THE PARADOX OF SURPRISE EXAMINATION

Igal KVART

I. In this paper an attempt will be made to resolve the above paradox whose formulation (for the case of two days) is the following: [1]

A teacher announced to his class one day that an exam will be given in one of the following two days at noon, and that it will be a surprise. By 'surprise' he meant that a rational student could not correctly tell, on the evening before, that the exam would be given on the next day. A student, however, reasoned as follows: If the exam were to be given on the last day, then clearly he could tell the evening before that it would be given on the next day, since it was not given on the previous day — hence it would not be a surprise. Hence there cannot be a surprise exam on the last day. Hence, it must be given on the first day, the only one left; hence he can rationally believe that on the evening before, and hence it would not be a surprise either. Hence a surprise exam could not be given to a rational student.

Clearly, we feel intuitively that a surprise exam can be given, and hence a paradox.

In order to resolve the paradox, we shall first prove that the belief that there will be a surprise exam in the next two days could not be held by a rational student. The limitation to the two-days case is for the sake of simplicity and clarity. The argument can be naturally extended to many-days cases.

[1] For references and surveys of previous literature Cf. the reviews of Jonathan Bennett and James Cargile, *The Journal of Symbolic Logic,* 1965, Vol. 30, pp. 101-103. See also J. Cargile's paper in the *Journal of Philosophy* (1967), Vol. 64, "Surprise Test Paradox", pp. 550-563.

II. The following is a logical truth:

Thesis:

The statement that the student believes that there will be a surprise exam in the next two days is inconsistent with certain very plausible assumptions on his rationality (that will be listed below). Hence, if the student is rational (according to those criteria) he does *not* believe that there will be a surprise exam in the next two days.

We shall now prove this thesis. First, some notation: let us indicate the second of the two days on which the exam could be given by '*n*' (to facilitate generalization to n-days cases), the day before by 'n — 1' and so on (here we shall deal with two days only, though). '$\overline{V}$' will be the sign for the exclusive disjunction connective.

E_i — the exam is given on day i.

B^ip — the student believes on the evening of day i that p.

An exam given on day i would be a *surprise* if the (rational) student does not believe on the evening of day i — 1 that it would be given on day i.

The formalization of the statement that the student believes (prior to the two days) that there will be a surprise exam in the next two days:

$$I — B^{n-2}[(E_n \overline{V} E_{n-1}) \ \& \ ((E_{n-1} \rightarrow \sim B^{n-2}E_{n-1}) \ \& \ (E_n \rightarrow \sim B^{n-1}E_n))]$$

i.e., the student believes on the day prior to those two days that there will be an exam on one of them, and that he would not believe on the day prior to the day on which the exam would actually be given that it would be given on the following day.

Now the assumptions on the rationality of the student that we shall make are the following:

A $B^{n-2}[B^{n-1}(E_n \overline{V} E_{n-1})]$

B $B^{n-2}[\sim E_{n-1} \rightarrow B^{n-1} \sim E_{n-1}]$

C $(B^ip \ \& \ p \rightarrow q) \rightarrow B^iq$

D $B^{n-2} \sim B^{n-2}p \rightarrow \sim B^{n-2}p$

E $B^ip_1 \ \& \ B^ip_2 \ \& \ldots \& \ B^ip_n \equiv B^i(p_1 \ \& \ldots \& \ p_n)$.

Let us comment on these assumptions. Of them only *A* will be related to statement *I*; that is, if *I* is true for a rational believer, A must be true for him too. The rest of the assumptions hold for a rational believer irrespective of whether he believes *I* or not.

A states that the student will believe prior to the two days that he will believe after the first day that there is an exam in one and only one of those two days. Now from his belief of I it follows that $E_n \overline{\vee} E_{n-1}$; thus, A states that he will still believe a day later a consequence of his belief in day n — 2. Thus this amounts to elementary confidence in memory and stability in belief.

Assumption *B* states that he believes on day n — 2 that if there will be no exam on day n — 1, he will believe it on the evening of that day — thus ,that he should be able to remember on the evening of that day that he had no exam if indeed he had not.

Assumption *C* states that on day i the student believes the logical consequences of his beliefs, which is to assume the deductive closure of his beliefs. The use of *C* in the proof shows how much in terms of logical consequences the student is expected *de facto*, for the purposes of our proof, to foresee, and I believe it is quite a modes extent. Thus, in principle, *C* could be limited to exactly that extent — that the student should be able to follow only such-and-such consequences of his beliefs.

Assumption *D* states that the student can tell correctly that he does not believe a certain statement if he does not: if he believes that he does not believe that p — then indeed he does not. Assumption *E* states that if the student believes a (finite) series of statements, he must believe their conjunction, and vice versa.

All these assumptions, given *I*, seem most natural and elementary assumptions concerning rationality of belief. They are certainly empirically true for a large number of people were they to be put in the above situation (at least when *C* is restricted to the scope of its actual use in the proof below). Conditions *B* and *D* involve elementary requirements on memory and correct identification of one's own beliefs which, in the way they are used in the proof, are very easy to meet, and thus (perhaps somewhat technically) can be also grouped under the leading 'rationality conditions'.

III. We can now reformulate the Thesis above and prove it:

Reformulated Thesis:

The statement that the student believes (on the day prior to the two days period) that there will be a surprise exam on exactly one of the following two days (formalized as *I*) is logically inconsistent with the conditions on rationality described in premises *A* to *E*.

Proof:

1 $B^{n-2}[E_n \mathrm{V} E_{n-1}]$ (from *I*, *E*)

2 $B^{n-2}[E_n \rightarrow \sim E_{n-1}]$ (from *1*, *C*)

3 $B^{n-2}[E_n \rightarrow B^{n-1} \sim E_{n-1}]$ (*B*, *2*, *E* and *C*)

4 $B^{n-1}E_{n-1}$ & $B^{n-1}(E_n \overline{\mathrm{V}} E_{n-1})$ $B^{n-1}E_n$ (by *E*, *C*)

5 $[(E_n \rightarrow B^{n-1} \sim E_{n-1})$ & $B^{n-1}(E_n \overline{\mathrm{V}} E_{n-1})] \rightarrow (E_n \rightarrow B^{n-1}E_n)$
(from *4* by: $[(p \,\&\, q) \rightarrow r] \rightarrow [[(s \rightarrow p) \,\&\, q] \rightarrow (s \rightarrow r)]$ —
a tautology).

6 $B^{n-2}(E_n \rightarrow B^{n-1}E_n)$ (*3*, *A*, *E*, *C*)

7 $B^{n-2}[(E_n \,\&\, \sim B^{n-1}E_n) \overline{\mathrm{V}} (E_{n-1} \,\&\, \sim B^{n-2}E_{n-1})$ (by *I*, *C*)

8 $B^{n-2}(E_{n-1} \,\&\, \sim B^{n-2}E_{n-1})$ (*6*, *7*, *E*, *C*)
(since '$E_n \rightarrow E^{n-1}E_n$' and '$E_n \,\&\, \sim B^{n-1}E_n$' are inconsistent).

9 $B^{n-2}E_{n-1} \,\&\, B^{n-2} \sim B^{n-2}E_{n-1}$ (by *8*, *E*)

10 $B^{n-2}E_{n-1} \,\&\, \sim B^{n-2}E_{n-1}$ (by *9*, *D*)

Hence: $p \,\&\, \sim p$; i.e. contradiction.

IV. Now the above theorem amounted to the impossibility that a rational student (judged by certain most plausible assumptions on rationality) would believe the teacher (by showing that if he believes that a surprise exam would be given in the next two days, certain most plausible rationality conditions result in a contradiction).

Hence, given that the student is rational, he cannot believe the teacher. Hence he may 1. either not believe that there would be any exam at all in the next two days, 2. or he may believe that there will be one, but that it need not be a surprise exam.

In the first case, clearly an exam given on any of the two days will be a surprise. In the second case, an exam on the first day will be a surprise, since the student would not be able to rule out on the evening before the first day that the exam will be given on the second day, hence he would not be able to believe (rationally) that the exam would be given on the first day.

Hence, the student can be surprised in either of those two cases, hence there can be a surprise exam. Hence the paradox is illuminated.

The paradoxical air results from the fact that the sentence utterred by the teacher cannot be believed by a rational student. Yet of course the sentence can still be true. If the student reflects on the inconsistency demonstrated in the above proof, he may realize that, as a consequence, there *can* be a surprise exam. Yet he would not be able to believe that there *will* be a surprise exam without a contradiction. But his belief that there *can* be a surprise exam and his lack of belief that there *will* be one (say through a suspended judgement on whether there will be a surprise exam)are certainly compatible.

V. The source of he paradoxical air is that were a rational student to entertain a belief in a surprise exam, he would be led to realize its impossibility, the holding of both of which is impossible for a rational student.

Let us explain why the line of reasoning of the student, which seemingly establishes that there cannot be a surprise exam, is faulty.

Notation: Read '*SE_i*' as: an exam on day i would be a surprise.

Then: $SE_i \equiv \sim B^{i-1}E_i$.

I- — $(E_n \mathbin{\&} B^{n-1}E_n) \overline{\vee} (E_{n-1} \mathbin{\&} B^{n-2}E_{n-1})$

(distinguish *I*- from *I* !) i.e., *I*- states that there would be a surprise exam on one of the two days. Thus: $B^{n-2}I\text{-} \equiv I$ (hence the name '*I*-': *I* without the prefix 'B^{n-2}').

The student's line of reasoning (presented in section I above) from the assumption *I*- can be formalized as follows, given that he is rational:

Assume *I*-.
Now: $E_n \rightarrow B^{n-1} \sim E_{n-1} \rightarrow B^{n-1}E_n \rightarrow \sim E_n$.
Hence: 1. $\sim (E_n \,\&\, SE_n)$;
2. $E_{n-1} \rightarrow B^{n-2}E_{n-1}$; but from
3. $B^{n-2}E_{n-1}$: $B^{n-2}E_{n-1} \rightarrow \sim SE_{n-1}$,
hence: $\sim (E_{n-1} \,\&\, SE_{n-1})$.
Hence: $\sim$ I-.

But 3. does not follow from E_{n-1} in 2. above. This is so since even if the exam would be given on day $\underline{n} — 1$ (and not on n), and the student would believe that $(E_n \overline{V} E_{n-1})$,still he would be able rationally *not* to believe on n — 2 that E_{n-1}. It would follow, however, if we were to assume that $B^{n-2}I$-.

If this reconstruction of the argument is correct, then either it is outright invalid, or else '$B^{n-2}I$-' is smuggled as an implicit assumption into the argument, even though it does *not* follow from *I*- and the rationality requirements (hence it is consistent for the student *not* to *assume* it).

Hence it seems that either the argument is outright invalid, or that assumption I (which is: $B^{n-2}I$-) has to be used. But *I* cannot be used by a rational student, as it contradicts cannons of rationality (as was shown above in the theorem). But self-contradictory assumptions can lead to any conclusion, in particular to $\sim$ *I*-. Hence it is not surprising that if *I* is a premise, $\sim$*I*- is a conclusion. The resolution lies in realizing that the premises are self-contradictory.

The argument, however, is suggestive, (which makes the paradox less conspicuous and also puzzling) in that the irrationality of holding *I* is not trivial in the sense of immediately recognizable, nor is its use as a premise for the argument made explicit.

VI. In the proof of the Reformulated Thesis above step *8* was:

8 $B^{n-2}[E_{n-1} \& \sim B^{n-2}E_{n-1}]$

which cannot be true for a rational believer by Moore's paradox (whose statement has the general form: $B(p \& \sim Bp)$. Hence, incidentally, we may replace, if we wish, condition *D* there by:

*D** $\sim B(p \& \sim Bp)$).

Hence it is clear why statement *I* cannot be rationally held: the statement of Moore's paradox is a *paraphrase* of it (under plausible rationality conditions). Hence the paradoxical result (that $\sim$ *I*-) results from a statement which cannot be true for a rational believer since the statement of Moore's paradox cannot. Hence the paradox is a *version* of Moore's paradox, obscured by some complications. The sting of the student's reasoning lies in either of two points: Either in a fallacious derivation (if the assumption is *I*-), or in a derivation from inconsistent premises.

That the source of trouble is Moore's paradox is obvious, since 'surprise' here means that the (rational) student could not believe on the night before the exam that there will be an exam the next day. But for him to believe the teacher is to believe that (for the general case of n-days rather than 2-days only):

(E ! i) ($n \geqslant i > 0$ & there will be an exam on i but the student would not believe on $i - 1$ that there will be an exam the next day) ≡

≡ (E ! i) ($n \geqslant i > 0$ & there will be an exam on i but the student would not believe it on $i - 1$) ≡

≡ (L) $(E\,!\,i)\ (n \geqslant i > 0 \ \&\ E_i \ \&\ \sim B^{i-1}E_i)$

which has the form:

(K) $(E\,!\,i)\ (n \geqslant i > 0\ p(i) \ \&\ \sim B^{i-1}p(i))$

When appended by 'B^o' (day 0 is the day just befor the days), (K) resembles Moore's statement, except for the quantifier on days and the difference in times of belief — the complications which make it less obviously irrational to believe. Thus, the logical feature revealed by the paradox is that

$$(M)\ B^{o}(E\,!\,i)\ (n \geqslant i > 0\ \&\ E_i\ \&\ \sim B^{i-1}E_i)$$

is logically impossible for a rational believer, and that this is a modification of Moore's paradox.

This formulation involves the possibility of more days than two. If, however, the number of days is restricted to one (i.e., if n = 1), then the above statement becomes exactly Moore's statement. For n = 2, the 2-days case, it reduces to statement *I* above.

The above modification of Moore's paradox should be provably inconsistent for a rational believer by generalizing over n in the rationality assumptions *A* to *E*.

A natural modification of Moore's paradox which seems to retain its paradoxical features is the generalization on the time of the embedded belief (in (M)):

$$B^{o}(E\,!\,i)\ (k)\ (n \geqslant i > 0\ \&\ i > k \geqslant 0\ \&\ Ei\ \&\ \sim B^{k}E_i).$$

Thus 'p but I *shall* not believe that p' is irrational to assert as well as 'p but I *do* not believe that p' (though less conspicuously so). Generalizing in a different direction, a more general form of this statement-form (M) results in appending to (K) the belief operator 'B°':

$$B^{o}(E\,!\,i)\ (n \geqslant i > 0\ \&\ p(i)\ \&\ \sim B^{i-1}p(i)).\ [2]$$

It is interesting to conjecture under what kind of restriction on p(i) this would be inconsistent for a rational believer — what restrictions on p(i) should qualify it as a generalization of Moore's paradox.

Brandeis University, Igal Kvart

[2] It seems indeed that 'i — 1' could be generalized as well under suitable to a universally bount j, for $i > j \geqslant 0$.