The Paradox of the Surprise Examination

Lucian Wischik[†]

[†] The author may be contacted at Queens' College, Cambridge CB3 9ET, or by e-mail to ljw1004@cam.ac.uk. He gratefully acknowledges the contribution of Dr. C Wischik and D Wischik. This work was done as part of an application for an M.Phil. in History and Philosophy of Science.

1. Introduction

1.1. A statement of the paradox

The students set off home for the weekend. Their teacher had been cruel in his sentencing: “You shall have an examination at nine in the morning one day next week; but you will not know in advance which day—it will be a surprise.”

“Suppose,” thought one to herself with the desperation available only to the damned and the examinable: “Suppose we were still unexamined until Friday. I know there will be an examination some time this week, and so I would be able to expect the Friday examination. Therefore the teacher cannot possibly leave it until Friday.”

“Suppose further,” she thought to herself: “Suppose that I am still unexamined on Thursday. I know the examination cannot be on Friday, so it would have to come on Thursday. But then too I would have been able to predict its day. So, the teacher cannot leave it until Thursday either.”

“... or Wednesday... or Tuesday... or Monday... In fact, he cannot possibly set any examination at all, since no day would be a surprise!”

And when the teacher handed out exam papers at nine o’clock on Wednesday morning—wasn’t she surprised!

1.2. The nature of past approaches to the paradox

This paradox has been the object of a steady stream of discussion since O'Conner brought it to public view in 1948 [1]. O'Conner called it a 'Class A Blackout' (class A practices being sprung on unexpecting soldiers), Quine [7] introduced the 'Condemned Man', Shaw [9] called it a 'Surprise Examination', Lyon [10] had a Hand of Cards that were revealed in order with one of them known to be an Ace, and there have been many more presentations. For consistency this paper shall henceforth refer to a surprise examination that has been announced for the next school week. The teacher announces:

A. There will be an examination next week (Monday to Friday);
B. You will not know in advance which day it will be.

There have been as many proposed resolutions to the paradox as there have been authors. The paradox remains fascinating because it has for so long withstood a definitive explanation, and because all aspects of it continue to be debated. O'Conner said [1]: "It is worthwhile for philosophers to pay a little more attention to these puzzles than they have done up to now even if their scrutiny does no more than make a little clearer the ways in which ordinary language can limit and mislead us." The paradox is also relevant to the Philosophy of Knowledge, since it would seem to suggest that either deductive reasoning is powerless to reason about the future, or that propositions regarding states of mind are not meaningful.

The author proposes a new resolution based on a logical notation revised to deal with epistemic concepts—an ‘Epistemological Notation’. This will be presented within a review of all extant literature for two reasons: an account of past developments provides an introduction to the problem; and the author feels that some of the key insights have come from papers that have been largely neglected.

The two most conspicuous features of the literature are that most articles credited with valid points have also been criticised for invalid ones, and that little consensus has emerged. In an attempt

to rectify these problems the new proposal has been justified with an approach that owes more to Computer Science than to Philosophy.

1.3. Key points to the 'Epistemological Notation' solution

In this solution, the position is taken that the students are in a different position from that of an external observer, and that all logical deductions must be predicated upon the person making that deduction. This notation makes explicit the fact that the students are unable authoritatively to reason about the pronouncement, while an external observer is so able. The teacher said: "*you* will not be able to know…" Many previous authors have been unable to resolve the paradox because their notations were unable to express the problem correctly. Key points are as follows:

1. The pronouncement is self-referring, or recursive;
2. The pupils are in a different position from an external observer;
3. Proposition *B* ("you will not be able to predict") is self-fulfilling;
4. The examination may still be a surprise even on the last day; and
5. The conclusions reached by anyone subject to an epistemic proposition are necessarily suspect.

1.4. Overview

This review presents past positions in their own terms and mirrors historical progress within each subject, since to constrain them all to a unified notation would leave some vacuous and others inexpressible.

Section Two covers the nature and intention of the pronouncement, and what it means to be a paradox. Some authors have dismissed the surprise examination as a simple contradiction similar to the Liar Paradox.

Section Three examines how the students should believe in the pronouncement, since it has often been suggested that they have no right to believe what their teacher says. Several authors have had the students selectively ignoring certain pronouncements: this renders the remainder non-contradictory.

Section Four makes general considerations about the reasoning process of the students, since this is necessary before developing a formal logic. It has been claimed that truths change through time, that the students are not in a position to reason sensibly and that modal logic is necessary.

Section Five looks at logical symbolic formalisations in the light of the self-referential nature of the pronouncement. It includes a streamlined version of the students' reasoning and considers several previous symbolic versions.

Section Six presents a new Epistemological Notation which permits reasoning about knowledge, and allows for a solution to the paradox. The notation is developed formally.

2. Contradictions, and types of Pronouncement

This section makes explicit what is implied by the terms 'Contradiction' and 'Paradox'. It then covers possible types of pronouncements: 'pragmatically paradoxical' announcements are rendered contradictory by their publication; and Scriven [4] drew further distinctions between 'ordainments' and 'statements'.

2.1. Contradictions

A contradiction occurs when the 'givens' cannot all be true at once: they are logically incompatible. Equivalently a system might be described 'inconsistent' or as having no 'valuations'. A valuation is the assignment of a value to each atomic proposition such that the pronouncements are made true. For example, the proposition $A \vee B$ has three possible valuations: (A **true**, B **false**), (**false**, **true**) and (**true**, **true**) whereas the proposition $A \wedge \neg A$ has none. In discussion about the Paradox, the propositions are generally taken as e_1 for 'Monday exam', e_2 for 'Tuesday exam' and so on.

Gardner [15] and Popper [16] claimed that the paradox is reducible to the classic Liar problem. This may be expressed a piece of card, one side saying "the reverse of this card is true" and the other, "the reverse is false", which like $A \wedge \neg A$ is contradictory. It is not clear that the reduction is possible and so their greeting cards are best regarded as they were intended: as frivolities.

Cargile [21] made the same claim, and presented a simple example similar to the Paradox:

D: "Either D is false, or *D* is true but K does not know it", or
D= $\neg D \vee (D \wedge \neg K)$

He writes that *D* is just the case of the Disjunct-Liar and that "knowledge is important only in that it entails truth". This would seem incorrect: a proposition may be either true or false; but knowledge has the three modes 'know that P is true', 'know that P is false', 'do not know P'. Additionally, knowledge makes the puzzle self-referential because the premises for knowing the statement is the statement itself—the simple entailment of truth does not have this property.

2.2. Paradox and Contradiction

McClelland and Chihara [37] explained the difference thus:

> "...a *paradox* is a convincing argument with an absurd or contradictory conclusion. It is natural to respond to a paradox by attempting to find gaps or questionable steps in the argument."

Thus, satisfactory resolutions of the paradox would be either a proof that the pronouncements are not possible and an explanation for the commonly held belief that they are; or a proof that they are possible and an explanation for the students' belief they are not.

2.3. Pragmatic Paradoxes

O'Conner [1,5] described a pragmatic paradox as one whose publication renders itself false. "I am unable to construct sentences in English" might be true, but not if communicated in English. He referred to Russell's 'egocentric particulars' [49] and Reichenback's 'token-reflexive words' [50]. However, he used these contradictions only as a description for the logic that the students use and did not consider the paradox generated when the examination is actually set. Cohen [2] distinguished between 'utterances' which contain egocentric particulars and 'propositions' which do not; and Alexander [3] noted that the Paradox seems to be a proposition and hence may not fall into the group

described by O'Conner. In fact, as shown in Section Six, the pronouncement do not render themselves contradictory when published and hence O'Conner's distinction is irrelevant to the Paradox.

2.4. Types of Pronouncement

Scriven [4] distinguished between a 'statement' and an 'ordainment'. In the original formulation, he said, the pronouncement is a statement. As an ordainment it would become:

A. The exam will be this week;
B. The exam will be a surprise to you;
C. It is in my power to ensure that both *A* and *B* are fulfilled;
D. I shall ensure that both *A* and *B* are fulfilled.

He claimed that the ordainment (*A*,*B*,*C*,*D*) is contradictory, but that the statement (*A*,*B*) works. Scriven's distinction is important in that it makes explicit the authority of the teacher and it is cogently argued, but Binkley [30] is the only author to have cited his work. It is hard to imagine that the pronouncement might be interpreted as a statement rather than an ordainment, and for the rest of this paper it is implicitly taken as the latter.

Scriven's and the other approaches above claim, but do not establish, that the ordainment is contradictory. In order to show that this is not the case it is necessary to develop a formal logical approach, which is postponed until Section Six.

Wright [28] came up with an alternative reformulation of the puzzle:

A. There will be an exam this week if the conditions are met.
B. It will be cancelled if the students predicted it for a particular morning.

This version again makes explicit the authority of the teacher to give and cancel tests, but it also allows for the event not to occur at all. The next section discusses the possibilities that the examination might not take place, or that it might not be a surprise.

3. Unreliable Pronouncements

It has been suggested that one or both of the teachers' pronouncements are unreliable, and that the students should ignore them, or consider them to be false. This allows for the possibility (but not the necessity) of a surprise examination. Ignoring one of the pronouncement has proved a popular strategy, but it suffers from inconsistency. This section commences with an explanation of why the students are not at liberty to dismiss the pronouncements. Then for historical interest the two possibilities are discussed: either the students could ignore the statement "*A*. There will be an exam this week"; or they could ignore the statement "*B*. The exam will be a surprise." A better solution, detailed later, is to accept the pronouncements as given but to deny that the students can reason with them fruitfully.

3.1. Normal Assertions

Were it stated, "The table is made of wood," then the assertion should be interpreted, "[It is true that] the table is made of wood." Other interpretations, such as "[It is false that] the table is made of wood," or "[I don't know whether or not] the table is made of wood" are simply incorrect. There is an implicit assertion of the truth of both of the teacher's pronouncements, then, from the simple fact that he pronounced them.

O'Beirne [12] made this explicit by considering the three propositions:

A. There will be an exam this week;
B. It will be a surprise to the pupils; and
C. The students should assume *A*,*B* to be true.

But now *C* suffers from the same problem, and he does not say how the students should believe *C* itself. (O'Beirne concluded that *A*,*B* were true but that the students were caught within a situation wherein they could not decide this. In the author's opinion this represents an important insight, and section Six demonstrates a formal way to reason about the students' beliefs without a third proposition.)

3.2. Not necessarily an exam this week

Alexander [3] modified the pronouncement:

A. There will be an exam if its conditions can be met.

Quine [7] developed the similar notion that the exam might simply fail to occur on any day of the week. At the start the students looked ahead and saw two possibilities:

1. The exam will have occurred at or before Thursday;
2. The exam will occur on Friday and the students will (in violation) be aware that it will be Friday.

They ruled out *2* and concluded *1*, and so their argument started. But, said Quine, they should have considered two more possibilities:

3. The exam will (in violation) fail to occur on Friday; and
4. The event will occur on Friday but the students cannot deduce this because they do not know whether the decree will be violated.

He wrote: "The students are not thereby assuming, even as a hypothesis for the sake of argument, that they know that the examination will take place."

This is to neglect the nature of *A*,*B* as axioms in the students' belief system, since what has been given as axiomatic must by definition be taken to be true. If the students reach a contradiction with

their axioms then they should tell the teacher that the axioms he gave were inconsistent, rather than selectively ignoring one.

To continue with Quine's reasoning it might be claimed that nothing could ever be deduced about the future. But what better position can there be from which to reason, than that of a prisoner sentenced to death or a student to an exam? (Actually, it is possible to *'reason'* about the future but still to be unable to *'know'* something. Section Five elaborates on this point.)

Moreover, the Paradox can be rephrased to guarantee that pronouncement *A* will necessarily be fulfilled with, for instance, the pack of cards introduced by Lyon [10]: a hand of card has one Ace inserted; and the cards are revealed one by one with the Ace—which must necessarily be in the hand—being the unexpected event. Other authors [35,37,47] have introduced an veritable toy cupboard of devices. As O'Beirne [13] says, "With all practical difficulties aside, the logical problem remains."

Thus Quine's tactic, of allowing no exam, is invalid for two reasons: it takes a non-sensical view of the nature of axioms; and it can be removed by reformulating the Paradox. It is surprising that his proposed solution should have become the most widely quoted paper on the subject.

3.3. Not necessarily a surprise

Cohen [2] was the first to say that the examination must be allowed to be non-surprising on the last day. Lyon [10] makes explicit two alternative versions of the pronouncement:

B1. The exam will be unexpected by the students *unless* it takes place on Friday.
B2. The exam will be unexpected by the students *even if* it takes place on Friday.

The first forestalls the argument used by the students. The second, he claims, is logically insistent because it asserts that even an exam on Friday must be unexpected: yet surely the students would be able to predict it if Friday were the only day left. In his view the paradox arises from a confusion between the two.

Nerlich [14] criticised this point with the argument that the original pronouncement simply does not sustain interpretation *B1*. He claimed that the original pronouncement '*B*. The exam will be a surprise' might *entail B1* and B2 for the simple reason that Friday will never occur, but that is not to say that it is the same as *B1* or *B2*.

However, the students are able to hypothesise about Friday and for this purpose they must take either *B1* or *B2*, since there is no third alternative. Nerlich's objection is therefore not valid. In Section Six it is proposed that even interpretation *B2* might be taken without leading to contradiction.

Nerlich's own suggestion [14] was that you cannot know the day, because *B* says you cannot: "the premise used states its inability to be used in that way." "The students' argument is invalid, since to demonstrate the negation of any of the statements is thereby to demonstrate its possibility, on a higher level one might say, owing to the queer sort of statement that *B* is."

One implication is that the pronouncement is self-fulfilling. In a clever analysis Edman [36] seemed to show that no self-fulfilling pronouncement *B* is possible. However, his argument is flawed because his version of the pronouncement does not even mention the reasoning processes of the students. The proposed Epistemological Language in Section Six captures in a formal notation the 'queerness' of *B*.

3.4. But what if the pronouncement had allowed the last day to be non-surprising?

The project of changing the pronouncement to "*B*. It will be a surprise unless it comes on Friday" has had great appeal. This is because the change prevents the students' argument at the start (since they cannot rule out Friday) and it still admits the solution of a surprise examination on, say, Wednesday.

This second redefinition has superseded the earlier "*A*. It may or may not occur this week" since different formulations can be conceived which remove uncertainty—for instance, with an Ace randomly inserted into a hand of cards.

It is the case, however, that by changing the pronouncement it becomes possible not to solve this puzzle, but a slightly different one. Instead, the solution to this puzzle must lie in deciding whether the students should believe the statements to be contradictory. This requires a careful account of the ability of the students to reason, and of the nature of their reasoning process. Both themes are developed in the next section.

4. About the reasoning of the students

This Section considers general aspects of the students' reasoning. Some authors have questioned the competence of the students. Even with the assumption that the students are able to reason, other objections have been raised: perhaps their assumptions fail; or perhaps there is a difference between imagining the future and actually experiencing it; or perhaps the students are in a different position from an external observer. Additionally, it is not clear precisely what it means 'to reason' or 'to deduce', and some authors have argued that valid deductions can change over time. Others have suggested that classical logic must be augmented with values other than just 'true' and 'false'.

Sharpe [22] suggested that if the students were stupid then the pronouncement would work—since they would not be competent to deduce. This suggestion will be discounted since it simply avoids the paradox. Instead, the assumption will be made that the students have all their faculties and reasoning powers at their disposal.

Wright [28] said that, since no proposition alone can cause events, it is not enough for the students simply to arrive at a conclusion: they must as well explicitly alert the teacher to the result of their deductions. But consider the counterexample: "If Fred deduces the truth of *X*, then I will do *Y*." If the teacher knows that Fred is a perfect deducer, and the teacher knows that Fred is able to deduce the truth of *X*, then the teacher knows that Fred will have deduced *X* because this is what perfect deducers do. Hence Fred's deduction alone is sufficient.

Meltzer and Good [23] claimed that rather than considering logical truth or falsehood, the teacher was talking about 'epistemic probabilities'—for instance, "I think that there is a twenty percent chance of a Tuesday exam." They wrote, "And if this meaning is used, the paradox becomes trivial." This is simply not true, since the students' argument could be run through with 100% substituted for a deduction of truth and 0% substituted for a deduction of falsehood and it would be exactly equivalent.

4.1. Initial hypotheses fails

The students' argument started: "Suppose we remain unexamined up to Friday. It would be a surprise. Therefore the examination can not be left until Friday." Schönberg [26] wrote: "The point is that we are entertaining the hypothesis that the week has passed. The whole argument rests on this and makes a mere verbal pretence of reconsidering the earlier days." Effectively, their argument is flawed because it does not properly discharge its first hypothesis. This is not true. The students first supposed that they were unexamined up to Friday, deduced a contradiction, discharged this assumption, and concluded that they could not be left unexamined until Friday. They then supposed that they were unexamined up to Thursday, deduced a contradiction, discharged this assumption, and concluded that they could not be left until Thursday. At each step in the argument the students are correctly discharging their assumptions.

Kirkham [44] claims that it is a different assumption that is at fault. He describes the assumption critical to the students' argument, which he calls the 'Key Premise':

A. The students expect an exam next week.

This "is true at the time the student begins his argument… But after the bright student completed his argument, the argument itself convinced the students that there would not be an examination the following week and, so, it changed their expectations in such a way as the render *A* false. The argument renders itself unsound by virtue of its own persuasiveness." Kirkham's claim is essentially the same as that of Quine and it fails for the same reason: at the end of the argument, the students still have the truth of the axiom 'there will be an exam next week' in contradiction with their conclusion 'we do not expect an exam next week'. Additionally, his claim that the truth of propositions can change over time is contrary to logic, as demonstrated in Section 4.3.

4.2. Difference between imagining actions, and experiencing them

Ayer [35] drew the distinction between two possible interpretations of the pronouncement "*B*. The exam will be a surprise":

> *B1*. For an examination on any day of the week, you will not be able to predict its day *before the week has started*.
> *B2*. For an examination on any day of the week, you will not be able to predict its day *at any time in advance*.

According to Ayer, *B1* is clearly true and *B2* is clearly false—because on Thursday evening, a Friday examination surely could be predicted. Janaway [47] claimed that these possibilities are not a resolution to the paradox, for the actual pronouncement was in fact:

> *B3*. For whichever will prove to be the actual day of the examination, you will not be able to predict its day at any time in advance.

That is, Janaway distinguished between the inability to predict for every single day in *B2* with the inability to predict for just a single day in *B3*. Effectively he tries to limit the strength of *B3* and prevent a student's hypothetical "suppose the exam were on Friday," by saying, "But if it is not going to be on Friday then you cannot draw any conclusions from the conjunction of your hypothesis with *B3*!"

Janaway's version is too restrictive. When the students are postulating an exam on Friday, they are considering the possible world which does have an examination on Friday and in which *B3* does apply to Friday. When they rule out this possible world, they go on to consider a second possible world with an examination on Thursday and in which *B3* does apply to Thursday. A position between *B2* and *B3* could be imagined:

> *B4*. For an examination on any day out of those days on which it is possible to have an examination, you will not be able to predict its day at any time in advance.

While Ayer might not claim this to be contradictory like his *B2*, this version does allows for the students' argument: they suppose the set of possible days to include Monday to Friday and arrive at a contradiction; they suppose it to include Monday to Thursday and arrive at a contradiction; and so on. In fact the distinction between *B2* and *B4* is unimportant since, as demonstrated with epistemological language in Section Six, Friday is not a contradiction and the set of possible days does cover the entire week.

Bennet [20] in his clear and concise review, drew a distinction between reasoning about the days in the future and actually experiencing them. So the pronouncement "*B*. The exam will be a surprise" could be interpreted in two ways:

> *B1*. Even imagining yourself on the eve of what would prove to be the examination, you still would not be able to deduce it;
> *B2*. Actually arriving at the eve of what would prove to be the examination, you still would not be able to deduce it.

Bennet suggested that *B2* is the correct interpretation and that it is "unproblematically true if a test occurs on Thursday or earlier." He seems to think that the students' argument is prohibited from the outset because they have not actually arrived at Thursday eve, and so they cannot reason about Friday.

(Actually, Bennet's formulations are a little more convoluted. For example "A_6. There are either zero or many days at which you could actually arrive and deduce an exam the next day." But presumably since something must be true for you to deduce it, the 'many' case can be discounted.)

The following argument is equivalent to Bennet's suggestion. "If we suppose that we arrive at Friday and deduce a contradiction, then we should conclude that we have not in fact arrived at Friday, rather than that we could never arrive at Friday".

Another objection that rests on the problem of deducing negations from contradictions is as follows. The students started with the supposition, "Were we to get to Friday…" and derived a

contradiction. From this they deduced the negation of their original supposition. But if the negation that they deduced was "We cannot get to Friday," then their original supposition would have had to have been "We can get to Friday". If instead the negation they deduced was "We have not arrived at Friday," then their original supposition would have had to have been "We have arrived at Friday". In fact it was neither: instead, the original supposition was a subjunctive 'Were we to get to Friday…' and it does not have a negation.

However, the students are actually arguing about possible future worlds. As mentioned earlier, the students imagine the situation in the possible world which a Friday examination. They deduce a contradiction and conclude that this particular possible future world is impossible and hence unreachable. Expressed in this manner it becomes clear that the students are able to imagine themselves in a situation and that it is possible to negate the subjunctive. Both objections above, therefore, are invalid.

4.3. How the students' knowledge is special

As mentioned in section 3.1, O'Beirne [12,13] made explicit a third pronouncement:

C. The students should assume *A,B* to be true.

He wrote: "The truth of the two statements and the justification of the right of the pupils to assume their truth, are two entirely different matters; and if the first two statements are in fact true, this very fact—logically and automatically—must prohibit the pupils from assuming that they are true." In the opinion of the author, the difference between the notions of the students and those of an external observer is crucial to solving the paradox.

This difference was ignored in the literature for twenty-four years, until Olin [40] introduced what Sorenson [42] later called an 'epistemic blindspot'. "A proposition p is an *epistemic blindspot* for a person a if and only if p is consistent, while Kap (for a knows that p) is inconsistent." That is, there are some true statements that a person simply cannot know (while an onlooker can.) These blind-spots are a useful notion. For instance, a problem arises since O'Beirne concluded as above that the students must not assume *A,B* to be true; but at the same time the students must assume them to be true simply because they were presented as axioms. Section Six develops the idea that epistemic blind-spots are be able to cover up this incongruity.

4.4. Deduction and consolidation

An important aspect of the Paradox is the process of deduction of the students, of how they arrived at what Nerlich [14] called "deductive correlates". Meltzer [19] says that a proposition is deducible if "it is able to be arrived at by a finite number of applications of the rules of inference." White [48] used the term 'reflection' for the act of the students arriving at the next step in their deduction.

The consolidation of knowledge, also called the 'principle of temporal retention' when the consolidation occurs over time, was first mentioned by Binkley [30] in 1968. He quoted from Cicero: "The Wise Man never opines, never regrets, never is mistaken, never changes his mind." This Binkley calls Cartesian epistemology: the "accumulation of knowledge as the slow but sure building up of a structure, brick by solid brick, upon some secure foundations."

Wright & Sudbury [38] criticised this: "The error in the Thesis argument is that of assuming a principle of temporal retention of beliefs." So the students might believe the pronouncements on Monday but not (should they remain unexamined) on Friday. Once this principle is seen to be mistaken, they said, the paradox disappears.

Kirkham [44] suggested that temporal retention of beliefs is in fact very rare. "Statements whose truth values change over time are familiar enough." He is quite wrong. His example "the shirt I am wearing at the moment is blue" appears to change over time only because it 'the shirt I am wearing' is a function of both a person and a time, whereas the proposition "The shirt that Mr Kirkham wore at 3 p.m. on April 22 1985 was blue" can never change. This difference had been noted much earlier by Cohen [2] and Alexander [3].

In a clever analogy, Sorenson [42] and Janaway [47] criticised Wright & Sudbury's argument. Sorenson gave the analogy of the Designated Student, which retains the paradox while removing time processes. He thereby demonstrated that the principle of temporal retention is irrelevant to the paradox.

However Sorenson's attempt was ill-conceived, for the consolidation of knowledge does not only apply to events in time. He and the other authors discussed above have suggested that a proposition might [correctly] be deduced as true at one stage, and then deduced false when more information is revealed. That would be equivalent to asserting that it is possible, for instance, for "*A* proves *D*" to be true and at the same time "*A* and *B* prove not-*D*." This view is logically incoherent. Deduction is defined such that the truth of the premises guarantees the truth of the conclusions, no mater what additional facts are made available.

Cicero's Wise Man would have used a notion of truth similar to that proposed by Tarski. A proposition P_1 has a set of valuations V_1 which make P_1 true. (A valuation is the assignment of a truth value to each atomic proposition). If a second proposition P_2 has a set of valuations V_2, then the proposition $P_1 \wedge P_2$ has valuations $V_1 \cap V_2$. That is, adding any further propositions to a conjunction monotonically decreases the set of possible valuations. If 'true' were the only possible valuation for P_1, then it is impossible that 'false' is the only possible valuation for $P_1 \wedge P_2$. The 'principle of temporal retention of knowledge' is simply the instance of this result where P_1 and P_2 are separated in time.

4.5. The excluded middle, and modal logic

Weiss [6] claimed that this paradox was an example of the invalidity of the Law of the Excluded Middle. "There is a great difference between 'it is true that either *x* or non-*x* is the case' and 'it is true that *x* is the case or it is true that non-*x* is the case', between $f(x \vee \neg x)$ and $f(x) \vee f(\neg x)$." He relates this to the problem in the paradox of isolating at the separate propositions from the disjunction $e_1 \vee e_2 \vee e_3 \vee e_4 \vee e_5$. In a similar vein Woolhouse [31] wrote that the paradox arose "because 'it is deducible that it will not occur on that day' is not the negation of 'it is deducible that it will occur on that day.'" He suggested that the use of multi-valued logic is necessary. Not so, according to Meltzer [23] writing several years earlier: "[it] only appears so because of an insufficiently sophisticated use of logic." Binkley [30] was the first to demonstrate a 'more sophisticated use of logic'—modal logic, with the predicate 'J' to stand for 'the students have justified belief that'. Various modal logics have been used extensively from this point on.

It should be noted that modal and multi-valued logics are not significantly different from classical logic: they are embedded within classical logic and give it structure [54]. In the case of this paradox there are a finite number of days and thus a finite number of modal propositions, and so they could be replaced with non-modal propositions. The only reason for using modal or multi-valued logics, then, would be for clarity of expression.

This section has established that the students may be considered ideal thinkers, that in their argument they did properly discharge their hypotheses, that students are able to make hypotheses about future possible worlds, and that deduced knowledge must always be consolidated. A complete formal argument, which would include the notions that knowledge for the students is different from that for

an observer and that they suffer from epistemic blind-spots, is postponed. The next section presents formally the argument used by the students, which did not consider these points.

5. Logical formalisms and self reference

Many authors have asserted that the pronouncements are contradictory. But reasoning with language is tricky and error-prone—hence the need for symbols. Symbolic arguments were used first by Kaplan and Montague [11] in 1960, and extensively thereafter. A good notation matches the problem domain: this section therefore starts with a discussion of self-reference, which is a key part of the paradox. A formal version of the students' argument follows.

5.1. Self-referential statements: Gödel

Shaw [9] was the first to identify a crucial aspect of the pronouncement: it refers to itself, or is 'recursive'. He illustrated what he considered a non-recursive pronouncement *P1* that is different from the original:

A. It will occur this week;
B. It will be unexpected, in that it is not possible to deduce its day from *A*.

and a recursive pronouncement *P2* which seems to correspond to the original pronouncement:

A. It will occur this week;
B2. It will be unexpected, in that it is not possible to deduce its day from *A* and *B2*.

and *P3*, which he thought was a non-recursive version of *P2*:

A. It will occur this week;
B3. It will be unexpected, in that it is not possible to deduce it's day from *A*.
B4. … from *A, B3*.
B5. … from *A*, B3, *B4*.
(up to as many as there were days in the paradox).

Curiously, Shaw claimed "it is clear that the origin of the paradox lies in the self-referring nature of rule *B2*." This is curious [14,20] because he himself claimed that *P3* is equivalent to *P2* but is not self-referring: if this is so, why would *P3* remain a paradox as he maintained? In fact, the argument falls down for two reasons: the chain should extend indefinitely; and *P1* and *P3* are actually both recursive because, taken as a whole, parts of it refer to other parts [57]. For instance, *P1* might be taken as the dual simultaneous identity:

(*A, B*) = (an exam will occur, you can not deduce an exam from *A*)

In any case, it is clear that many recursive statements are not contradictory. Lyon [10] gave the example, "This sentence is written in black type." And as Nerlich [14] says, "Gödel would seem to have demonstrated that at least one [statement] can [refer to itself]."

An intriguing pun was suggested by Cargile [29] between this Prisoner's Paradox, and the Prisoner's Dilemma. After all, both puzzles have two parties each of whose actions and beliefs depend recursively on those of the other. But his solution for maximal co-operative rewards is incorrect (the actual probability is 1/12), and this Paradox differs in that the teacher or judge is in a position of power and of authoritative ordainment, whereas the two prisoners in the dilemma are equal.

There have been many references to Gödel, starting with Kaplan and Montague [11]. This would seem reasonable, since the pronouncement does refer to the way in which the students would interpret its textual content: and Gödel allows us to manipulate the textual content of a proposition as well as its value. The proposed epistemic language in Section Six uses fixed points, which underpin Gödel's Theorem.

Edman [36] makes the very important point that any recursive statement is a description of a proposition, rather than a definition. Descriptions may admit none, one, or several solutions: "It is in

fact possible that there are several correct predictions. Another possibility is that there are no correct predictions."

To make this more clear the simple mathematical example of the factorial function will be considered. It is recursively described as follows:

factorial(n) = n x factorial(n-1), or 1 if n=0

There are multiple versions for the function 'factorial' that satisfy the above description. One possible factorial function evaluates to zero when given any negative number. Another possible factorial function evaluates to '*undefined*' for all negative numbers since, for instance *undefined* x -5 = *undefined*.

Thus, the teacher gave a description of a pronouncement rather than a pronouncement itself. This must be reflected in any notation. An attractive resolution to the paradox might appear to be that two contradictory solutions might satisfy the description: for example, the students' "no exam is possible" and the teacher's action "exam on Wednesday." Unfortunately this is not the case. As demonstrated below, it is only the students' conclusion that fits the recursive description.

5.2. Notation

There is no consistent notation in the literature. The author has transcribed expressions from past papers into the following symbols:

D	the set of days. $D \equiv \{0,1,2,3,4,5\}$, where 0 is Sunday.
e_i	the proposition "the exam is on one of days i"
$K_{a,i}(P)$	the proposition "at times i, people a know propositions P"
$H_i(P)$	the event described by propositions P occurred at time i
$A \vdash B$	B is logically deducible from A
$\rightarrow$	classical implication, ie. $\supset$
$\otimes$	exclusive or operator
"X"	the Gödelisation of proposition X
[["X"]]	meaning of Gödel string "X", which is just X

5.3. The students' argument, formally

This section contains a formal recursive definition of the pronouncement and a note on the students' solution. It is useful to have a symbolic version of the argument rather than a version in English because it makes explicit all assumptions.

The argument relates to a set of possible days for the exam, since it would seem that this is what the teacher was telling the students. (In fact this is not necessarily justified, as explained in Section Six). Using sets is an easier route than natural deduction because it ensures that no rules or their possible consequences have been accidentally omitted. Moreover, sets are the natural language for reasoning about recursive statements. Actually, the teacher is strictly talking about valuations—assignments of truth values to atomic propositions—rather than possible days, but since the atoms are e_1, e_2, ... e_5 and they are mutually exclusive the two are equivalent. Let the teacher's proposition, as a set of possible days, be $\mathcal{Z}$. If there exist multiple solutions for $\mathcal{Z}$, then multiple pronouncements that satisfy the teacher's description would be possible.

A typical step in the argument is as follows. The act of supposing that the students had reached Tuesday evening would be equivalent to hypothetically ruling out e_1 and e_2. Before this supposition they knew that $\mathcal{Z}$ days were possible solutions. Given the supposition, the possible days would be narrowed down to $\mathcal{Z} \cap \{e_3, e_4, e_5\}$. If it happened that the intersection gave only one result, say $\{e_3\}$,

then in this hypothetical situation they would have deduced the exam to be on Wednesday, since they also have been told that one of e_1, e_2, … e_5 is true. The hypothetical situation would be then ruled out.

A subtlety in the approach below is the division of the pronouncements into two descriptions of $\mathcal{Z}$ that do not refer to each other. The first states at least one day must be possible—the exam will then occur on that day. The second states that the exam will be a surprise, and already has the first proposition built into it. Hence they are not mutually referring.

$$
\begin{array}{lll}
|\mathcal{Z}| & \geq 1 & \text{(the number of elements in } \mathcal{Z}\text{)} \\
 & & \\
\mathcal{Z} & = & \{e_1\} \text{ if } \mathcal{Z} \cap \{e_1,e_2,e_3,e_4,e_5\} \neq \{e_1\} \\
 & \cup & \{e_2\} \text{ if } \mathcal{Z} \cap \{e_2,e_3,e_4,e_5\} \neq \{e_2\} \\
 & \cup & \{e_3\} \text{ if } \mathcal{Z} \cap \{e_3,e_4,e_5\} \neq \{e_3\} \\
 & \cup & \{e_4\} \text{ if } \mathcal{Z} \cap \{e_4,e_5\} \neq \{e_4\} \\
 & \cup & \{e_5\} \text{ if } \mathcal{Z} \cap \{e_5\} \neq \{e_5\}
\end{array}
$$

The solution is trivial, and can be found constructively by supposing $\mathcal{Z}$ to contain first e_5 and deducing a contradiction, and then e_4 and so on. Alternatively it could be found by listing all 32 possible values for $\mathcal{Z}$ and checking which of them satisfied the equations.

In fact, $\mathcal{Z} = \{\}$ is the only solution that satisfies the second description. This is how the pupils deduced that no day could possibly have an examination. They should have gone a step further and concluded that no $\mathcal{Z}$ satisfied both propositions, and so the teacher's pronouncement was self-contradictory.

However, as discussed in the next section, the above argument is flawed. This is because rather than "*you* cannot know the day", it assumed the more general "the day cannot be known." To capture the paradox, the notation must be extended to what is deduced by whom.

6. A new epistemological notation

In this Section the author proposes a resolution to the paradox. Essentially, a distinction is made between the thought processes of the students and those of an external observer. The conclusions reachable by the external observer are taken as objectively true. The reasoning process is represented as a step-by-step process that as a minimum presumably allows for natural deduction. Minimal properties of 'knowledge' are presented: if a person can deduce a thing without possibility of contradiction, they might be said to know that thing. In a simple case such as "A but you do not know that A", several forms of reasoning may take place: the students' reasoning process might continue in circles; or they might recognise the fact and conclude that the axioms were contradictory; or they might look further and see what an external observer would conclude. But the last case requires that the students recognise that their powers of reasoning are limited, and this contradicts the earlier supposition that reasoning as a minimum must contain natural deduction. Hence, in the case of the paradox, the bright students will eventually accept the axioms and will simply not apply further deductions to them.

The remainder of this section developes a comprehensive framework for epistemological logic and expresses the explanation above within this framework. It first describes the chain of deductive steps in the process of reasoning. It is shown that, for a bright student to be aware of her situation, she must also be able to recognise that natural deduction is not an appropriate form of reasoning. The single-day and for the two-day forms of the paradox are then considered in the light of this limitation. A mathematical description of epistemological language follows, with a formal expression of the teacher's pronouncement, and ramifications are explored in the conclusion.

6.1. Mental processes in the students

As O'Beirne [12] noted, the pupils are in a different position from an external observer. The teacher pronounced: "*you* will not be able to know," so in the statement of the problem itself propositions are predicated upon people. It therefore seems not unreasonable to choose a notation that corresponds to the problem domain, and to predicate all other propositions upon people: for instance, "P deduces that A is true" rather than just "A is true". Indeed, if a notation was used that did not predicate its statements thus, it would simply be unable to express the teacher's original pronouncement. Therefore, such a strategy is a valid approach the problem. It follows that "P deduces that A is contradictory" may be different from "A is contradictory." Were it to be shown that the two always in fact turn out the same, then it might be concluded that the distinction did not have a valid place in logic. However, it will be shown that the two are not always the same. It should be noted that the introduction and exploration of such distinctions is necessary part of logic. For instance, modal predicates must be introduced to distinguish between 'A is true' and 'P knows that A is true', since the negation of the first is 'A is false' whereas the negation of the second is 'P does not know that A is true.' Without modal predicates, it would be impossible to reason about the difference.

Consider the starkest form of the paradox:

$$A \wedge \neg K(A)$$

There are different possible interpretations of the predicate 'to know':

1. The student feels that to deduce a thing is to know that thing. She deduces from the axioms that A is true, and this deduction means that she knows that A is true. She therefore concludes that the axioms were contradictory.
2. The student has dismissed the axioms and can make no other conclusions. Therefore the student does not know anything about A.

If it were taken that the deduction of a fact was the same as knowing that fact, then the student would be right. But the second conclusion suggests that the normal usage of the word is more refined than simple deduction. Consider the following axioms:

1. $A \wedge \neg A$

From the axioms it is a valid step to deduce 'A'. But the student could not be considered to *know* A in this case: she should rather conclude that the axioms were contradictory. Therefore the author suggests that for a person P to know A, it must be the case that P can deduce A without the possibility of further contradiction.

The remained of this section examines the reasoning process. Logical reasoning is the process of going through various states of mind. There are a great many possible states of mind $\mathcal{T}$, but not all will be reachable from a given starting point. Let $\mathbb{T}$ be the set of only those states of mind reachable from a given set of axioms. Reasoning starts at some state of mind $t_0 \in \mathbb{T}$ and, via suppositions and deductions, reaches other states of mind. Deterministic reasoning would imply a linear chain of states of mind; non-deterministic reasoning would generate a branch at any point which has more than one action available. This paper considers a tree rather than a chain, since the tree subsumes both cases. The tree is also given the label $\mathbb{T}$.

$$t_0 \rightarrow^f t_1 \rightarrow^f t_{2,1} \rightarrow^f t_{3,1} \rightarrow^f t_{4,1} \rightarrow^f \ldots$$
$$\qquad\quad \llcorner\!\rightarrow^f t_{2,2} \rightarrow^f t_{3,1} \rightarrow^f t_{4,1} \rightarrow^f \ldots$$

The starting point t_0 is the state of initial knowledge about the problem, and may be considered to be a textual form of the axioms as given without any deductions yet applied. Given a model of reasoning, the tree is uniquely defined by this initial state t_0. The relation $f: \mathcal{T} \rightarrow \mathcal{T}$ describes what possibilities there are for the next step of reasoning. Thus, the set $\mathbb{T}$ of reachable states is the closure of $\{t_0\}$ under f. A natural model for the reasoning relation f might the Sequent Calculus, which has the same properties as Gentzen's Natural Deduction [52] except that it makes clear the discharging of assumptions. If the sequent B comes from A through a single application of one of the rules of natural deduction, then it is the case that f(A)=B. Note that this implication is in a single direction: there may be more to the reasoning process than simple natural deduction, but the assumption for now is that natural deduction is a bare minimum for rational deduction. It will, however, later be suggested that the reasoning process may not even be allowed to contain this.

It may be claimed that reasoning is a continuous process rather than a series of discrete jumps. However, the mental state may be sampled just after each time a rule of inference has been applied, giving rise to a discrete chain of reasoning embedded within the continuous process. The following arguments would then apply to this discrete chain.

In general it is the attitude of the subject towards a particular set of propositions that is of interest, rather than their entire state of mind. One representation of attitudes, due to Ginsberg [53], is with the four possibilities 'true', 'false', 'undefined' and 'over-defined / ambiguous'. These are equivalent to the sets of possibilities that a person may still entertain about a proposition: {t}, {f}, {t,f}, {}. For instance, if it is only proposition A that is of interest, then the infinite chain of states can be projected onto a chain over just four states:

(A is true)	(A is undefined)
(A is false)	(A is overdefined)

A simpler representation is possible, and shall be used in this paper. When it is only the truth of a proposition that is of interest, then all other states may simply be projected onto a single state ⊥, pronounced 'floor', and the 'is true' is assumed:

(A [is true])	(⊥)

Let $\mathcal{S}$ be the set of propositions whose truth may be considered, let $\mathbb{A} \subseteq \mathcal{S}$ be a set of mutually exclusive assertions about a particular proposition, and define $\mathbb{A}_\perp = \mathbb{A} \cup \{\perp\}$. States of mind about a particular proposition are elements of $\mathbb{A}_\perp$, which is also called 'attitude space' since each element describes the attitude of the subject to the truth of of a proposition. For each assertion space $\mathbb{A}$ there is a projection m, mentioned above, which maps entire states of mind onto states of mind about a particular proposition:

$$m_{\mathbb{A}}: \mathcal{T} \rightarrow \mathbb{A}_\perp$$

Since the propositions in assertion space are mutually exclusive, m must map every $t \in \mathcal{T}$ onto no more than one assertion $a \in \mathbb{A}_\perp$. Since all states that do not make such an assertion are mapped onto ⊥, m must map every $t \in \mathcal{T}$ onto at least one element of $\mathbb{A}_\perp$. Therefore, m is a function. Note that attitudes must be stated explicitly: at a state in the tree at which 'A∧B is true', it may be the case that the deduction rule for deducing 'A is true' has not yet been applied, and so this state would map to ⊥ rather than to 'A true'.

The tree of states of mind $\mathbb{T}$ generated from t_o gives rise to a collection of sequences over $\mathbb{A}_\perp$. For instance, if $\mathbb{A}_\perp = \{x,y,\perp\}$, then the sequence corresponding to one path through the tree might start '⊥,⊥,⊥,x,x,⊥,y…' while the sequence corresponding to an alternate path to the tree might start '⊥,x,x,y,⊥…'

A reduced sequence a_i is the sequence generated by mapping from the tree $\mathbb{T}$ to $\mathbb{A}_\perp$ only those states which have discharged all their assumptions. In Gentzen's Natural Deduction this is hard to determine; in the Sequent Calculus, any conclusion which has discharged all its assumptions has the form $\Gamma \Rightarrow \Delta$ where Γ contains only the initial axioms. A 'conclusion' in this paper is taken to mean the deductive results that require no assumptions other than the axioms.

To Cicero's description of the Wise Man, a further description may be added: "The Wise Man never forgets." Locke [58] in his essays on Human Understanding put it thusly: "The next faculty of the mind, whereby it makes a farther progress towards knowledge, is that which I call retention, or the keeping of those simple ideas which from sensation or reflection it hath received." The reasoning relation f is taken to be cumulative upon its conclusions. That is, whenever a conclusion has reached with all assumptions discharged, that conclusion is added to the store of knowledge. Hence in a reduced sequence over $\{x,y,\perp\}$, the only symbol that can follow an x is a further x or a ⊥, and likewise for y.

The requirement for knowledge suggested above said that, to know a thing, the student must be able to deduce that thing without further possible contradiction. This definition may be expressed more formally. The proposition A is 'knowable' when, over all assertion spaces $\mathbb{A}_\perp \supseteq \{A,\perp\}$, the following hold:

1. There is at least one reduced sequence containing 'A is true'; and
2. Every reduced sequence containing 'A is true' continues with it.

Equivalently, no matter which route is taken through the reasoning tree, the conclusion 'A is true' once reached is 'stable under deduction'. In those sequences that do contain A, it is said to be a 'fixed point' of the sequence. Another way to phrase this is to say that a knowable proposition A 'is a fixed point of reduced deduction onto $\mathbb{A}_\perp$'. If a person came to know something, then it is not

possible for him, through continuing his reasoning upon the same axioms, to come to know its negation. If he did come later to deduce its negation, then he can not have 'really known' it in the first place. As Cicero said: "The Wise Man never changes his mind."

Further requirements to knowledge have been proposed, such as the requirement that a person must as well recognise the fact that P is 'knowable' before they could be said to know it. Locke [59] wrote: "Knowledge then seems to me to be nothing but the perception of the connection and agreement, or disagreement and repugnancy, of any of our ideas." He presumably meant that to know a fact is to *recognise* that all relevant trains of thought must be in agreement on the truth of the proposition. After all, when following through just a single train of reasoning to reach an answer, one is often left with the worrying feeling that an alternative route may have lead to a different answer. However, there is the possibility that the 'recognition' of the fact is itself only one possible conclusion down one of the many branches; and that the opposite conclusion might have been drawn down a different branch. Maybe, then, for a person to know P requires that he knows that he knows that he knows P, and so on. Further discussions on knowledge are possible but shall not be pursued.

Consider the attitude space:

(A is true) ($\perp$)

Here, $\perp$ is 'not maintaining that A is true', rather than 'maintaining that A is not true': the subject may simply have failed yet to consider the truth of A; or there might have been insufficient information for any judgement to be made on A.

In the particular attitude space given, knowledge has the following properties:

K(A) $\Leftrightarrow$ 'A is true' is a reachable fixed point for some reduced sequence.

$\neg$K(A) $\Leftrightarrow$ Either all sequences have a fixed point at $\perp$, or in some sequence that reaches 'A is true', there are further valid deductive steps that lead to $\perp$.

6.2. Epistemological propositions

This section briefly reviews the syntax of non-quantified predicate logic with only the knowledge predicate K(·). This is motivated by the desire for a formal definition for 'assertion space' and by the need to describe valuations. K is subscripted $K_{person,"t"}$, where "t" is a description of the starting point of a person's chain of reasoning. It is shown here as the Gödelization of a statement of the axioms, but other descriptions are possible. The argument P in K(P) is some logical proposition. In the discussion below, $\aleph$ is taken to be entire set of predicates $K_{p,"t"}$, over all p,t. Note that description of the language is recursive, since the "t" is itself any string in the language. This section is only a summary of relevant results: texts such as [54] would provide a better introduction for those not familiar with the area. It can be safely omitted for those not seeking a formal explanation of the problem.

A propositional language L (without predicates) is a pair (P,O) consisting of a set of atomic sentences P together with a set O of operators—generally $\vee$, $\wedge$, $\neg$, $\rightarrow$. The set S_1 of sentences in L is the smallest set containing P that is closed under O. This is an infinite set. Here are some elements, where P is the set {A,B}:

$S_1 \quad \supseteq \quad \{ A,\ B,\ \neg A,\ \neg B,\ A \vee B,\ A \wedge B,\ A \rightarrow B,\ B \rightarrow A \}$

The set S_2 of sentences in first order epistemological language, which describes knowledge about first order propositions such as $K_{p,t}$ (A$\wedge$B), is given by

$S_2 = S_1 \cup (\text{closure of } \mathcal{K} \times S_1 \text{ under O})$

where $\mathcal{K} \times S_1$ prefixes all elements in S_1 with all possible predicates $K_{p,t}$. Here are some elements of S_2:

$S_2 \supseteq \{ A,\ B,\ \neg A,\ A \vee B,\ K_{p,t}(A),\ K_{p,t}(\neg A) \wedge K_{p,t}(A \vee B) \}$

Higher order epistemological language allows predicates to apply to themselves. For instance, $K_{p,t}$ ($A \wedge K(B)$) first appears in S_3. Let h: $S_i \rightarrow S_{i+1}$ be a function generating the next higher order language:

$h(X) = X \cup (\text{closure of } \mathcal{K} \times X \text{ under O})$

Then the set of epistemic sentences $\mathcal{S}$ for a set of atomic propositions A is given by:

$\mathcal{S} =$ closure of A under h

This set contains all possible sentences that can be made with the knowledge predicate K. The teacher's pronouncement must be one of them.

A valuation is the assignment of a value to each variable. A pronouncement is a set of possible valuations. There may be several pronouncements that satisfy the description given by the teacher. A 'pronouncements set' is the set of all possible pronouncements. For instance, a sentence such as 'A' which contains only a single atom has possible valuations defined by the following pronouncements set:

{ {t}, {f}, {}, {t,f} }

That is, the sentence 'A [is true]' has an answer that is one of the following: either 'A is true', or 'A is false', or 'A can be neither true nor false', or 'A can be either true or false'. In fact, the correct answer is the first. A sentence with two atomic symbols such as 'A∧B' has a larger pronouncements set:

{ {ff,ft,tf,tt },
{ff,ft,tf}, {ff,ft,tt}, {ff,tf,tt}, {ft,tf,tt},
{ff,ft}, {ff,tf}, {ff,tt}, {ft,tf}, {ft,tt}, {tf,tt},
{ff}, {ft}, {tf}, {tt},
{} }

In fact, if there are *n* terms in a sentence, then a prononcement (which is a set of possible valuations) is an element of the powerset $\mathcal{P}(\mathcal{B}^n)$, where $\mathcal{B}$ is the set of booleans {true,false}.

In a sentence in $\mathcal{S}$, such as 'A ∧ K(B) ∧ K(K(C))', the terms are all those predicates used. So the example would have terms { A, B, C, K(B), K(C), K(K(C)) }. The pronouncements set is again the power set $\mathcal{P}(\mathbb{B}^{|\text{terms}|})$, where $\mathbb{B}$ are booleans {t,f} and '|terms|' is the number of elements in the set of terms. Thus, the teacher's description says which of the elements of this power set are possible pronouncements.

6.3. The teacher's pronouncement, and solutions

Suppose the teacher had made the pronouncement: "You will be examined today, but you will not know the fact before the examination." Formally, this is:

$Z = A \wedge \neg K_{\text{students,"z"}}(A)$

The relevant assertion space is

(A is true) ($\perp$)

Now, (A is true) cannot be a fixed point of any student's chain of reasoning starting with Z. This is because at any stage the student is at liberty to take the second axiom, $\neg K_{\text{students,"z"}}(A)$, and from the description of knowledge above conclude that any sequence with an 'A is true' in it must later have a $\perp$.

Suppose that the student looked at the statement and deduced what an external observer would deduce: namely, that X: (A=true, $K_{"Z"}$(A)=false) was the only possibly valuation. The state of mind that says that X is the only possible valuation, must fall when projected into one of the states { (A is true), (⊥) }. Clearly it falls into the first. However, as described above, the first state cannot be a fixed point of any student's chain of reasoning. Therefore the students are unable to *know* that X is the only possible valuation. But note that X remains a possible fixed point for the chain of reasoning of anyone who is not a student.

Suppose that the student arrived at ⊥, and concluded that the axioms were contradictory. The answer X given above is still consistent with the conclusion of the students: for it is true that they do not know A. This is a case where the conclusion 'contradictory axioms' on the part of the students is not the same as the conclusion 'contradictory axioms' on the part of an external observer.

Suppose the students had read this paper, and knew what was happening: could they not step above and reach a definite solution? Suppose that the students are so able. As shown above, the fixed point cannot be at (A is true); for similar reasons it cannot be at (A is false). Thus, if it is required that the student is always able to reach a definite conclusion, then it must at (⊥). This definite conclusion does not assert either that A is true or that A is false, since assertion space is mutually exclusive. Therefore it must either assert that A is a third different value in some multi-valued logic, or the definite conclusion that the student reaches must be incapable of making any claims about A. Perhaps the conclusion reached is that natural deduction is pointless in this situation. That one of these possibilities is the case comes inevitably from the hypothesis that definite conclusions are always possible. Alternatively, the students may be unable to reach a definite conclusion and so be condemned to perpetual logical circles.

Does a Z exist, which satisfies the description? Trivially it does, for Z is simply a proposition, or a sentence which is an element of $\mathcal{S}$, and the question means as little as the equivalent, "does the proposition A∧B exist?" Are there any valuations which satisfy the description? Trivially there are: (A=true, $K_{students,"Z"}$(A)=false) makes Z true. The fact that the students are not able to know this fact does not alter the fact of its validity. Indeed, any observer not covered by the term 'students' in the knowledge predicate K is able to know that this is a solution.

Thus, even the starkest form of the paradox can be resolved. In the case of the full week, the students' argument breaks down since they are unable to rule out Friday. The conclusion that they must reach instead is that all days are possible for an examination. Using the higher order epistemological language described above, the statement of the paradox is lengthy and difficult to work with. The epistemological language was introduced to make explicit all assumptions and to reason formally about the paradox, and the resolution of the paradox above has justified its introduction. However, the language is less useful for expressing the teacher's pronouncement. The solution given below, therefore, is a very rough informal sketch. The original pronouncement was as follows:

A. There will be an examination next week (Monday to Friday);
B. You will not know in advance which day it will be.

Effectively, before any deduction has occured, there are five possible days for the examination. In the case that it will be on Monday, the students will be unable to know on Sunday evening that the examination will be on Monday. In the case that it will be on Tuesday, the students will be unable to know either on Sunday evening or on Monday evening that it will be on Tuesday. In the case that it will be on Wednesday, the students will be unable to know the fact on Sunday evening, Monday evening or Tuesday evening. Similarly for the rest of the week. The notation e_1,e_2,e_3... for Monday,

Tuesday, Wednesday... is used. K_s abbreviates '$K_{students}$,' and 'A+"B"' abbreviates '"A∧B"'. ⊗ is the exclusive 'or' operator.

$$
\begin{array}{lll}
Z & = & (e_1 \otimes e_2 \otimes e_3 \otimes e_4 \otimes e_5) \quad \Lambda \\
 & (& e_1 \ \Lambda \ \neg K_{s,"Z"}(e_1) \\
 & \vee & e_2 \ \Lambda \ \neg K_{s,"Z"}(e_2) \ \Lambda \ \neg K_{s,\neg e1+"Z"}(e_2) \\
 & \vee & e_3 \ \Lambda \ \neg K_{s,"Z"}(e_3) \ \Lambda \ \neg K_{s,\neg e1+"Z"}(e_3) \ \Lambda \ \neg K_{s,\neg e1+\neg e2+"Z"}(e_3) \\
 & \vee & e_4 \ \Lambda \ \neg K_{s,"Z"}(e_4) \ \Lambda \ \neg K_{s,\neg e1+"Z"}(e_4) \ \Lambda \ \neg K_{s,\neg e1+\neg e2+"Z"}(e_4) \\
 & & \Lambda \ \neg K_{s,\neg e1+\neg e2+\neg e3+"Z"}(e_4) \\
 & \vee & e_5 \ \Lambda \ \neg K_{s,"Z"}(e_5) \ \Lambda \ \neg K_{s,\neg e1+"Z"}(e_5) \ \Lambda \ \neg K_{s,\neg e1+\neg e2+"Z"}(e_5) \\
 & & \Lambda \ \neg K_{s,\neg e1+\neg e2+\neg e3+"Z"}(e_5) \ \Lambda \ \neg K_{s,\neg e1+\neg e2+\neg e3+\neg e4+"Z"}(e_5) \quad)
\end{array}
$$

One of e_1, e_2, e_3 ... must be true to fulfill Z. If, for instance, e_2 were true, then the third line would imply $\neg K(e_2)$. Since every other e_i is false, and since it is not possible to know something that is false, every other $K(e_i)$ must also be false. That is, every valuation that satisfies Z must have all K predicates false and one e_i true. Using the same analysis as for the starkest case, all of these solutions are indeed consistent under the definition of knowledge. This paper does not attempt to give all analyses, since they are all lengthy and all are similar to the single day version.

In fact, there are 31 possible sets of valuations for Z, or 31 different pronouncements that satisfy the teacher's description. A valuation strictly is a list of assignments to every predicate. For instance:

(f,f,t,f,f,f,f,f,f,f,f,f,f,f,f,f,f,f,f,f,f)

This is the particular valuation that has true for e_3, and false for every other e_i and false for every $K_{s,...}(e_i)$ predicate. This notation is unwieldly, so below the shorthand is used where each set $\{e_1,e_2,e_3\}$ represents one pronouncement, or a collection of valuations:

$\{e_1,e_2,e_3,e_4,e_5\}$,
$\{e_1,e_2,e_3,e_4\}$, $\{e_1,e_2,e_3,e_5\}$, $\{e_1,e_2,e_4,e_5\}$, $\{e_1,e_3,e_4,e_5\}$, $\{e_2,e_3,e_4,e_5\}$,
$\{e_1,e_2,e_3\}$, $\{e_1,e_2,e_4\}$, $\{e_1,e_2,e_5\}$, $\{e_1,e_3,e_4\}$, $\{e_1,e_3,e_5\}$, $\{e_1,e_4,e_5\}$, ...
... $\{e_1\}$, $\{e_2\}$, $\{e_3\}$, $\{e_4\}$, $\{e_5\}$

However, with his description the teacher simply indicated that one or more of the above is the case; and since all are subsumed within the first term, the first term is the maximal proposition that can be drawn by an external observer with respect to possible examination days:

C. The examination may occur on any day of the week.

6.4. Conclusion

It is significant that the students may have deduced that the axioms were contradictory, while an external observer would have deduced (correctly, as it turned out) that they are not. For instance, in the single day case the external observer is able to see that the exam will occur on that day and that the students will be unable to predict it. The students on the other hand will either fail to conclude anything, or they will conclude that the axioms were contradictory. It is crucial to the resolution of this paradox that only those conclusions made by an external observer are authoritative. The conclusions drawn by anyone subject to the proposition are necessarily suspect, because their reasoning falls within the logic. This is similar to the situation in Quantum Mechanics, where the experimenter is himself part of the experiment and is unable to percieve the true values of, for instance, both momentum and position of a particle.

It has been shown that within the system of epistemological logic as described, there are statements which are true, and which a subject has been told are true, but which the subject cannot *know* are true. It should be noted that this result bears a striking resemblance to Gödel's theorem, which states that in a formal system there are statements which are true, but cannot be *proved* true.

Additionally, the proof of epistemic blind-spots bears a resemblance to the proof for Gödel's theorem which also uses a fixed-point argument.

An external observer is one to whose reasoning no knowledge predicate refers. The thought processes of this external observer are not limited by any propositions, and the conclusions of the observer are not suspect. It is interesting to note that if God is conceived as that which 'knows all truths,' or is an 'external observer,' then by definition it is impossible to apply the same predicate 'knows' to God as is generally applied to humans.

The Principle of Consolidation of Knowledge must be considered with care in epistemological logic. From the axiom 'A∧B' a person might conclude that A is true. However, if an additional axiom '¬A' were added, then they can no longer hold the same conclusion. Thus, the first caveat to the principle is that the addition of apparently contradictory axioms can stop a person from knowing something. It remains the case, however, that if they initially concluded A, then no axiom could be added to make them conclude ¬A.

The second caveat is that the Principle applies only to the non-epistemological sentences. After all, it might initially be the case that a subject concludes that he does not know A simply because no axioms have mentioned A. However, if the axiom 'A' were introduced, then the subject would conclude that he does know A. 'Know A' is an epistemic sentence, and so is not subject to the Principle of Consolidation of Knowledge: hence its truth value may well change under addition of further axioms.

7. Discussion

The conclusion rests upon the definition given of knowledge, as the deduction of a statement which does not admit contradiction. It may be that alternative understandings of knowledge may be possible, which lead to different results. The author hopes that the epistemological notation presented above may prove useful in formulating precise definitions for deductive knowledge.

There is a strong correspondance between Philosophy of Knowledge and Computer Science. The Curry-Howard isomorphism, a description of the parallel between computation and proof transformation in intuitionistic logic, relates constructs in programming languages to logical operators. The step-by-step reasoning described above, governed by 'deduction', is exactly analogous to the step-by-step execution of a computer program which is governed by the programming language's 'operational semantics' [55]. The tree of reasoning corresponds to 'denotational semantics' [56] which ascribes meaning to an entire computer program. The teacher's original pronouncement is very similar to 'Floyd-Hoare' partial logic [51], which is used to specify the start and end conditions of a section of code. When a computer program becomes stuck in an infinite loop, its denotation is given by the symbol $\perp$. Much research in the field of computer semantics has been motivated by the need to cope with non-termination. Results in this area are important because they help ensure the well-functioning of safety-critical computer systems, such as those in nuclear power plants and airplanes.

It should be noted that a possible approach to a paradox that has troubled philosophers for many years should come Computer Science. The author suggests that a great potential exists: for philosophical puzzles like this Paradox to find practical applications such in, for instance, airplane safety; and for techniques in Computer Science to stimulate fresh approaches to philosophical questions.

8. Bibliography

[1] O'Conner DJ. *Pragmatic Paradoxes*. Mind 57 1948 pp. 358-359

[2] Cohen LJ. *Mr O'Connor's "Pragmatic Paradoxes"*. Mind 59 1950 pp. 85-87

[3] Alexander P. *Pragmatic Paradoxes*. Mind 59 1950 pp. 536-638

[4] Scriven M. *Paradoxical announcements*. Mind 60 1951 pp. 403-407

[5] O'Connor DJ. *Pragmatic paradoxes and fugitive propositions*. Mind 60 1951 pp. 536-538

[6] Weiss P. *The prediction paradox*. Mind 61 1952 pp. 403-407

[7] Quine WvO. *On a so-called paradox*. Mind 62 1953 pp. 65-67

[8] Ebersole FB. *The definition of "pragmatic paradox"*. Mind 62 1953 pp. 80-85

[9] Shaw R. *The paradox of the unexpected examination*. Mind 67 1958 pp. 382-384

[10] Lyon A. *The prediction paradox*. Mind 68 1959 pp. 510-517

[11] Kaplan D, Montague R. *A paradox regained*. Notre Dame journal of formal logic 1 1960 pp. 79-90

[12] O'Beirne TH. *Can the unexpected never happen?* New Scientist 25 May 1961 pp. 464-465

[13] O'Beirne TH. (letters and replies). New Scientist 25 May 1961 pp. 597-598

[14] Nerlich GC. *Unexpected examinations and unprovable statements*. Mind 70 1961 pp. 503-513

[15] Gardner M. *A new prediction paradox*. British Journal for the Philosophy of Science 13 1962 p. 51

[16] Popper KR. *A comment on the new predicition paradox*. British Journal for the Philosophy of Science 13 1962 p. 51

[17] Fitch F. *A Gödelized formulation of the prediction paradox*. American Philosophical Quarterly 1 1964 pp. 161-164

[18] Medlin B. *The unexpected examination*. American Philosophical Quarterly 1 1964 pp. 66-72. Corrigenda p. 333

[19] Meltzer B. *The third possibility*. Mind 73 1964 pp. 430-433

[20] Bennet J. (review). Journal of Symbolic Logic 30 1965 pp. 101-102

[21] Cargile J. (review). Journal of Symbolic Logic 30 1965 pp. 102-103

[22] Sharpe RA. *The unexpected examination*. Mind 74 1965 p. 255

[23] Meltzer B, Good IJ. *Two forms of the prediction paradox*. British Journal for the Philosophy of Science 16 1965 pp. 50-51

[24] Chapman JM, Butler RJ. *One Quine's so-called paradox*. Mind 74 1965 pp. 424-425

[25] Kiefer J, Ellison J. *The prediction paradox again*. Mind 74 1965 pp. 426-427

[26] Shönberg. *A note on the logical fallacy in the paradox of the unexpected examination*. Mind 75 1966 pp. 125-127

[27] Fraser JT. *Note relating to a paradox of the temporal order*. "The Voices of Time", JT Fraser ed New York, Barziller 1966 pp. 524-526, 679

[28] Wright JA. *The surprise exam: prediction on last day uncertain*. Mind 75 1967 pp. 115-117

[29] Cargile J. *Surprise test paradox*. Journal of Philosophy 64 1967 pp. 550-563

[30] Binkley R. *The surprise examination in modal logic*. Journal of Philosophy 65 1968 pp. 127-136

[31] Woolhouse R. *Third possibilities and the law of the excluded middle*. Mind 76 1968 pp. 283-285

[32] Harrison C. *The unanticipated examination in view of Kripke's semantics for modal logic*. "Aspects of philosophical logic: some logical forays into central notions of linguistics and philosophy" ed. Davis J, Hockney D, Wilson W pub. D Reidel Dordrecht 1969

[33] McLelland J. *Epistemic logic and the paradox of the surprise examination*. Internalional Logic Review 3 1971

[34] Woodall DR. *The paradox of the surprise examination*. Eureka 30 pp. 31-32

[35] Ayer AJ. *On a supposed antimony*. Mind 82 1973 pp. 125-126

[36] Edman M. *The Prediction Paradox*. Theoria 40 1973 pp. 166-175

[37] McClelland J, Chihara C. *The surprise examination paradox*. Journal of Philosophical Logic 4 1975 pp. 71-89

[38] Wright C, Sudbury A. *The paradox of the unexpected examination*. Australasian Journal of Philosophy 1977

[39] Kvart I. *The paradox of the surprise examination*. Logique et Analyse 1978

[40] Olin D. *The predicition paradox resolved*. Philosophical studies 1982

[41] Sorenson R. *Recalcitrant versions of the prediction paradox*. Australasian Journal of Philosophy 1982 pp. 355-362

[42] Sorenson R. *Conditional blindspots and the knowledge squeeze: a solution to the prediction paradox*. Australasian Journal of Philosophy 1984

[43] Chihara C. *Olin, Quine and the surprise examination*. Philosophical Studies 1985

[44] Kirkhan R. *The two paradoxes of the unexpected hanging*. Philosophical Studies 1986

[45] Olin D. *The prediction paradox: resolving recalcitrant variations*. Australasian Journal of Philosophy 1986

[46] Hapern J, Yoram M. *Taken by surprise: the paradox of the surprise test revisited*. Journal of Philosophical Logic 15 1986 pp. 289-304

[47] Janaway C. *Knowing about surprises: a supposed antinomy revisted*. Mind 98 1989 pp. 391-410

[48] White G. *Davidson and an Unexpected Examination*. (Unpublished) Clare Hall Cambridge August 1992

[49] Russell. *An enquiry into meaning & truth*. ch. vii

[50] Reichenback. *Analysis of conversational language*. Elements of Symbolic Logic ch. vii para. 50

[51] Floyd RW. *Assigning meanings to programs*. "Mathematical Aspects of Computer Science, Proceedings of Symposia in Applied Mathematics 19 (American Mathematical Society)" ed. Schwartz JT pub. Providence 1967 pp. 19-32

[52] Gentzen, G. *Investigations into logical deductions*. "The Collected Papers of Gerhard Gentzen" ed. Szabo ME pub. North-Holland Publishing Company 1969 chap. 3 pp. 68-129

[53] Ginsberg M. *Bilattices and modal operators*. Journal of Logic and Computation 1(1) 1990

[54] Ryan M, Sadler M. *Valuation systems and consequence relations*. "Handbook of Logic in Computer Science" ed. Abramsky S, Garbbay DM, Maibaum TSE pub. Oxford Science Publications 1992 vol. 1 pp. 1-30

[55] Winskel G, Nielsen M. *Models for Concurrency*. "Handbook of Logic in Computer Science" ed. Abramsky S, Garbbay DM, Maibaum TSE pub. Oxford Science Publications 1995 vol. 4 pp. 22-25

[56] Ong C-HL. *Correspondence between oeprational and denotational semantics: the full abstraction problem for PCF*. "Handbook of Logic in Computer Science" ed. Abramsky S, Garbbay DM, Maibaum TSE pub. Oxford Science Publications 1995 vol. 4, pp. 270-282

[57] Forster, T. *The significance of Yablo's paradox without self-reference*. (Unpublished) Dept of Pure Mathematics and Statistics Cambridge January 1996.

[58] Locke J. *Of Retention*. “An Essay Concerning Human Understanding” 1688 book II chap. X sec. 1.

[59] Locke J. *Of Knowledge in general*. “An Essay Concerning Human Understanding” 1688 book IV chap. I sec. 2.