CHARLES S. CHIHARA

OLIN, QUINE, AND THE SURPRISE EXAMINATION

(Received 15 March, 1984)

The instructor announces that he will give one (and only one) more examination on one of the following n class days (at the usual class time) and that the examination will 'surprise' the students in the following sense: they will not know, before the class meets on the day of the test, that the test is going to be given that day. Student S concludes that the examination cannot be given on the nth class day since if it were, he would know on the day of the examination (before class) that the examination had not been given on any of the previous $n-1$ class days and that there was only one day left on which the examination could be given, so he would then know that it was to be given on that day (contrary to what was announced). S then concludes that the examination cannot be given on the $n-1$st class day either since if it were, he would know on that day before class that no examination had been given on the previous $n-2$ class days and that there was only one day left on which the examination could be given, so he would then know that it was to be given on that day (contrary to what was announced). S continues in this way, concluding that the announcement of the instructor cannot be true. Evidently, we can construct a *reductio ad absurdum* of the assumption that the students know that what was announced is true. This is a paradoxical result since we seem to have very strong intuitions that the students could very well know just that.

This paper analyzes a 'solution' to this well-known paradox recently put forward by Doris Olin[1] and compares Olin's resolution with one I favor and also with Quine's. My analysis will be given within the framework of a propositional epistemic logic described in an earlier paper I co-authored with James McLelland.[2] The language of this logic is obtained by taking just the sentential letters

$$p_1, p_2, \ldots, p_n$$

as atomic sentences. Thus, interpreting 'p_i' to mean

Philosophical Studies **47** (1985) 191–199. 0031–8116/85.10

The examination is given on day i.

the teacher's announcement that there will be an examination on one of the following n class days can be expressed by the disjunction

(1) $\quad p_1 \vee p_2 \vee \ldots \vee p_n$

and letting 'T_j' stand for

$$p_1 \vee p_2 \vee \ldots \vee p_j$$

(for any j), the above announcement can be shortened to 'T_n'.

The language also contains the operators

$$K_1, K_2, \ldots, K_n$$

where '$K_i p$' is to be interpreted as saying

> It is known by student S on class day i (before the usual class time) that p.

Then, the teacher's announcement that the students (in particular S) will not know, on the day of the examination before that class, that the examination is to be given on that day can be expressed by the conjunction

(2) $\quad \bigwedge_{i=1}^{n} (p_i \rightarrow -K_i p_i)$

It is clear that the reasoning of the paradox presupposes more than just the truth of the announcements. For example, it is assumed that S would be able to identify and keep track of the relevant class days and would know if an examination had already been given on any of the previous class days.[3] So we can also assume as given

(3) $\quad \bigwedge_{i=1}^{n} (-T_i \rightarrow K_{i+1} -T_i).$

Now the paradox seems to show the absurdity of assuming that the students (and in particular S) know (1), (2), and (3). Thus, the reasoning seems to show the absurdity of the *initial premise*

$$K_1(1) \,\&\, K_1(2) \,\&\, K_1(3).$$

The system of Epistemic Logic I use here has, in addition to the usual proposi-

tional calculus, the axioms:

For all i,

(K1) $K_i\phi \rightarrow \phi$

(K2) $K_i(\phi \rightarrow \psi) \rightarrow (K_i\phi \rightarrow K_i\psi)$

and the rule of inference

(The K rule) Infer $K_i\phi$ from ϕ.

A *theorem* is a formula that can be obtained from the axioms using the K rule and *modus ponens*. A formula ϕ is *deducible from* a set Γ of formulas if ϕ can be obtained from Γ using theorems of the system and *modus ponens*.

These principles cannot be considered to be self-evident truths or even accurate representations of what ordinary mortals know, but in our earlier paper, we gave a detailed justification for using this system in our analysis of the paradox;[4] so I shall not repeat the justification here. I would like to note simply that the use of these principles in this context involves a strong idealization, but this idealization is one that is useful and, we argued, logically harmless.

Now it can be shown that if one accepts the additional principles

(K3) $K_i\phi \rightarrow K_iK_i\phi$ for all i

(K4) $K_i\phi \rightarrow K_j\phi$ for $i < j$

one can derive a contradiction from the initial premise of the paradox

$$K_1(1) \,\&\, K_1(2) \,\&\, K_1(3).$$

(K3) is the well-known "*KK* axiom" of Hintikka's epistemic logic, and (K4) is the *temporal retention principle*, which says (roughly) that no relevant knowledge will be lost during the period in question. These principles are discussed in detail in our previous paper,[5] so I shall be brief here. These principles are not being singled out as ones that ought to be accepted. I only note that they have seemed plausible to some philosophers and that one can deduce a contradiction from the initial premise if one allows their use. These facts suggest that we may be able to classify suggested 'solutions' to the paradox by determining which of these principles are singled out for rejection. Clearly, any solution must provide some way of obviating the derivation of the contradiction, so at least one of these principles, it would seem, must be rejected by any plausible 'solution'.

Consider now Olin's resolution of the paradox. How does Olin propose to block the derivation of the contradiction? I believe that she, in effect, rejects (K4). She does not deny that S knows (1), (2) and (3), at least on the first day of class. But she does attack S's argument that the examination cannot be given on the last day, by rejecting a key assumption of the reasoning: one cannot assume, she in effect argues, that the student will know (1) and (2) on class day n.

Thus, although Olin never explicitly rejects the temporal retention principle nor states specifically that the knowledge S has on day 1 would be lost on class day n if no examination were given up to then, she does claim that S would not be justified in believing (1) and (2) on day n if we take into account the student's "total available evidence".[6] Her own diagnosis of the erroneous reasoning of the paradox is that the following principle is incorrect:

(P5) If A is justified in believing ϕ, ϕ strongly confirms ψ A sees this and has no other evidence relevant to ψ, then A is justified in believing ψ.

But her argument that it is (P5) that should be rejected is, itself, quite questionable; for the falsity of (P5) is supposed to follow from:

(*) Even though on the nth class day the student would have good evidence for (1) and (2), he would not be justified in believing (1) and (2).[7]

Let us grant Olin (*). Does this show that (P5) is false? Well, on the nth class day, S would still have good evidence for the hypothesis that the instructor has decreed (1) and (2); and this hypothesis strongly confirms (1) and (2). Let us also grant that S would not, on this day, be justified in believing (1) and (2). Do we now have a counter-example to (P5)? Not at all. For to have such an example, we would also need to satisfy the condition that *the student has no other evidence relevant to the truth of* (1) *and* (2); and this condition is surely *not* met in the example under consideration. In the example, the student knows that no examination has been given on the previous $n - 1$ class days, and this is surely relevant to the truth of (1) and (2). So I question her argument that (P5) has been shown to be false. As I see it, what she has done is to provide a convincing case for the position that even if S is justified in believing (1), (2) and (3) on the first class day, it does not follow that S would be justified in believing these things on class day n – especially if S

were to obtain new information relevant to the truth of the propositions during the intervening days. This is why I regard her position as, in effect, the rejection of (K4).

Much of Olin's paper is given over to supporting her rejection of the key assumption that S would be justified on class day n in believing (1) and (2). On this point, as I indicated above, I have no quarrel with her. And I agree that she is able to undermine, by this device, the reasoning of at least one form of the paradox. But I do question her suggestion that she has thereby completely resolved the paradox. For I believe that there is more to the puzzle than she allows. To support this suggestion, I shall attempt to resurrect the paradox – but in a form that is not treatable in the way Olin advocates. To do this, we need to imagine that the instructor also guarantees that, in addition to what was announced earlier, S will not lose any knowledge pertaining to the examination before the examination is given. We can imagine that the instructor is absolutely reliable in these matters and that there is more than ample evidence for trusting him regarding his announcements. In this situation, it would be reasonable to attribute a knowledge of (K4) to S, at least on class day 1, for we can stipulate that, in the imagined situation, the instructor's pronouncements are true and S believes they are true. Thus, in this situation, we have as our initial premise

$$K_1(1) \,\&\, K_1(2) \,\&\, K_1(3) \,\&\, K_1(4)$$

and it is no longer open to Olin to reject (K4).

To see more clearly that there are indeed situations in which one can have the sort of knowledge being attributed to S, consider the following *card situation*: It is verified by all that the dealer has a standard deck of playing cards. He then arranges the cards and places the deck face down on the table. He announces that he will turn up, one at a time from the top, each card in the deck until the Jack of Spades is turned up and that we will be 'surprised', i.e., we will not know before that card is turned up that it will be the Jack of Spades.[8] He also guarantees that we will not lose any knowledge relevant to this situation during this sequence of events (the turning up of the cards). Again, we can suppose that the dealer is absolutely reliable, etc. Now imagine that he turns up the Jack of Spades on the 32nd card-turning and that we are all 'surprised' by this. Would we not have a situation in which the initial premise holds? And would it not be implausible to suppose that we just couldn't know (K4) in such a situation and that our knowledge of (1) and (2)

would have to be lost? And yet we seem to be able to produce a convincing argument – the paradox – that the initial premise is unsatisfiable!

Let us return to the examination situation and reexamine Olin's solution. Her strategy is to block the very first step of the reasoning: even the nth class day cannot be ruled out, she suggests, since (K4) is questionable. But in the present version, given the above initial premise, we can indeed deduce T_{n-1},[9] so the examination cannot be given on the last day. Now if we can deduce T_{n-1} from our initial premise, so can S (who can be assumed to be a reasonable logician): he too can conclude, it would seem, that the examination cannot be given on the last day. So he too would be in a position to know T_{n-1}. We would thus have $K_1 T_{n-1}$, and by (K4), we could infer $K_{n-1} T_{n-1}$. Now it can also be proved that

$$K_{n-i} T_{n-i}, (2), (3) \vdash T_{n-(i+1)}; \; i = 0, 1, ..., n-1.^{10}$$

So it would seem that by repeating the above line of reasoning, we can obtain

$$T_{n-2}, T_{n-3}, ..., T_2, T_1$$

and thus, evidently, an absurdity – an absurdity that Olin's analysis is powerless to block. I should add that I have propounded this version of the paradox to graduate students of philosophy, logic, and mathematics, and without exception they have found it puzzling, so there is some (crude) confirmation of the hypothesis that the paradox is not completely laid to rest by Olin's solution.

It is interesting to compare Quine's well-known resolution of the paradox with Olin's.[11] Quine too rejects the very first step of the reasoning. He too denies that we can rule out the very last day, for he denies (as did Olin) that we can affirm that S would know (1) on the last day.[12] But Quine's reason for rejecting the acceptance of $K_n T_n$ does not involve the rejection of the temporal retention principle. In the version of the paradox Quine considers, we are not even granted $K_1(1) \,\&\, K_1(2)$, i.e., it is not even stipulated that the student knows that the announcements are true. All we are allowed to assume in Quine's version is that the student knows that the instructor has decreed (1) and (2).[13] It is no wonder that Quine finds the reasoning of the paradox less than persuasive: from such weak initial premises, there is no way one can plausibly deduce a contradiction. But it is also no wonder that so many philosophers have continued to work on the paradox; for Quine's 'solution' works only for the weakest of versions.

Well, what then is wrong with the reasoning of the paradox? We have seen that this strong version cannot be blocked in the way advocated by Quine and Olin. So what step in the reasoning should be questioned? As I see it, the faulty inference is the one from T_{n-1} to $K_1 T_{n-1}$. We can indeed make the deduction of T_{n-1} from the initial premise, and we believe that student S can deduce what we can deduce and, in this way, come to know what we know. Since it appears that we know T_{n-1}, it seems reasonable to conclude that S would know T_{n-1} too. But that should be rejected. For our knowing T_{n-1} comes down to our knowing that T_{n-1} holds in the imagined situation; and we know this because we know that the initial premise holds in the imagined situation – after all, the situation was stipulated to be one in which the initial premise obtains. But S does not know that the situation he is in has been stipulated to be one in which the initial premise holds.

To see this clearly, suppose it is asked: 'How do you know that the initial premise is true?' It is natural to respond, 'It was given' or 'Didn't you stipulate that the situation was to be one in which the initial premise held?' And to the question, 'How do you know T_{n-1} holds in this situation?' we are apt to respond, 'We deduced T_{n-1} from what we were given.' But this indicates that our supposed knowledge of T_{n-1} can be analyzed to be knowledge of a hypothetical: as a result of our deduction we know that *in any situation in which the initial premise holds, T_{n-1} holds also*. Now S can arrive at this knowledge too; but this knowledge, i.e., the knowledge that S would have, would not amount to his knowing T_{n-1}. For S cannot obtain more knowledge from his deduction than we can from ours. In other words for S to know T_{n-1}, he would have to know that the initial premise was true. It is not enough that he be in a situation in which the initial premise holds, *he would also have to know that he is in such a situation*. And this is not given by the initial premise.

To amplify the basic point a bit further, recall that we can respond to the question 'How do you know the initial premise is true?' by saying 'It was given' or 'We stipulated that the situation was to be one in which the initial premise is true.' But the student in the situation cannot so respond. He cannot say, 'I know that the initial premise is true because I know that you have stipulated this to be such a situation.' That would be absurd!

The confusion detailed above can be made more difficult to avoid if, as is so often done, we imagine during the reasoning that we, ourselves, are the students in the situation described. We would then imagine that we are giving

the argument of the paradox. Then when we infer T_{n-1} from the initial premise, it is easy to imagine that we, the students, would know T_{n-1}. It would then be difficult to notice the slipping back and forth between the viewpoint of the theorizer and that of the student.

There are other ways in which one might be misled into thinking that S would also know T_{n-1} as a result of the kind of deduction given above. For example, if one accepted some form to the *KK principle*, then one could use the principle to make the deduction of $K_1 T_{n-1}$. Thus, it is possible that some would find the reasoning of the paradox irresistable because of the acceptance of an invalid principle. Such a possibility (as well as other possible confusions) is discussed at some length in our earlier article.[14] But I believe that the central fallacy of the paradox is the one presented above.

I should like now to take up a type of response to my treatment of the paradox that I have gotten from several people. Why, it has been asked, could one not simply strengthen the initial premise of the paradox even further to fill the gap in the reasoning that I have been stressing? Suppose this time that student S knows not only the initial premise of the previous version, but even *that he knows this*. In other words, take as a new initial premise

$$K_1(K_1(1) \& K_1(2) \& K_1(3) \& K_1(4)).$$

Now, in this situation, one can make use of the added power to infer, from T_{n-1}, $K_1 T_{n-1}$ and then $K_{n-1} T_{n-1}$. So we can indeed arrive at T_{n-2}. But in this new situation, the 'gap' of the previous situation does not disappear: it merely moves over one step. It now appears in the step from T_{n-2} to $K_1 T_{n-2}$. So even this new initial premise is satisfiable. Well, in that case, why not strengthen the initial premise still further to

(**) $\quad K_1 K_1 \ldots K_1(K_1(1) \& K_1(2) \& K_1(3) \& K_1(\mathrm{K4}))$

(where the number of 'K_1's appearing to the left of the left parenthesis in $n-1$)? In this case, we can derive a contradiction and conclude that this premise is unsatisfiable. But is that really shocking? After all, (**) is an extraordinary attribution of knowledge – it is not at all the sort of thing that 'the man-on-the-street' would be likely to ever say. And how clear and trustworthy are our intuitions about such attributions? In the version I considered earlier, the student was hypothesized as knowing certain things that we believe a person could know: these are things that even non-philosophers would assume they know as a result of being given the information described in the paradox. But I do not find (**) comparable: I certainly do not regard it as

obviously satisfiable. Indeed, I find it difficult to conceive of a situation in which it would be natural to describe someone as having the kind of knowledge needed by (**). What is clear is that the implications of having this kind of knowledge are far-reaching and difficult for most people to anticipate. The unsatisfiability of (**) seems to me to be of a piece with the unsatisfiability of various complicated self-referential sentences – perhaps surprising to some people at first but, I believe, not so puzzling when the implications of the sentences are clearly laid out.

In summary, if one wishes to regard (**) as giving rise to a paradox, alongside the one described earlier, then my diagnosis of these paradoxes comes to this: the earlier one rests on fallacious reasoning, and its initial premise is satisfiable, as we all thought from the start; whereas (**) is simply unsatisfiable, contrary to what some may have thought.[15]

NOTES

[1] Doris Olin, 'The prediction paradox resolved', *Philosophical Studies* 44, pp. 225–233.
[2] J. McLelland and C. Chihara, 'The surprise examination paradox', *Journal of Philosophical Logic* 4, pp. 71–89.
[3] See pp. 72–72, *ibid.*, for details.
[4] See especially p. 76, *ibid.*
[5] See pp. 80–82, 84–85, *ibid.*
[6] Olin, *op. cit.*, p. 228.
[7] *Ibid.*, p. 229.
[8] See McLelland and Chihara, *op. cit.*, p. 74, for another formulation of the surprise card paradox.
[9] This can be seen from Lemma 3 on p. 79, *ibid.* However, it should be pointed out that there is a misprint in the statement of Lemma 3: the subscript of the first occurrence of 'T' should be '$n - i$' instead of '$n - 1$'.
[10] This is just Lemma 3.
[11] W. V. Quine, 'On a supposed antinomy', in his *The Ways of Paradox*, Random House, New York, 1966, pp. 21–23.
[12] See p. 23, *ibid.*
[13] *Ibid.*, p. 22.
[14] See McLelland and Chihara, *op. cit.*, pp. 83–86.
[15] The stimulation for this paper came from a discussion I had with Doris Olin at the *Seventeenth World Congress of Philosophy* held in Montreal in August of 1983, where she gave an oral presentation of her paradox paper. Her paper was also discussed in a seminar I gave in Berkeley this fall. I am grateful to Ms. Olin, the students in the seminar, and George Myro, for their helpful criticisms.

Department of Philosophy,
University of California,
Berkeley, CA 94720,
U.S.A.