ON PARADOXES AND A SURPRISE EXAM

RICHARD L. KIRKHAM

In ascending (steeply!) order of importance, I want to defend my solution to the surprise exam paradox, update Margalit's and Bar-Hillel's (1983) survey of work on that paradox,[1] and take issue with some commonly held views on what constitutes a good solution to a paradox.

One version of the paradox of the Surprise Exam is: A teacher tells his class "I am going to give you a surprise exam sometime next week, very possibly on that last day. I do not believe in announcing the date of examinations in advance. For this reason I shall never give you an exam does not come as a surprise. Accordingly, I am not going to tell you when I shall be administering the test, except that it will be on the first moment of class on one of the five weekdays. You will not find out until that moment that the test is scheduled for that day. You will be surprised, even if I administer it on the last weekday."

At that point a bright student says: "But you cannot give us an exam next week: If you have not given us an exam by Thursday, then we shall expect on Thursday night that the exam will be the next day; hence, we shall not be surprised when you administer it. But you have said that you will never give us an exam that does not come as a surprise, so you cannot give us the exam at all on Friday. Neither can you give us an exam on Thursday; for, believing that it cannot be on Friday, we would believe by Wednesday night that it must be given the following morning, so, if it occurs, it will not be a surprise. For similar reasons you cannot give us a surprise exam on Wednesday, Tuesday, or Monday. Therefore, since you will not give us an un-surprising exam, you cannot give us any exam all all." The first moment of Friday's class, the teacher gave his students an exam and they were surprised.

O'Connor's (1948) original presentation of the story ended after the teacher's announcement. It was O'Connor himself who made the reasoning now attributed to a student in the former's attempt to show that certain contingent future tense statements cannot possibly come true. Call this the ur-paradox. But Scriven (1951) pointed out that the announcement can come true. He was right. Note, for example, that if the students suffer amnesia, a logical possibility, immediately after the teacher's announcement, they will cease to expect an exam the following week and, thus, be surprised at getting one. More intriguingly, it is possible, and indeed probable, that the student's argument, specifically, the conclusion that the teacher cannot give an exam, will itself cause the students to cease to expect an exam the following week; thereby setting them up for a surprise. This has two implications: First, the ur-paradox is something of a fraud. Since the scenario in the modern version is logically possible, and in that version everything the teacher says comes true, the teacher's announcement is not self-contradictory or self-refuting or of any particular logical or semantic interest at all. At most, we can say that it has a kind of epistemic peculiarity and perhaps also a kind of pragmatic (in the linguists' sense) peculiarity since the teacher has no right to assume that the circumstances which would allow it to come true are going to obtain. Second, since the student's (O'Connor's) reasoning at least seems to be sound, but, in the logically possible story, has a false conclusion; there is another, genuine, paradox here. The story is a *prima facie* counterexample to our standard canons of logic. Hence, to oversimplify for a moment, resolving this geniune paradox means defending standard logic; which, in turn, means showing that, appearances to the contrary, the student's reasoning is not sound after all. So the task is to find an invalid step in his reasoning or a false premise.

One might think that philosophers would respond to this by concentrating their efforts on the genuine paradox. This did not happen. Some philosophers did not grasp either of the implications of Scriven's work. Most saw the second, albeit with blurred vision in many cases. Until well into the 1960's most, including Scriven, did not see the first at all. As a result, much of the literature, especially in the early years, displays an odd preoccupation with the ur-paradox and it mostly consists in unintelligible attempts to provide a common, integrated diagnosis and solution for two paradoxes, one of which is not in fact a paradox at all. The failure to

see the first implication, in particular, has led many to reach conclusions that I once rather delicately called "exotic". Margalit and Bar-Hillel, tougher characters, call them "mind-boggling". (For examples and citations, see Kirkham (1986, nl) , and Margalit and Bar-Hillel (1983, *passim*).)

What are the criteria for an acceptable solution to a paradox? The most commonly accepted criterion arises from the recognition that all paradoxes have variations. The Surprise Exam Paradox is no exception. The story is not always about an examination. Sometimes it is about a surprise hanging or some other event. More importantly, there is a "know" variation in which the bright student's second sentence reads "...then we shall *know* on Thursday night...etc," and the rest of the story is rewritten *mutatus mutandus*. The "justifiably believe" variation has the same clause read "...then we shall justiifiably believe on Thursday night...etc," and, again, other changes *mutatus mutandus*. Since an expectation is a belief, the version I presented above can be called the doxastic version. There are also versions which are non-temporal in a certain sense (see below). This represents another whole axis of variation, since each non-temporal version has a "know", "justifiably believe", and "expects" sub-version. The fact that there are variations raises the possibility that the inferential step or premise which a proposed solution identifies as problematic is not even present in every variation of the paradox. Thus, a commonly accepted adequacy condition is that a proposed solution must be applicable to every variation of the paradox. At the end of this paper I shall argue that this commonly accepted criterion is mistaken. But, readers may take it as a given in the meantime; for independent reasons which I make clear at the end, I am going to argue anyway that my solution to the Surprise Exam is applicable to all versions of that paradox.

Below we shall see some perfectly good ways to fill in the details of the story, the announcement, and the argument, with a non-ordinary sense of 'surprise'; but we are wise to begin with a version which uses the ordinary sense. We say S expected event e if and only if, S believed, for some period, however brief, immediately prior to e's occurrence, that e was going to occur. But not all events that are merely not expected count as surprises. If they did, then the vast majority of events we observe would come as surprises. Manifestly, they do not. Event e comes as a surprise to S if and only if, S believed, for some period, however brief, immediately prior to e's

occurrence, that e would not occur (that is, S expected not-e). If S did not expect either e or not-e, then e is neither expected nor surprising. A quick amendment: If a stranger hands you $1000 *in* the next minute, you will be surprised even though you do not have an expectation that this will not occur. You have not even thought about the matter, so you have no expectations one way or the other. But if you were to consider the possibility, you would believe that this event will not take place. You have a *non-occurrent* belief that the event will not happen. So let us insert 'occurrently or non-occurrently' after 'believed' in the definition of surprise.

Since the students were surprised, it must be that at the time of the exam they were expecting that there would be no exam. Thus, something happened to reverse the expectation they had immediately after the teacher's announcement; the expectation that there *would* be an exam the following week. The most natural hypothesis is that it was the conclusion of the bright student's argument which had this effect.[2]

Let us proceed to reconstruct the student's argument and see where, if anywhere, it goes wrong. Let *t* range over both periods of time and discrete time slices, which I'll call moments. (I sometimes use 'during *t*' as an abbreviation for 'at or during *t*'.) And let 'classtime Friday' be the name of the first moment of Friday's class, and let Z be the name of the last moment before classtime Friday. Lines 1-6b and 19 are the ultimate premises. Between them, 2, 19, and the embedded sentence in 1, express the content of the teacher's announcement.

1. We now believe there will be an exam on or before classtime Friday.
2. For all times *t*, if an exam will be held at or during *t*, then we shall believe at the last moment before *t* that there will *not* be an exam at or during *t*.
3. For all times *t*, and all times *t'* later than *t*, if no exam is given before or during *t*, then we believe at or during *t'* that no exam was given before or during *t*.
4. For all times *t*, if we believe at or during *t* that not-*p*, then we do not believe at or during *t* that *p*.
5. For all events *e*, times *t*, and sequences *S* of times, if we believe at or during *t* that *e* will occur at or during some member of *S*, then (((if we believe at or during *t* that *e* does not occur at or

during any but the last member of *S*, then we believe at or during *t* that *e* happens at or during the last member of *S*).

6a. For all epistemic agents (real or ideal) *a*, all times *t*, and all times *t'* later than *t*, if *a* believes at or during *t* that *p*, then *a* believes at or during *t'* that *p*.

6b. For all times *t* before classtime Friday, if we believe at or during *t* that *p*, then we believe that *p* at Z.

7. Therefore, If we now believe there will be an exam on or before classtime Friday, then we shall believe at Z that there will be an exam on or before classtime Friday. {UI, 6a or 6b}

8. Therefore, We shall believe at Z that there will be an exam on or before classtime Friday. {MP, 1,7}

9. Therefore, If we believe at Z that there will be an exam on or before classtime Friday, then (if we believe at Z that no exam has been given through Thursday, then we believe at Z that there will be an exam at classtime Friday). {UI, 5}

10. Therefore, If we believe at Z that no exam has been given through Thursday, then we believe at Z that there will be an exam at classtime Friday. {MP, 8,9}

11. Therefore, If there has been no exam through Thursday, then we believe at Z that no exam has been given through Thursday. {UI, 3}

12. Therefore, If there has been no exam through Thursday, then we believe at Z that there will be an exam at classtime Friday. {Hyp. Syl., 10,11}

13. Therefore, If there will be an exam at classtime Friday, then we believe at Z that there will *not* be an exam at classime Friday. {UI, 2}

14. Therefore, If we believe at Z that there will *not* be an exam at classtime Friday, then we do *not* believe at Z that there *will be* an exam at classtime Friday. {UI, 4}

15. Therefore, If there will be an exam at classtime Friday, then we do *not* believe at Z that there *will be* an exam at classtime Friday. {Hyp. Syl., 13,14}

16. Therefore, If we do believe at Z that there *will be* an exam at classtime Friday, then there will *not* be an exam at classtime Friday. {Contrapositive, 15}

17. Therefore, If there has been no exam through Thursday, then there will *not* be an exam at classtime Friday. {Hyp. Syl. 12,16}

From this point, 17, along with a definition of 'next week' yields:

18. If there has been no exam through Thursday, then there will be no exam next week.

By a series of parallel steps for each day of the week, the bright student tries to arrive at: "Therefore, if there has been no exam through Sunday, then there will be no exam at all next week." (The 'Sunday', of course, refers to the Sunday *before* the proposed exam week.) This last line, plus

19. There will be no exam through Sunday.

yields via *modus ponens* his ultimate conclusion: "There will be no exam next week." Of course, it would l o easy trick to get from 18 to the next parallel step: "If there been no exam through *Wednesday*, then there will be no exam next week." The bright student would need, as a further premise, a belief-iteration principle reading "if we believe, during *t*, that *p*, then we believe, during *t*, that we believe that *p*" (Wrig d Sudbury (1977, 46)). But any mistakes the student makes after 18 are not essential to the paradox[3] for two reasons. First, we have already reached a falsified sub-conclusic vith 17. (See the story on the first page of this paper.) Second, there is a single-day version in which the teacher tells his students that there will be a surprise exam on the following Friday. A student argues that the students will, on Thursday night, expect an exam on the next day. So, given the teacher's intention not to give an *un*surprising exam there will in fact be no exam Friday. The next Friday the students are surprised to get an exam. In this version, "Sunday" in 19 is replaced with "Thursday" and the student reaches the final conclusion immediately from 18 and *modus ponens*.

A "know" (or "justifiably believe") version can be created by substituting "know" (or "justifiably believe") for "believe" wherever the latter appears in lines 1-19. These versions give a stipullated sense to 'surprise' such that an event is a surprise whenever we do not actually know (or justifiably believe) in advance that it will happen.[4] This is not *in principle* objectionable. As long as 'surprise' is used consistently and we have a seemingly sound argument to a false conclusion, we have a paradox. The reader is invited to make the suggested substitutions for lines 1-6b and note that many of the premises which are plausible on my doxastic version become wildly implausible on the other versions. Philosophers who favor these versions handle this problem by stipulating that the students are

ideal epistemic agents. Again, this is not objectionable in principle, but we shall see below that some writers have idealized the agents in question to the point that their versions fail to describe a *prima facie* counterexample to a logic that is of any interest. Finally, note that, for my doxastic version, we need not suppose that the teacher made an announcement at all. It is logically possible that the bright student believed these premises and made this argument, even if the teacher had said nothing. We would wonder why he would. (To reassure a jittery peer who fears a surprise exam?) But lack of motivation for the protagonist is a mere literary failure. There would still be a seemingly sound argument with a false conclusion; hence, there would still be a paradox here. The fact that there is at least one "no announcement" version should have an inhibiting effect on those tempted to think that analysis of the announcement's peculiar characteristics as a public utterance is going to reveal much about the *essence* of the paradox. And this also gives us further reason to think that O'Connor's ur-paradox has nothing to do with the geniune paradox; for, on the 'no announcement' version, the whole of the ur-paradox, the whole of what was originally the "paradox" drops out completely.

My contention is that in all versions of the paradox (save those which are counter-examples to logics no one has any reason to defend anyway (see below)) the bright student assumes some version or other of a false principle I call a projection principle and that no other mistake on his part is present in all versions. A projection principle says that if some agent in some time and place believes (or knows or justifiably believes) that *p*, then some agent (possibly the same agent) in another time/place/possible world also believes (or knows or justifiably believes) that *p*. In the argument above, 6a and 6b express projection principles. I include them both so readers can see the range of strength such principles can have. (Both are *temporal* projection principles because the bright student is reasoning *now*, immediately after the teacher's announcement, about the future moment *Z*. We see below that in the non-temporal versions of the paradox, he assumes a non-temporal projection principle.) Line 6a expresses a very strong projection principle. Line 6b, which quantifies over fewer variables and is thus less general, is one of the weakest principles which would entail 7.[5] We could even imagine the student just assuming 7 itself, which, not being

quantified over at all, would be the weakest relevant projection principle.[6]

But 7, and hence anything that entails it, is false. The students are not going to believe (or know or justifiably believe) at Z that there will be an exam sometime or other that week, because something happens between "now", when the student is making his argument, and Z that will cause them to change their minds about (and lose their justification for) the proposition that there will be an exam: The student is going to reach the end of the argument he is "now" making and his conclusion is that there will not be an exam.

This is essentially the same solution as Wright and Sudbury (1977), save that they were unaware of the non-temporal versions and, thus did not sufficiently generalize their solution.

What makes the student's reasoning in the original story so initially plausible is that he does not (and *we*, in being taken in by the argument do not) index the belief (or knowledge or justification) operator to a particular time. Thus, the implicit line 1 he (and we) assume does not have "now" or any other time index in it. This line is then allowed implicitly to do the work of 8 in deriving (along with a version of 9 with an un-indexed antecedent) line 10. Thus, no explicit projection principle makes an appearance in the original story.[7]

To what extent my solution is *better* than others, if I am right in contending that the assumption of an illicit projection principle is the one and only mistake common to all versions of the paradox, is an issue I take up below. First let me try to justify my contention.

Most of the mistakes others have identified as the student's key error are not applicable to the multi-day version I presented above: None of the premises is self-referring or refers to any of the other premises.[8] And none of the premises is self-contradictory nor can a contradiction be deduced from them. The common belief that inconsistent premises is the student's *essential* error is based on the mistaken impression that he must, in all versions, assume as a premise "there will be an exam on or before Friday" (which contradicts his ultimate conclusion) or that he must assume something which entails this, like "we know that there will be an exam on or before Friday".[9] In fact, in my doxastic version, he need only assume premise 1.

None of the solutions, except mine, which *are* applicable to my multi-day version are applicable to the single day version. For

example, McLelland and Chihara (1975), working with a "know" version, say that the infamous KK principle – if we know p, then we know that we know p – is the false premise of the argument. But Wright and Sudbury (1977, 46) show that an iteration principle of this sort is not needed for the first part of the argument in which only the last day is eliminated and, hence, is not needed at all in the single day case. This is confirmed by the fact that my doxastic version has no "BB" principle through line 18.[10]

One interesting school of thought, originated by Binkley (1968) and more recently defended by Kvart (1978), Olin (1983 and 1986), and Sorensen (1988), holds that the student's argument assumes that he can believe a statement which, although logically consistent and true, cannot be believed. (Or, according to some within this camp, he assumes that he justifiably believes [or knows] a statement which, though true, he cannot justifiably believe [or know].) Sorensen calls such propositions 'blindspots'. Some examples of blindspots are "p, but I do not believe it", "p, but I am not justified in believing it", and "p, but I do not know it". None of these can be known. The first two cannot be justifiably believed, but the third can be. The last two can be believed, but the first cannot. On Sorensen's well developed account, two or more propositions which are not themselves blindspots can entail an unknowable blindspot. In such cases, the propositions are 'semi-blindspots' and cannot be simultaneously known (1988, 328f). In the multi-day "know" version the students falsely assume that they can simultaneously know both (A) There will be an exam Monday and we do not know it, or there will be an exam Tuesday and we do not know it, or...,or there will be an exam Friday and we do not know it, and (B) It is not the case that (there will be an exam Monday and we do not know it, or there will be an exam Tuesday and we do not know it, or..., or there will be an exam Thursday and we do not know it). Sorenson says

> The essential flaw in the clever student's argument is that it assumes that their initial knowledge of (A) can be combined with knowledge of the falsity of the first four disjuncts of (A). Although each of the two propositions are knowable, they are not coknowable. The pair are semi-blindspots.[356]

To see why the blindspot approach (in which camp I include Binkley, Kvart, and Olin, as well as Sorensen) is not applicable to

the single day version of the paradox, note first a point which advocates of this approach often fail to keep consistently in mind: Resolving the paradox means finding an invalid inferential step or a false premise in the student's argument; but identifying a premise as a blindspot does not in itself accomplish this, for blindspots can be true. To have even a semblance of a solution, the blindspot approach must identify a premise which *itself asserts that* the students (i) *know* a certain p, where that p is an unknowable blindspot (or know a certain p & q, where p and q are a pair of co-unknowable semi-blindspots), or (ii) justifiably believe an unjustifiable blindspot, or (iii) believe an unbelievable blindspot. Thus the blindspot analysis must suppose, as Sorensen seems to recognize in the above quotation, that the crucial lines of the student's argument are not (A) and (B) as such, but rather (A) and (B) embedded within the doxastic or epistemic operator appropriate to the version. I.e. K_i(A) and K_i(B) for the "know" version. To produce the corresponding lines in the four day version we simply drop the last disjunct from each of K_i(A) and K_i(B). A similar editing gets us the three day version, etc. This leads the blindspot school to think that, in the single day case (with Monday instead of Friday as the designated exam day), K_i(B) just drops out entirely and K_i(A) shortens to the self-contradictory "We know [or believe or justifiably believe, depending of the version] that there will be an exam Monday and we do not know [or...etc.] it."[11] But as Wright and Sudbury first pointed out (1977, 45 & 50) there is a crucial disanalogy between the multi-day and single day versions. As we move from the two day version to the single day, even the doxastic or epistemic operator in which (A) is embedded can be allowed to drop away without harming the validity of the argument. The corresponding line of the argument shortens all the way to the mere claim that there will be an exam on that particular day and that it will be surprise; that is, the premise shortens to just the *true* statement "There will be an exam Monday and we do not know [or...etc.] it." This is confirmed by the fact that, on my version, which gives, in effect, the *whole* of the single day argument, lines 1 and 2 are *not* conjoined and then embedded inside a operator. One could, of course, attribute to the student in the single day case stronger premises than he needs, but that would miss the point that there are some single day versions to which the blindspot approach is irrelevant. None of this is to deny that the students do, perhaps irrationally, believe both my lines 1 and 2. My, and Wright

and Sudbury's, point is that a statement asserting that they believe (or know or justifiably believe) these things need not itself be a premise in their argument (in the single day case) and, hence, such a statement cannot be the *common* source of unsoundness in all versions.

Indeed, it is arguable that the student in my version does not make *any* mistakes (through line 18) other than his assumption of 6a or 6b. *Ex hypothese* lines 2, 3, and 19, are true in this logically possible story. Line 1 seems reasonable since the teacher has just told the students there will be an exam. As for 4, most philosophers have thought that we cannot intelligibly attribute beliefs to a being at all if those beliefs do not conform to 4, and note that 4 bans only explicit contradictions from the belief set, it does not ban all inconsistent sets of beliefs.[12] Line 5 is more controversial; but, since the student explicitly uses this sort of disjunctive syllogism even in the original brief story, 5 seems to be true in this context (where 'we' refers to only the students) even if it is not a general truth about all humans.

There are non-temporal versions of the paradox: A teacher and student S are standing on top of city hall with a telescope through which S is looking at the south gate to the city. The teacher says "There are many possible worlds in which you (or your counterpart) are also *now* standing on city hall looking at either the north or south gate to the city. In a certain subset of those worlds, call it W, Johnson (or his counterpart) is *now* standing in the north or south gate. And, there is a non-empty subset of W, call it *w*, in all of whose members you (or your counterpart) do not *now* believe [or know, or justifiably believe, depending on the version] that Johnson is standing in one of the gates unless you (your counterpart) *now* looking at Johnson." To this S responds "By definition, in all members of W, Johnson is in one or the other gate; so in all of the worlds in W in which I (my counterpart) am not actually looking at Johnson right now, I (my counterpart) believe [know, justifiably believe] that Johnson must be in the other gate. So *w* is empty." But S's conclusion is false. In many of the worlds in W, S (or S's counterpart) is *not aware* that his world is a member of a set in which Johnson is standing in one of the gates. And in many of those worlds, S (or his counterpart) is looking at an empty gate and does not believe that Johnson is standing in either gate. But the latter set of worlds just is *w*. So S is falsely assuming that information he

has in the actual world is also possessed by him (or his counterparts) in all the worlds in W. He is trans-modally projecting his information. The assumption of an illicit projection principle is less obvious in the non-temporal versions found in the literature only because they do not have the protagonist argue *explicitly* about possible worlds. Instead he argues with conditional statements about hypothetical situations (e.g. Sorensen (1988, 317-324 & 333-335)). But I have not seen a version yet in which there is not an implicit projection of information into the hypothetical situation.[13] (Arguably, temporal projection is just a special case of trans-modal projection, since a hypothetical future is just a possible world.)

Is a projection principle present in all versions? It is usually explicitly present in the various formal or semi-formal presentations of the student's argument. When it is not, it is there implicitly. A sampling of the latter: Medlin (1964), who presents only that part of the argument in which the student imagines himself on Thursday night, has the student attribute to himself at that future time the information that there will be an exam sometime or other that week (67-69). If the whole argument were presented, that attribution would have to become a premise itself and it is, of course, a projection principle.[14] Sainsbury (1988) offers a complete argument which contains *no* explicit projection principle, but he thinks the student can validly (!) deduce $K_T p$ from Kp alone, where the 'T' is a time index referring to the morning before class on the last day in which the exam can be given (99). But the second of these involves the same unintelligible, timeless knowledge operator discussed in note 7. In the context in which Sainsbury uses it, 'Kp' is relevant and intelligible only if if means $K_S p$, where 'S' is the time the student is making his argument, and $K_T p$, will not follow from this premise without the help of a projection principle.[15] Sorensen (1988) does not ever present a detailed version of the student's argument, but it is clear that a projection principle is involved. Reread the quotation from him above. The root assumption the students are making is that they can combine "initial" knowledge with knowledge they are not going to have until after Thursday's class. They are assuming, that is, that they will *still have* the initial knowledge. This is, of course, a projection principle.

Binkley (1968, 135-6) claims, but does not show, that the student need not be interpreted as assuming, where i and j are time indices (and where D is substitutable for whatever doxastic or

epistemic operator is appropriate to the version under discussion), '$D_ip \supset D_jp$, for all j>i'; which is roughly paralled to my 6a. Rather, he can be interpreted as assuming '$D_ip \supset D_iD_jp$, for all j>i'. Binkley defends the plausibility of his principle by saying that the student has no reason to think that anything relevant to the teacher's credibility is going to change. Considering that the student is in the process of arguing for a conclusion that contradicts the teacher, I find Binkley's remark astonishing. Moreover, although there are *some* parts of the argument which can be restuctured so as to let Binkley's principle do the work of the regular projection principle, McLelland and Chihara (1975, 83-4), Wright and Sudbury (1977, 57), and I have searched and none of us can produce a version of the argument which is even seemingly valid and which does not use the regular projection principle at least once. It hardly matters to my solution if someone finds such an version: Binkley's principle is itself a projection principle. (Indeed, where D is the knowledge operator, Binkley's principle, along with the principle that knowledge implies truth, entails the regular projection principle.)

Chihara (1985), in a move endorsed by Sorensen (1988, 312 & 338), creates a multiday "know" version designed precisely to undercut the sort of solution I advocate. In addition to stipulating students who are already highly idealized epistemic agents, he stipulates that (i) the teacher says as part of his announcement that the students will not lose any knowledge during the course of the week (thus, a projection principle is true of them), and finally, Chihara stipulates, (ii) that everything the teacher says is true and the students initially *know* that it is (195). Now in this situation we cannot claim that the student's assumption of a projection principle is a false premise, since it is stipulated to be true. But something is fishy here. If it is open to Chihara to stipulate that a projection principle is true of the agents, it is equally open for me to create an otherwise identical situation save that I stipulate instead that the KK principle, which Chihara thinks is the false premise the student assumes, is true of the agents. For that matter, we could make both stipulations, in which case the student has not made *any* mistakes and we have in Quine's (1962, 5-10) terms, not just a paradox but an antinomy, a *genuine*, not merely a *prima facie*, counter-example to classical logic. (We would also have an antinomy on the single day variation of Chihara's case, since in that variation the KK principle is not used [see above].)

Or do we? Consider a situation in which an agent makes an argument for a false conclusion from premises which we all accept as true and which is valid save that in one step of the argument he makes the fallacy of affirming the consequent. Now consider a similar situation save that directly or indirectly we stipulate that the agent/situation is such that affirming the consequent is valid for this agent in this situation. By the standards of classical logic, this is not a logically possible situation; but it is logically possible given a logic that allows affirming the consequent. To create an antinomy or paradox for a given logic, we need a situation which is logically possible, by the standards of the logic, and in which a sound or seemingly sound argument to a falsified or contradictory conclusion is made.[16] So the situation just described is an antinomy for this strange logic. Whether or not there is any importance to the paradox or antinomy, or hence to a solution thereof, depends on whether or not anyone has any inclinaton to think that the logic in question is correct. Even if we did not think the logic was correct for humans, the paradox would be important if we were inclined to think that the logic is correct for a particular kind of agent/situation which itself is of *independent* interest (e.g. computers). Now recall that any logic can be formulated as a natural deduction system with no axioms or, at the other extreme, as an axiom system with a substitution rule as its only rule of inference, or as a mixed system containing both axioms and a number of rules. If we stipulate that one (or both) of the KK principle or a projection principle is true in the Surprise Exam situation, then we are in effect creating a partly axiomatized epistemic logic and the paradox or antimony is a counter-example only to this new logic, not to classical logic or even to classical logic extended by more reasonable epistemic axioms. Thus, I concede that my solution is inapplicable to Chihara's version; since by definition the student did not *mistakenly* assume a projection principle. But Chihara owes us a reason for thinking that the bizarre logic which his stipulations create is true of any agents/situations in which we have any interest. Even computers suffer loss of memory and, hence, knowledge. To put the point another way, we do not *want* to solve *this* version. Indeed, we are pleased to have counter-examples to logics we think are mistaken.

If it is true, as is commonly assumed, that proposed solutions which are applicable to all variations of a paradox are to be preferred

over those applicable to only a subset of variatons, then my solution is superior to all others yet proposed for the Surprise Exam.

But this common assumption is false. The set of all variations of all paradoxes forms a multi-dimensional array whose members are tied by non-transitive similarity relations. Some regions of this array have been given generic names; e.g. epistemic paradoxes, self-reference paradoxes. Smaller regions have been given singular species names; e.g. The Surprise Exam Paradox, The Liar Paradox. But it is ultimately just historical accident that we have drawn the regional boundaries where we have. From a strictly logical view, these boundaries are arbitrary. The paradox variations just inside the boundary of the Surprise Exam region have more features in *common* with those just outside the boundary than they do with those in the center of the region. And some regions overlap with each other: The fact that the students assume a false projection principle does not mean they do not also, in some versions, commit a mistake of vicious self-reference. So the region of the Surprise Exam overlaps with the region of self-reference paradoxes. Similarly, in the multi-day "know" variations, the argument assumes the false KK principle as well a projection principle. Thus part of the Surprise Exam region overlaps the region of epistemic paradoxes. Similarly, many of the variations of the Surprise Exam (possibly *all* multi-day variations) involve a blindspot as well as an illicit projection principle. The region of epistemic paradoxes in which an argument assumes an illicit KK principle may well be as large as the region of the Surprise Exam. The same holds for the region of self-reference paradoxes and the region involving illicit belief or knowledge of a blindspot. The latter region may well include many epistemic paradoxes such as the Lottery Paradox (Olin, 1983, 231). Whether or not any of these regions conforms to the boundaries of some preconceived class of paradoxes is irrelevant. For this reason, I do not claim that my solution is superior to all of the others. It is a good solution because it solves a lot of paradox variations, but the fact that the region of variations it solves conforms with the boundaries of the Surprise Exam region is just a coincidence which does not in itself contribute to making my solution a good one. I have argued that it solves all variations only because this is a useful way of showing that it solves a big chunk of the array of paradoxes.

Accordingly, philosophers ought to cease to argue about whether this or that paradox is or is not a member of this or that species of

paradox (e.g. Sorensen (1988, 327 & 340f)). And they should stop worrying about whether this or that solution is applicable to all the members of some preconceived species or genus of paradox.[17]

Why would it have ever seemed plausible to think that a good solution must solve every member of some generic or specific class? Perhaps it is because, from a global perspective, the discovery of some one common error in all the mathematical and logical paradoxes is an ideal we aim at. Russell (1908) thought they all violated what he called the Vicious Circle Principle: "Whatever involves all of a collection must not be one of the collection. [63]" But one might wonder if any philosopical work is accomplished by such an ambiguous principle, and it is not obvious that his principle is relevant to all the paradoxes discovered since he formulated it. At any rate, it might seem a natural inference from the fact that it is desirable that the whole zoo of paradoxes eventually be tamed by some one solution to the conclusion that, in the interim, we should seek solutions species-by-species and, later, genus-by-genus. But this does not follow. Even our *interim* goal is to cover the *whole* of the paradoxical floor one way or another. Russell wants to do it with a wall-to-wall carpet, but his carpet is so thin as to be virtually transparent. Sorensen and most others want to draw boundaries around regions of the floor and find throw rugs which exactly conform to the boundaries of one or another region. My suggestion is that we just start tossing throw rugs around until the whole floor is covered. There will be time enough later to discover which are completely overlapped by others and can be safely removed.

UNIVERSITY OF OKLAHOMA
NORMAN, OKLAHOMA 73019
USA

BIBLIOGRAPHY

Ayer, A.J. (1973) "On a Supposed Antnomy," *Mind* 82:125-6.

Binkley R. (1968) "The Surprise Examination in Modal Logic," *Journal of Philosophy* 65:127-136.

Bunch, B. (1982) *Mathematical Fallacies and Paradoxes* New York: Van Nostrand.

Cargile, J. (1979) *Paradoxes* Cambridge: Cambridge University Press.

Chapman, J.M. & Butler, R.J. (1965) "On Quine's "So-called Paradox'." *Mind* 74:424-425.

Cherniak, C. (1986) *Minimal Rationality* Cambridge, Mass.: MIT Press.

Chihara, C.S. (1985) "Olin, Quine, and the Surprise Examination," *Philosophical Studies* 47:19-26.

Guiasu, S. (1987) "Prediction Paradox Revisited," *Logique et Analyse* 30:147-154.

Halpern, J.Y. & Moses Y. (1986) "Taken By Surprise: The Paradox of the Surprise Test Revisited," *Journal of Philosophical Logic* 15:281-304.

Holtzman, J.M. (1987) "An Undecidable Aspect of the Unexpected Hanging Problem," *Philosophia* 17:195-198.

Hughes, P. & G. Brecht (1975) *Vicious Circles and Infinity: An Anthology of Paradoxes* Harmondsworth: Penguin.

Kirkham, R.L. (1986) "The Two Paradoxes of the Unexpected Examination," *Philosophical Studies* 49:19-26.

Kvart, I. (1978) "The Paradox of Surprise Examination," *Logique et Analyse* 21:337-344.

Lyon, A. (1959) "The Prediction Paradox," *Mind* 68:510-517.

Margalit A. & M. Bar-Hillel (1983) "Expecting the Unexpected," *Philosophia* 13:263-288.

McLelland, J. (1971) "Epistemic Logic and the Paradox of the Surprise Examination," *International Logic Review* 3:69-85.

McLelland J. and C. Chihara (1975) "The Surprise Examination Paradox," *Journal of Philosophical Logic* 4:71-89.

Medlin, B. (1964) "The Unexpected Examination," *American Philosophical Quarterly* 1:66-72.

O'Carroll, M.J. (1967) "Improper Self-Reference in Classical Logic and the Prediction Paradox," *Logique et Analyse* 10:167-172.

O'Conner, D.J. (1948) "Pragmatic Paradoxes," *Mind* 57:358-359.

Olin, D. (1982) "The Prediction Paradox Resolved," *Philosophical Studies* 44:225-233.

------ (1986) "The Prediction Paradox: Resolving Recalcitrant Variations," *Australasian Journal of Philosophy* 64:181-189.

------ (1988) "Predictions, Intentions and the Prisoner's Dilemma," *Philosophical Quarterly* 38:111-116.

Quine, W.V.O. (1962) "The Ways of Paradox," in Quine's (1976) *The Ways of Paradox and Other Essays* Rev. Ed., Cambridge, Mass.: Harvard University Press, 1-18.

Russell, B. (1908) "Mathematical Logic as Based on a Theory of Types," in Russell's (1956) *Logic and Knowledge* New York: G.P. Putnam's Sons, 59-102.

Sainsbury, R.M. (1988) *Paradoxes* Cambridge: Cambridge University Press.

Schoenberg, J. (1966) "A Note on the Logical Fallacy in the Paradox of the Unexpected Examination," *Mind* 75:125-127.

Scriven, M. (1951) "Paradoxical Announcements," *Mind* 60:403-407.

Sharpe, R.A. (1965) "The Unexpected Examination," *Mind* 74:255.

Shaw, R. (1958) "The Unexpected Examination," *Mind* 67:382-384.

Smith, J.W. (1984) "The Surprise Examination and the Paradox of the Heap," *Philosophical Papers* 13:43-56..

Sorensen, R.A. (1982) "Recalcitrant Variations of the Prediction Paradox," *Australasian Journal of Philosphy* 60:355-362.

------ (1983) "Conditional Bindspots and the Knowledge Squeeze: A Solution to the Prediction Paradox," *Australasian Journal of Philosophy* 62:126-135.

—— (1986a) "The Bottle Imp and the Prediction Paradox," *Philosophia* 15:421-424.

------ (1986b) "A Strengthened Prediction Paradox," *Philosophical Quarterly* 36:504-513.

—— (1986c) "Blindspotting and Choice Variations of the Prediction Paradox," *American Philosophical Quarterly* 36:337-352.

—— (1987) "The Bottle Imp and the Prediction Paradox, II," *Philosophia* 17:351-354.

—— (1988) *Blindspots* Oxford: Clarendon Press.

Wreem. M.J. (1983) "Surprising the Examiner," *Logique et Analyse* 26:177-190.

------ (1986) "Passing the Bottle," *Philosophia* 15:427-444.

Wright, C. & A. Sudbury (1977) "The Paradox of the Unexpected Examination," *Australasian Journal of Philosophy* 55:41-58.

Wright, J.A. (1967) "The Surprise Exam: Prediction on the Last Day Uncertain," *Mind* 76:115-117.

NOTES

1 Another history of the paradox is chapter 7 of Sorensen (1988). I shall also update Margalit and Bar-Hillel's bibliography by including works I do not cite in the text and even a few to which I do not have access. Between them, my bibliography and Margalit and Bar-Hillel's include every relevant work in English (published in a philosophical journal) of which I know. Sorensen (1988) refers in his text to relevant papers by Craig Harrison and Paul Dietl, but he left out the bibliographical data from his notes and bibliography.

2 Strictly speaking, my solution does not depend on either the announcement or the bright student's argument having any affect whatever on the students' expectations. It depends only on something which we are given in the story anyway; that premise 7 in the argument below is false. But I speak mostly as if the announcement convinced the students that there would be an exam and as if the bright student's argument reversed that expectation; because this is the most natural interpretation to put on the story.

3 *Contra* Shaw (1958); Lyon (1959); Sharpe (1965); Schoenberg (1966); and Wright (1967).

4 Thus, in premise 2, besides making the substitution just described, these versions would move the "not" so that it modifies "know" (or "justifiably believe"). A side effect of this is that there is no need for lines 4, 14, and 15, when the changes are completed.

5 Other weak principles would also entail 7. E.g. A principle just like 6b save that it quantifies over agents, instead of times-before-classtime-Friday.

6 One of the ways 7 differs from 6a is that 7 does not suppose that the belief in question is continuously held throughout the period from "now" through Z. It allows that the students may have temporarily lost the belief in between (and regained it at Z). A little reflection shows that the students do no really have to have the belief at "now" at all. So long as they have it at Z, the argument from 8 onward succeeds. So 8 itself is in one sense the weakest assumption the student could make, and were he to just assume it he could do without any projection principle. But as everyone, save possibly Medlin (1964), has realized 8 is too strong in another sense: If it is left

without any justification, then we do not even have the appearance of a sound argument here; and, hence, no paradox.

[7] Accordingly, I would not be adverse to the suggestion that my real solution to the paradox is that the students use an unintelligible timeless doxastic or epistemic operator. This is just verbally different from saying they use an intelligible time-indexed operator in conjunction with a projection principle.

[8] *Contra* Shaw (1958) and Halpern and Moses (1986). See Margalit and Bar-Hillel (1983) and Chapter 7 of Sorensen (1988) for references to the many self-reference approaches published between these two papers.

[9] See, for example, Chapman and Butler (1965); McLelland and Chihara (1975); and Schoenberg (1966).

[10] Wright and Subdury (1977, 52) also show that an iteration principle is arguably *true* if the operator, call it D, is not "know", "justifiably believe", or "believe"; but is, instead, "would be justified in believing, if he does believe". The thought is that if the agent's evidence for p is enough to mak $D_i p$ true, then that same evidence makes $D_i D_i p$, $D_i D_i D_i p$, ... etc., true. But a projection principle is not true even for the D operator (53).

[11] Binkley (1968, 130); Kvart (1978, 344): Olin (1983. 230); and Sorenson (1988, 331-2).

[12] See Cherniak (1986) for a discussion of this issue and references to other works.

[13] The fact that these conditionals are usually expressed in indicative mood (e.g. "...if I am in..." in Sorensen's grid game paradox (1988, 321)) instead of present tense subjunctive mood (...if I were in...) further disguises the fact that the arguer in these non-temporal versions is reasoning about hypothetical situations. In common parlance this is acceptable, since the subjunctive has become moribund in ordinary English. But philosophers ought be more precise. Indicative mood is appropriate only when the speaker believes the antecedent of the conditional actually obtains. (And in the case where he does, even to use the word "if" is misleading. When the speaker believes that p, he ought to say "since p, q" not "if p, then q.)

[14] So I was dead wrong to say in (Kirkham 1986) that Medlin has a version which does not make the mistake I have identified. Since my version commits no error of self-reference, I concluded *then* that there is no one mistake common to all versions and, thus, the surprise exam story hides two distinctly different paradoxes within it. I now renounce these claims.

[15] Similarly, line 4 of Sainbury's next version (100-101) is a *non-sequitur* without a projection principle.

[16] An exception: A para-consistent logic allows contradictions and statements which are both true and false. So a situation in which such a conclusion is reached from true premises by a para-consistent logic would not constiitute a paradox or antinomy for that logic. Instead we would need a situation which is logically possible by the standards of para-consistent logic and in which a conclusion is reached which even advocates of a para-consistent logic would regard as alarming. The statement "anything can be derived from a contradiction" would be such a conclusion, since advocates of para-consistent logic regard this as alarming and claim that it is not true on their logic.

[17] Thus, Holtzman (1987) is not principle doing anything wrong by acknowledging right from the start that his solution is not applicable to the single day case (196).