\Box_i\varphi_1 \wedge ... \wedge \Box_i\varphi_m) \rightarrow \Box_i\psi}$$

In the $m = 0$ case RK_i is the standard rule of Necessitation Nec_i. As Holliday points out, there are multiple philosophical objections possible to the use of RK_i even for ideally rational agents. However, he claims that his usage of the RK_i rule illuminates more than it distorts the analysis of the surprise exam paradox. And I would say the same could be said for the instances of RK_i which I will use in the analysis of Kaplan and Montague.

Further on in this thesis, I will refer to axiom schemas which extend **K**. Which axioms they are will become clear when they are needed. As to what it means for an axiom schema to extend **K**, I will adhere to the following definition, given by Holliday:

> Given schemas $\Sigma_1, ..., \Sigma_n$, $\mathbf{K}\Sigma_1...\Sigma_n$ is the smallest extension of **K** that includes all instances of $\Sigma_1, ..., \Sigma_n$. A sentence β is *provable* in $\mathbf{K}\Sigma_1...\Sigma_n$ from a set of sentences (premises) Γ, written '$\Gamma \vdash_{\mathbf{K}\Sigma_1...\Sigma_n} \beta$', iff there is a sequence $\langle\chi_1, ..., \chi_l\rangle$ of sentences with $\beta = \chi_l$ such that for all $1 \leq k \leq l$, one of the following holds:
>
> (i) χ_k is an instance of a propositional tautology;

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

(iii) χ_k is one of the sentences in Γ

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

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

(Holliday, 2014b)

For background reading on epistemic logic see (Pacuit, 2013a), (Pacuit, 2013b) and (Holliday, 2014a).

3. EXPOSITION OF THE ANALYSIS OF KAPLAN & MONTAGUE

One influential analysis of the paradox, which, coincidentally also led to the discovery of a new paradox is that of Kaplan and Montague. What Kaplan and Montague aimed to do in their analysis of the paradox, is to try and formulate it in such a way that it is genuinely paradoxical rather than trying to find a solution to the paradox, which is what most analyses do. Their analysis is built up from a simple version of the paradox due to Quine (1953), to which they then add more elements until they believed they had a genuine paradox. The first addition to Quine's formulation was made by Shaw (1958), both he and Kaplan and Montague were convinced that the paradox was caused by a self-referential element in the teacher's utterance, which was not accounted for in Quine's version of the paradox. However, Kaplan and Montague believed Shaw's version was no longer a paradox, but merely an announcement incapable of fulfillment, and so they came up with their own version of the paradox. In their paper Kaplan and Montague talk of the Hangman paradox, rather than the Surprise Exam paradox. But this should not matter, since these paradoxes are essentially the same, but instead of a teacher announcing to her student that he will be given a surprise exam in the next n days; a judge announces to a prisoner that he will be executed within the next n days, and that it will be a surprise. Because there are no essential differences between the two, I will, for the sake of consistency and ease of comparison, convert their formulation of the Hangman to a formulation of the Surprise Examination. Another difference is the way in which Kaplan and Montague formalize the paradox. However because Holliday's way of formalizing it is much clearer, I have decided to convert Kaplan and Montague's proofs to their counterparts formalized in the language mentioned in the previous chapter. What is also important to note is that Kaplan and Montague argue that their approach works on an arbitrary n, and thus to keep the proof short, they give an $n = 2$ version of the paradox.

Below is the $n = 2$ version of the Surprise Examination as Kaplan and Montague would formulate it:

> A teacher announces to her student that: Unless he knows on morning of day one that the current announcement is false; one of the following conditions will be fulfilled:
>
> (1) He will be given a test on day one (and not on day two) and on the morning of day one he will not know on the basis of the current announcement that he will be given a test on day one.
>
> (2) He will be given a test on day two (and not on day one) and on the morning of day two he will not know on the basis of the current announcement that he will be given a test on day two.

While it may seem a bit redundant to mention that when the test is on day one it cannot be on day two, it is essential for the proof, as Kaplan and Montague do not capture this necessary feature of the paradox in any of their premises.

Below at (1) is a formalization of the teacher's announcement, which is referred to as D_4, can be found. [1]

$$(1)\ \vdash D_4 \equiv \quad \Box_1\neg D_4 \vee (p_1 \wedge \neg p_2 \wedge \neg\Box_1(D_4 \rightarrow p_1)) \vee (\neg p_1 \wedge p_2 \wedge \neg\Box_2(D_4 \rightarrow p_2))$$

To prove the paradoxicality, some assumptions on the knowledge of the student must be made. Kaplan and Montague have eight of these assumptions, but when converting their assumptions from their notation to the notation used in this thesis I found that 6 of their assumptions are merely instantiations of the rule RK_i and since we are using $\mathbf{K}$ for this proof I will leave these assumptions out. After trimming down the list of assumptions, what we are left with are rules F_1 - F_2.[2]

$(\mathrm{F}_1)\ \Box_1\neg D_4 \rightarrow \neg D_4$

[1] Kaplan and Montague work in a setting where the modalities are treated as predicates of sentences. In this setting, they prove the existence of the self-referential formula D_4 using the Gödel Fixed Point Lemma. To understand the reasoning connected to the Surprise Exam Paradox, however, only the existence of the fixed point is relevant, plus, as we shall see modal reasoning. Hence, in the present context, the existence of the fixed point is simply stipulated.

[2] These correspond with assumptions C_1 and C_2 in (Kaplan et al., 1960)

(F_2) $\neg p_1 \rightarrow \Box_2 \neg p_1$

Note that F_1 is an instance of the $\mathbf{T}_i$ axiom: $\Box_i \varphi \rightarrow \varphi$, which means that if we were to use for example $\mathbf{KT}_1$ as our proof system we would only need premise F_2 to derive a paradox.

(2)	$D_4 \rightarrow \neg\Box_1 \neg D_4$	from F_1 using PL
(3)	$\neg p_1 \rightarrow (D_4 \rightarrow p_2)$	from (1) and (2) using PL
(4)	$(D_4 \wedge p_2) \rightarrow \neg\Box_2(D_4 \rightarrow p_2)$	from (1) and (2) using PL
(5)	$(D_4 \wedge p_2) \rightarrow \neg p_1$	from (1) and (2) using PL
(6)	$(D_4 \wedge p_2) \rightarrow \Box_2 \neg p_1$	from F_2 and (5) using PL
(7)	$(D_4 \wedge p_2) \rightarrow \Box_2(D_4 \rightarrow p_2)$	from (3) and (6) using RK_2
(8)	$D_4 \rightarrow \neg p_2$	from (4) and (7) using PL
(9)	$(D_4 \wedge \neg p_2) \rightarrow p_1$	from (1) and (2) using PL
(10)	$(D_4 \wedge p_1) \rightarrow \neg\Box_1(D_4 \rightarrow p_1)$	from (1) and (2) using PL
(11)	$D_4 \rightarrow p_1$	from (8) and (9) using PL
(12)	$\Box_1(D_4 \rightarrow p_1)$	from (11) using Nec_1
(13)	$D_4 \rightarrow \neg p_1$	from (10) and (12) using PL
(14)	$D_4 \rightarrow (\neg\Box_1 \neg D_4 \wedge \neg p_1 \wedge \neg p_2)$	from (2), (8) and (13) using PL
(15)	$\neg D_4$	from (1) and (14) using PL
(16)	$\Box_1 \neg D_4$	from (15) using Nec_1
(17)	$\Box_1 \neg D_4 \rightarrow D_4$	from (1) using PL
(18)	D_4	from (16) and (17) using PL

As shown above, under the assumptions F_1 and F_2, the student could show (and know) that the announcement is uncapable of fulfillment yet at the same time, this would make it possible for the teacher to fulfill her announcement by giving an unexpected exam, and thus we have a paradox.

The interesting thing, is that this proof of Kaplan and Montague works for $n = 1$ as well and even for $n = 0$.[3] While I will not present these proofs, as they are simply shorter versions of the above proof, I will in section 5, go into how Kaplan and Montague come to these slightly counter-intuitive results.

[3] In the $n = 0$ case what they are left with, is what Kaplan and Montague call the Knower paradox

4. EXPOSITION OF THE ANALYSIS FROM HOLLIDAY

In this section I will summarize the approach made in Holliday (2014b). Because Holliday makes a lot of use of the Designated Student paradox (a variation on the Surprise Exam introduced by Sorensen (1982)), I will introduce it here and refer to this version of the paradox for most of the chapter. Why Holliday gives preference to analysing the Surprise Exam, through this paradox will be explained below, after I have presented it.

The Designated Student, as formulated by Holliday:

> *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. 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 he has it. Student 1, at the front of the line, reasons as follows: "Since (I know) the gold star will be a surprise, (I know) the teacher did not give the star to student n; because if she did, then student n, seeing all the silver stars in front of him will know he has the gold star—so it will not be a surprise. Moreover (I know) the teacher did not give the gold star to student $n-1$; because if she did, then student $n-1$, knowing the gold star will be a surprise, will 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, will know he has the gold star—so it will 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 is surprised.* (Holliday, 2014b)

The main argument in favor of using this variation on the paradox is that, instead of involving one student reasoning about his knowledge and his future knowledge, it involves multiple students reasoning about each other's knowledge at the current time. Because this removes the temporal aspect of the Surprise Exam, a lot of analyses that do not get to the heart of the matter will not hold for the Designated Student. However, the sameness

of paradox is a delicate matter here.

When analysing the Designated Student, a different interpretation of the $\Box_i$'s and p_i's is in order. Therefore, within the analysis of the Designated Student $\Box_i$ and p_i should be read as follows:

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

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

In his analysis Holliday makes two distinctions. The first being a distinction between the $n = 2$ and $n > 2$ cases of the paradox. The second is a distinction between what he calls the "Promised Event" and the "Inevitable Event", however, I will only look at the Inevitable Event for my thesis, and I will explain why later on in this section.

I will begin with Holliday's analysis of the $n = 2$ cases of the paradox. Holliday's analysis rests on the use of 3 assumptions about the knowledge of the first student in line, student 1.[4] The assumptions for $n = 2$ would be the following: (A) The first student in line, student 1, knows that the gold star is a surprise; (B) 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 basis of his seeing a silver star on student 1's back); (C) student 1 knows that student 2 knows that one of them has the gold star.

What Holliday claims is the strength of his proof, is that with these common assumptions and a weaker proof system[5] (**K**) he is able to provide a simpler proof than the ones found in other analyses.

What Holliday does in his analysis is, from his assumptions, he derives a sentence containing Moorean knowledge of the form "I have the gold star but I don't know it". In reaction to deriving this intuitively inadmissable sentence he decides that we should be able to pinpoint which of the assumptions (or perhaps his proof system) is to blame for creating this paradox, and we should then reject this assumption(or proof system).

[4] Or in the case of the Surprise Examination, they would be assumptions on the knowledge of the student on the morning of day 1.

[5] Relative to that used in most other analyses.

I will now present his proof, after which I will show his considerations in deciding which of the above should be rejected.

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

(1) $\Box_2(p_1 \vee p_2) \wedge \Box_2\neg p_1) \to \Box_2 p_2$ by PL and RK_2
(2) $\Box_1((\Box_2(p_1 \vee p_2) \wedge \Box_2\neg p_1) \to \Box_2 p_2)$ from (1) by Nec_1
(3) $\Box_1(\Box_2\neg p_1 \to \Box_2 p_2)$ from (C) and (2) using PL and RK_1
(4) $\Box_1\neg(p_2 \wedge \neg\Box_2 p_2)$ from (B) and (3) using PL and RK_1
(5) $\Box_1(p_1 \wedge \neg\Box_1 p_1)$ from (A) and (4) using PL and RK_1

Holliday notes that while from (5) we cannot derive a contradiction yet, that with the addition of an extra axiom such as $\mathbf{J}_i$: $\Box_i\neg\Box_i\varphi \to \neg\Box_i\varphi$ we would be able to. But he suggests that the fact that we want to extend $\mathbf{K}$ with an extra axiom to disallow the sentence represented by (5) reflects that (5) is already paradoxical by itself.

As mentioned before, to respond to the derivation of (5) we must reject one of the premises (A), (B). (C), or we must reject the rule RK_i. Holliday notes that it is likely that for any of the premises you could find a way of filling in the Surprise Exam so that the premise should be rejected. However some ways of filling in the Surprise Exam are more natural than others. It seems rather natural to assume that the student in the Surprise Exam scenario has good memory, and that student 1 in the Designated Student scenario knows that the students behind him all have good eyesight. With these assumptions Holliday claims that allowing (B), which reflects the above assumptions, should be unproblematic, so the blame should be laid on one of the other principles.

According to Holliday, the question as to whether we should reject (A), (C) or RK_i has differing answers for the $n = 2$ case, depending on whether we find ourselves in an instance of the Promised Event, or the Inevitable Event. Below is an explanation of what Holliday means when he talks of the Inevitable Event and the Promised Event, and I will also explain why I will only consider the Inevitable Event for the purposes of this thesis.

Holliday explains the Inevitable Event as being essentially the same as the scenario given at the beginning of this chapter, however to be extra safe, he adds a few details. For example that the teacher will display all the stars, clear for every student to see, and that all students see that the other students have seen the stars, and that they communicate this with each other,

etc. And that when they are lined up and the stars are put on their backs, they reach around to feel that there is indeed a star on their back, and again communicate this to each other.

In Holliday's Promised Event, the teacher never shows the students any stars, there is only the teacher's claim that she will put a gold star on the back of one student, and silver stars on the backs of the remaining students. And thus the students have a lot less knowledge available to them in this Promised Event, as their only information comes not from their own perception, but from the teacher's announcement.

While the distinction between the two is quite apparent, I feel that the intriguing element of the paradox does not come from having to doubt the teacher's intentions, or the possibility that the teacher is some sort of magician or wizard. Because of this conviction I will disregard the Promised Event and focus on the Inevitable Event in this thesis.

Continuing with Holliday's search for a faulty assumption in the $n = 2$ case, having already acquitted (B) we must look at (A), (C) and RK_i. First Holliday examines the consequences of allowing (C). In the Surprise Exam (C) might cause some problems. In the Surprise Exam (C) is to be interpreted as meaning that the student knows on the morning of day one, that he will know on the morning of day two, the part of the teacher's announcement of there being an exam. The problem this would cause, is that some philosophers might deny the possibility of any future knowledge, let alone knowledge of future knowledge, and reject (C) on the basis of that. However Holliday argues that in the Designated Student paradox, (C) does not involve future knowledge, but rather, knowledge of another person's knowledge. To deny that student one knows that student two knows that one of them has the gold star, he claims, would be radical skepticism about social knowledge and not a solution to the paradox. And so Holliday decides that (C) is free from blame.

This leaves us with (A) and RK_i. If RK_i were to prove problematic Holliday suggests that it would have to reveal itself as such in one of the steps (1) - (5). To examine RK_i's involvement in the creation of (5) Holliday uses the Designated Student to interpret each step, and then, for each step argues whether the use of RK_i is problematic. First off is step (1), a step that reflects the idea that student 2, a perfect logician, will not at the same time believe $p_1 \vee p_2$ and $\neg p_1$ and refuse to believe p_2; this does not seem like a misuse of RK_i. Step (2) reflects that student 1 knows that student 2 is a perfect logician capable of the abovementioned reasoning, which again should not be problematic according to Holliday. Holliday notes that step (3) might raise an objection if the justification for (C) were to rely on some

justification for believing $\Box_2 p_1$; this because student 1 could then 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* $\Box_2\neg p_1 \rightarrow \Box_2 p_2$. Holliday (2014b)

However, Holliday points out that student 1's justification for believing that student 2 knows that one of them has the gold star, relies on his justification for believing that student 2 saw both the gold star and the silver star, and saw and heard the teacher walk behind each of them and stick the stars on their backs etc., which isn't incompatible with the supposition for the conditional proof that $\Box_2\neg p_1$. He gives two more arguments suggesting that even if student 1 believing $\Box_2(p_1 \vee p_2)$ rested on $\Box_2 p_1$ that it should not be an issue, but I will not repeat them here as I believe the usage of RK_i in step (3) has been defended well enough by the above argument. According to Holliday steps (4) and (5) should not be problematic either as, (4) merely reflects student 1 using 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 $p_2 \rightarrow \Box_2 p_2$ by transitivity of implication. And to get to step (5) student 1 uses his (supposed) knowledge of (A) and the knowledge he gained from (4) (which is the negation of the right disjunct of (A)) to deduce and come to know the left disjunct by disjunctive syllogism, again no foul play by RK_i to be found.

Holliday concludes that RK_i is not to blame and concludes that we must reject (A). As a consequence it is not possible for student 1 to know the teacher's announcement, given that he has the knowledge represented in (B) and (C), and given that he's clever enough to engage in the backward elimination argument. The reason being that this 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 a gold star".

Now that we've seen Holliday's solution to the $n = 2$ case I will continue by presenting his solution to the $n > 2$ case. While there is a distinction between the two, Holliday will also show that the solution to the $n = 2$ case will lead to a solution for $n > 2$. To analyse the $n > 2$ case Holliday first needs to enhance his premises(A), (B) and (C) so that they work for an arbitrary n. The generalizations of the premises are below, where $S_k := (p_k \wedge \neg\Box_k p_k)$.

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

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

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

While with $n = 2$ these assumptions would lead to a contradiction in $\mathbf{KJ}_1$, with $n > 2$, Holliday shows in (Holliday, 2014b) that these premises are consistent even in a much stronger proof system such as **S5**.

Seeing as Holliday cannot arrive at a paradox or contradiction with the current premises and proof system, there must be something missing, as the Designated Student is clearly still paradoxical with $n > 2$. Presumably it is because of this that Holliday decides to extend his proof system **K** with another axiom. The axiom he adds is the $\mathbf{4}_1^<$ axiom, which represents that student 1 knows that whatever is known to him is also known to any student in line behind him.[6] This seems like a fair assumption, and is not unlikely to be something student 1 uses in his reasoning. With this new proof system of $\mathbf{K4}_1^<$, like in his proof for $n = 2$, Holliday will be able to obtain a Moorean sentence of the form "I have the gold star but I don't know it". Because of $\mathbf{4}_1^<$, Holliday no longer needs (C^n) to find this Moorean sentence, as from (A^n) with $\mathbf{4}_1^<$, (E^n) can be derived. (E^n) is essentially a stronger variant of (C^n) which states that student 1 knows that the students behind him know the teacher's announcement of there being a surprise gold star, rather than them knowing only that there is a gold star.

Holliday's proof is as follows:

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

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

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

Now we show that student 1 can rule out a gold star on the back of the last student n, and if he has ruled out student $k+1$ and all students further back in line, then he can rule out student k and all further students (for $k \geq 2$)

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

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

(k, 1) $(\Box_k(S_1 \vee ... \vee S_n) \wedge \Box_k\neg(p_1 \vee ... \vee p_{k-1}) \wedge \Box_k\neg(S_{k+1} \vee ... \vee S_n)) \to \Box_k p_k$ by PL and RK_k

(k, 2) $\Box_1[(\Box_k(S_1 \vee ... \vee S_n) \wedge \Box_k\neg(p_1 \vee ... \vee p_{k-1}) \wedge \Box_k\neg(S_{k+1} \vee ... \vee S_n)) \to \Box_k p_k]$ from (k, 1) by Nec_1

(k, 3) $\Box_1(\Box_k\neg(p_1 \vee ... \vee p_{k-1}) \to \Box_k p_k)$ from (E^n), (k, 0) and (k, 2) using RK_1 and PL

[6] Or in the Surprise Exam it would suggest that the student knows that everything he knows on day 1 he will continue to know the subsequent days.

$(k, 4)$ $\Box_1\neg(p_k \wedge \neg\Box_k p_k)$ from (B^n) and $(k, 3)$ using RK_1 and PL

$(k, 5)$ $\Box_1\neg(S_k \vee ... \vee S_n)$ from $(k+1, 5)$ and $(k, 4)$ using RK_1 and PL

Repeating the above reasoning we eventually obtain:

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

$(2, 6)$ $\Box_1(p_1 \wedge \neg\Box_1 p_1)$ from (A^n) and $(2, 5)$ using RK_1 and PL

After deriving the paradoxical $\Box_1(p_1 \wedge \neg\Box_1 p_1)$ again, Holliday once more attempts to find a solution through rejecting either one of the premises or (part of) his proof system. His solution to the $n > 2$ case is closely tied to his solution to the $n = 2$ case. Holliday illustrates his solution again through an example of the Designed Student paradox. Imagine the following: if student $n-1$ were to see only silver stars in front of him, a possibility which we will call (*), then he is in what is essentially the same epistemic position as that of student 1 in the $n = 2$ case. And as has been proven earlier, the first student in de $n = 2$ case cannot know the teacher's announcement, and so neither can student $n-1$ in (*). Since student 1 in $n > 2$ does not initially know whether he finds himself in (*) (before he starts his reasoning about student $n-1$) he cannot know that student $n-1$ knows the teacher's announcement, and so we must reject (E^n). This has two important consequences. First off it is that if student 1 cannot know whether student $n-1$ knows the teachers announcement then he cannot eliminate student $n-1$ and thus his backward elimination argument is blocked. Secondly, from what we have seen it is possible for student 1 to know the teacher's announcement (so unlike in $n = 2$ rejecting (A^n) is not an option here). So we must reject $4^<$ which together with (A^n) is responsible for creating the inadmissable (E^n). While in previous analyses, within the scenario of the Surprise Exam, the rejection of $4_1^<$ might have been dismissed as worries about "temporal retention", from Holliday's analysis it seems clear that it cannot be allowed in the Designated Student scenario, which is free of temporal aspects.

5. CRITICAL COMPARISON OF THE ANALYSES

There are two main differences we stumble upon when presented with these analyses. The first being that Kaplan and Montague make use of a self referential formulation of the paradox, whereas Holliday seems to disregard this aspect in his solution to the paradox. The second large difference is their opposing views on whether the case of $n = 2$ is different from $n > 2$. Holliday argues there is indeed a distinction to be made, whereas Kaplan and Montague claim that all relevant features are preserved when going from $n = 3$ to $n = 2$, they even go as far as to say that they can still find a paradox with $n = 0$.

5.1 Self-Referentiality

Kaplan and Montague feel that the self-referential element is an integral part of the paradox. At first glance I would say that they are right, because the intuitive reason that we perceive the Surprise Exam as a paradox seems to be that the student's susceptibility to being surprised, comes from his ruling out the possibility there being such an exam in the first place. However, I have my doubts as to whether the kind of self-referentiality as displayed in the analysis of Kaplan and Montague is the same kind of self-referentiality which lies at the heart of the Surprise Exam paradox. My doubts stem from the fact that the formulation of Kaplan and Montague is capable of finding a paradox for $n = 0$. This $n = 0$ case is an instance of another paradox which they call the Knower paradox and it creates an announcement in the form of "This sentence is known to be false". Intuitively in an $n = 0$ case I'd argue that the teacher announces nothing, and surely it should not be paradoxical for a teacher to announce nothing to her students.[7]

In his solution to the paradox, Holliday does not need any self-referentiality. The question is whether this means that his analysis inaccurate, or whether self-referentiality simply doesn't play a central role in the creation of the paradox. Fortunately, in his paper, Holliday does explain why he does not include the self-referentiality of the paradox in his solution. He suggests that the bad reasoning exposed in section 4, is what lies at the heart of

[7] albeit of course less than optimal if the students are supposed to learn something from their teacher

the problem, and not the self-referential elements. While he acknowledges that the student engages in self-undermining reasoning when ruling out the possibility of surprise exam (and thus making himself susceptible to one), Holliday suggests that if the student did not engage in this bad reasoning in the first place, that he would not have fallen into this self-referential trap. Another reason why Holliday thinks there must be more to it than just an issue of self-referentiality is that, while in the Surprise Exam it seems to play an important role, it his hard to make a similar case for the Designated Student. As suggesting that student 1's reasoning for eliminating the possibility of a gold star on anyone's back, under the assumption that student $n-1$ knows the teacher's announcement, would in turn lead to student $n-1$'s to not believe the teacher's announcement after all seems odd.

No doubt there is more to be said about the role of self-referentiality in the paradox, but I conclude this matter with the observations that Kaplan and Montague, giving the Knower, which is quite a strong paradox in its own right, a central role in the explanation of the Surprise Exam, is at the very least questionable, whereas Holliday has a few strong arguments for disregarding the self-referentiality when solving and analyzing the paradox.

5.2 $n = 2$ versus $n > 2$

As I have mentioned previously, the two analyses have differing opinions on whether the $n = 2$ and $n > 2$ cases are significantly different. I will now present a $n > 2$ formalization to a formulation by Shaw (1958), on which Kaplan and Montague based their formulation, because Kaplan and Montague did not provide a proof for the claim that:

> *All relevant features of D_2[8] are preserved if only two dates of execution*[9] *are considered.* (Kaplan et al., 1960)

The reason I use Shaw's formulation is purely to shorten the proof a little, extending this to a proof for Kaplan and Montague's formulation should be simple.[10] I hope that by showing that Shaw's proof holds for an arbitrary n it becomes clear to see that Kaplan and Montague's proof would also hold for an arbitrary n.

[8] In their paper, Kaplan and Montague refer to the announcements as instances of D_i where each incrementation of i they've either added something to the announcement or taken something away

[9] As mentioned earlier, in their paper Kaplan and Montague discuss the Hangman paradox, which is in practically every way identical to the Surprise Examination, only now regarding a judge, a prisoner and a (surprise) execution

[10] Shaw's formulation is basically identical to that of Kaplan and Montague, only Kaplan and Montague added a segment to the announcement which states that there does not have to be an exam if the student knows on day one that the announcement is false.

The proof would be as follows:

$(\mathrm{D}^n)\ \vdash D^n \equiv \quad (p_1 \wedge \neg p_2 \wedge ... \wedge \neg p_n \wedge \neg\Box_1(D^n \to p_1)) \vee ... \vee (\neg p_1 \wedge ... \wedge \neg p_{n-1} \wedge p_n \wedge \neg\Box_n(D^n \to p_n))$

$(\mathrm{G}^n)\ \bigwedge_{1<k\leq n} (\neg p_1 \wedge ... \wedge \neg p_{k-1}) \to \Box_k(\neg p_1 \wedge ... \wedge \neg p_{k-1})$

Now we show that on the morning of day 1 the student can rule out an exam on day n, and if he has ruled out an exam on day $k+1$ and all days after that, then he can rule out an exam on day k and all earlier days.

$(k+1, 8)$ $D^n \to (\neg p_{k+1} \wedge ... \wedge \neg p_n)$ (if $k = n$ let $\neg p_{k+1} \wedge ... \wedge \neg p_n := \top$)

$(k, 1)$ $D^n \to (p_1 \vee ... \vee p_k)$ from $(k+1, 8)$ and (D^n) using PL

$(k, 2)$ $(\neg p_1 \wedge ... \wedge \neg p_{k-1}) \to (D^n \to p_k)$ from $(k, 1)$ using PL

$(k, 3)$ $(D^n \wedge p_k) \to \neg\Box_k(D^n \to p_k)$ from (D^n) using PL

$(k, 4)$ $(D^n \wedge p_k) \to (\neg p_1 \wedge ... \wedge \neg p_{k-1})$ from (D^n) using PL

$(k, 5)$ $(D^n \wedge p_k) \to \Box_k(\neg p_1 \wedge ... \wedge \neg p_{k-1})$ from (G^n) and $(k, 4)$ using PL

$(k, 6)$ $(D^n \wedge p_k) \to \Box_k(D^n \to p_k)$ from $(k, 2)$ and $(k, 5)$ using RK_k

$(k, 7)$ $D^n \to \neg p_k$ from $(k, 3)$ and $(k, 6)$ using PL

$(k, 8)$ $D^n \to (\neg p_k \wedge ... \wedge \neg p_n)$ from $(k+1, 8)$ and $(k, 7)$ using PL

Repeating the above reasoning we eventually obtain:

$(1, 8)$ $D^n \to (\neg p_1 \wedge ... \wedge \neg p_n)$

$(1, 9)$ $D^n \to \neg D^n$ from (D^n) and $(1, 8)$ using PL

This should make it clear that for Kaplan and Montague's formulation there is no significant difference between a $n = 2$ proof compared to a $n > 2$ proof, as unlike with Holliday we need no added axiom's to make the desired derivations when going from $n = 2$ to $n > 2$.

These varying results on whether or not there is a significant difference between $n = 2$ and $n > 2$ stem from a difference in assumptions. Comparing Holliday's assumptions to their Kaplan and Montague counterparts, I came up with the following observations. Premise (A) is to Holliday's proof what the teacher's announcement is to Kaplan and Montague's proof.[11] However, while in the proof of Kaplan and Montague the disjunction of the teacher's announcement is easily accessible, in Holliday's proof

[11] I count the teacher's announcement as an assumption when I speak of Kaplan and Montague's proof, as they need to assume the announcement is true, to derive the Knower paradox

the teacher's announcement is locked behind a $\Box_1$. As a result, Holliday derives a Moorean sentence such as $\Box_1(p_1 \wedge \neg\Box_1 p_1)$ (but no contradiction), whereas Kaplan and Montague can easily find $p_1 \wedge \neg\Box_1 p_1$, which with Necessitation easily becomes $\Box_1 p_1 \wedge \neg\Box_1 p_1$. Premise (B) is Holliday's variant of the assumption that the student has good memory, and plays the same role as premise (F_2) in the proof of Kaplan and Montague. While these premises are not equivalent, they have the same impact on the proof, when in combination with the other premises. The difference being that, on its own (F_2) does not exclude the possibility of multiple exams, whereas (B) does. However, Kaplan and Montague use multiple conjunctions of $\neg p_i$'s in the teacher's announcement to capture that there can be only one exam, and so the impact of (F_2) and (B) are the same. As I've mentioned in section 3 (F_1) is simply an instance of the $\mathbf{T}_i$ axiom: $\Box_i\varphi \to \varphi$, Holliday has no equivalent in his proof, which does not use $\mathbf{T}_i$. This is of no importance though, since, as you can see from the proof earlier this section, if we remove the Knower from the teacher's announcement, Kaplan and Montague's proof no longer needs (F_1) either. Holliday makes use of assumption (C) to make a derivation for $n = 2$ and $\mathbf{4}_1^<$ for $n > 2$, while Kaplan and Montague can do without. What seems to be only possible explanation for this is that the assumption of the teacher's announcement in the proof of Kaplan and Montague, without the $\Box_1$, is strong enough to eclipse the need for (C) and $\mathbf{4}_1^<$. What we can draw from this is that, because Kaplan and Montague's premises are much stronger, they have a much easier time to come to a contradiction (in Shaw's variant) or instance of the Knower paradox, which is the reason why they see no difference between the $n = 2$ case and the $n > 2$ cases. Holliday with his use of weaker premises on the other hand, allows us to take a closer look at where the boundaries of this paradox lie, one of which is apparantly $n = 2$.

6. CONCLUSION

From what we have seen in the above comparisons, the main differences are the lack of self-referentiality in Holliday's analysis, and the difference in the strength of the assumptions, which leads to a different answer to the question as to whether there is a difference between $n = 2$ and $n > 2$. I believe that on the basis of this, combined with the observation that Kaplan and Montague have added an instance of the Knower paradox onto Shaw's analysis of the Surprise Exam, that we can conclude that these analyses are definitely not incapable of complementing eachother. For it should not be difficult to add the self-referential element of the Knower, from Kaplan and Montague onto the teacher's announcement in Holliday's analysis, just as Kaplan and Montague have added it onto Shaw's. Whether this would be a meaningful addition however is an entirely different subject, and I would argue that it would not be so. While the Knower is a rather interesting paradox in its own right, I am of the opinion that the paradoxicality expressed in the Knower, is different from the paradoxicality of the Surprise Exam. And so, I would argue that being able to reduce the Surprise Exam to the Knower, as Kaplan and Montague are, would indicate a defective representation of the Surprise Exam. Perhaps, if we find better ways to represent them, elements of self-referentiality can be added on to analyses such as that of Holliday, although maybe, as Holliday claims, it is unnecessary to do so.

BIBLIOGRAPHY

Chellas, B. F. (1980). *Modal logic: an introduction*, volume 316. Cambridge Univ Press.

Holliday, W. (2014a). Epistemic logic and epistemology.

Holliday, W. (2014b). Simplifying the surprise exam.

Kaplan, D., Montague, R., et al. (1960). A paradox regained. *Notre Dame journal of formal logic*, 1(3):79–90.

Pacuit, E. (2013a). Dynamic epistemic logic i: Modeling knowledge and belief.

Pacuit, E. (2013b). Dynamic epistemic logic ii: Logics of information change.

Quine, W. (1953). On a so-called paradox. *Mind*, 62(245):65–67.

Shaw, R. (1958). The paradox of the unexpected examination. *Mind*, pages 382–384.

Sorensen, R. A. (1982). Recalcitrant variations of the prediction paradox. *Australasian Journal of Philosophy*, 60(4):355–362.