JOSEPH Y. HALPERN AND YORAM MOSES

TAKEN BY SURPRISE: THE PARADOX OF THE SURPRISE TEST REVISITED

A teacher announced to his pupils that on exactly one of the days of the following school week (Monday through Friday) he would give them a test. But it would be a surprise test; on the evening before the test they would not know that the test would take place the next day. One of the brighter students in the class then argued that the teacher could never give them the test. "It can't be Friday," she said, "since in that case we'll expect it on Thurday evening. But then it can't be Thursday, since having already eliminated Friday we'll know Wednesday evening that it has to be Thursday. And by similar reasoning we can also eliminate Wednesday, Tuesday, and Monday. So there can't be a test!"

The students were somewhat baffled by the situation. The teacher was well-known to be truthful, so if he said there would be a test, then it was safe to assume that there would be one. On the other hand, he also said that the test would be a surprise. But it seemed that whenever he gave the test, it wouldn't be a surprise.

Well, the teacher gave the test on Tuesday, and, sure enough, the students were surprised.

1. INTRODUCTION

Was the teacher telling the truth? It seems that he was. After all, the test was given, and the students were surprised. Yet where was the flaw in the student's reasoning?

The story described above is the well-known Surprise Test Paradox, also known as the Class A Blackout, the Hangman Paradox, the Prediction Paradox, etc. It was circulated by word of mouth in the 1940s, and was first discussed in print in 1948 [OC]. Interestingly enough, the first few authors who discussed it viewed it simply as an example of a statement that could not be fulfilled, and were unaware of the potential "twist" at the end. It was not until 1951 that Scriven pointed out that the teacher can give the test and surprise the students [Sc]. Since then, numerous authors have discussed the problem and presented solutions, although none apparently definitive. (See [Ga] for an eminently readable introduction to the paradox, and [MB] for a thorough survey of the literature, with a bibliography listing 40 papers).

Journal of Philosophical Logic **15** (1986) 281–304.

With some trepidation, we offer yet another solution to this paradox. The solution uses elements present in a number of other solutions, yet puts them together in what seems to be a novel way.

The most important step in our analysis (and, we would argue, the most important step in *any* analysis of this paradox) is making precise exactly how we should interpret the phrase "on the evening before the test they would not know . . . ". Of course, the crucial word here is *know*. One approach we can take is to try to formulate this in terms of a modal logic, with a modal operator K for knowledge. This approach has been taken by a number of authors (e.g. [Bi, Kv, Le]), and tends to reduce the problem to a variant of Moore's "pragmatically paradoxical" sentence.[1]

Our approach, which goes back to Shaw [Sh], is to interpret "the students know" as "the students can deduce from the information given by the teacher". But with this approach a number of questions need to be answered. Exactly what information did the teacher give the students? And what rules of deduction can the students use? In what logic are they working?

As has been noticed by many previous authors (the first of which again seems to have been Shaw [Sh]), the teacher's statement has some self-referential features (this is discussed in more detail in the next section). We enrich our logic with a fixed-point operator to capture these self-referential features. This is not the first use of fixed-point operators to capture self-reference; they have been used frequently before, particularly in the computer science literature (cf. [SD, Pa, Pr, Ko]).

As Margalit and Bar-Hillel point out [MB]:

> Within the logical framework, Shaw and others demonstrated convincingly several things: (i) that there are more than one set of rules which can be substituted for the teacher's decree; (ii) that the differences between them are often subtle and easy to overlook, even with careful analysis, although they are crucial to the issue at hand; (iii) that some, though not all, of the rules are genuinely contradictory; (iv) that some, though not all, allow the possibility of a surprise test being given.

The four points made above are all readily apparent in our discussion. In fact, we suggest four ways to translate the teacher's statement formally into logic, all subtly different. The first of the translations turns out to be contradictory, the second is consistent and allows the

possibility of a surprise test being given on any day (including Friday!), the third rules out the possibility of the test being Friday, but is consistent with the test being any other day, while the fourth does not even admit of a reasonable semantic interpretation, and is paradoxical in that it is consistent if and only if it is inconsistent. All of the translations attempt to directly capture the spirit of the teacher's statement, rather than avoiding it as some authors seem to have done. The fact that the teacher's statement allows four translations with such wildly different properties may help explain the "staying power" of this puzzle.

We discuss our translations informally in the next section, and formalize the details in Sections 3 and 4. We compare our approach to the modal logic approach in Section 5, and conclude in Section 6 with some general observations on the logical analysis of this and similar paradoxes.

2. TRANSLATING THE PUZZLE INTO LOGIC: AN INFORMAL DISCUSSION

How can we represent the puzzle in a formal system in order to analyze it? The first step is to capture the information given by the teacher. Certainly one piece of information given by the teacher is:

I_1 The test will take place on exactly one of the days of the following school week.

I_1 can easily be represented in a propositional logic, using propositions of the form T_D standing for "the test is given on day D".

Representing the information that the test will be a surprise is somewhat trickier. As suggested by the story, we take "surprise" to mean that the students will not know the evening before the test that the test will take place the next day. But this does not seem to help very much. How do we capture what the students know? Instead of using a modal logic with a knowledge operator, we will take "know" here to mean "can deduce from the teacher's announcement". But with this reading, the teacher's announcement then becomes, more or less, that the students will not be able to deduce from the teacher's

announcement when the test will be. Clearly there is a case of self-reference here!

Roughly speaking, following Shaw ([Sh]), we can capture this information by

I_2 If the test is held on day D, then on the previous evening the students will not be able to deduce from I_1 and I_2 that the test will take place on day D.

Suppose we could represent I_2 formally in our logic (we will show in the next section how to do this using a provability operator and a fixed point operator). Let 'I_2' be the formula that represents I_2. (In general we will use 'I_j' to denote the formula that represents the information I_j.) Not surprisingly, it will turn out that 'I_1' $\wedge$ 'I_2' is not satisfiable; in fact, it will be provably false in our logic. The proof will follow exactly the same lines of reasoning as the student's argument in the story.

But in classical logic, from a false statement one can deduce any thing at all. So (assuming we are working in the framework of classical logic), the students can, *still within the logic*, now construct a proof (using the assumptions 'I_1' $\wedge$ 'I_2') that the test will be held Monday. And on Tuesday and on Wednesday, for that matter! If we take "surprise" to mean "not deducible from 'I_1' $\wedge$ 'I_2'", then the test will not be a surprise at all![2]

According to this translation of what the teacher said, the teacher is not telling the truth. The students are not surprised, because every evening they can deduce that the test will be held on the next day. But there seems to be something not quite right with this notion of surprise. After all, if we "prove" the test is on Monday, and then discover it isn't, can we honestly say we are not surprised if we "prove" the test is on Tuesday, and it turns out to actually be Tuesday?

This suggests that we should slightly reinterpret out definition of surprise. The crucial point here seems to be: can you really be said to know something as a result of having deduced it from inconsistent information? The classical logician may rub his hands with glee on receipt of inconsistent information, since he can now prove anything. But in everyday life, we don't take this approach at all. If someone gives us information we know to be inconsistent, we simply discard

it as useless (and consider with some doubt all the other information given to us by this apparently unreliable person!).

However, we are still interested in providing a logical analysis of the paradox. If we grant that our first attempt at capturing the notion of "surprise" is not quite right, we are still faced with the problem of providing a translation that is more in the spirit of our understanding of the story. One approach we might take is to adopt a non-classical logic, such as *relevance logic* or one of its cousins (cf. [AB]), where it is *not* the case that from an inconsistency you can deduce everything.

We prefer a solution that captures somewhat more closely the idea that you "discard" inconsistent information. There are two directions we can go from here in translating the teacher's information. The first is for the students to be somewhat charitable to the teacher and, despite the fact that he may have said I_2, to interpret his second piece of information as

I_3 If (I_1 and I_2) is consistent and the test is held on day D, then on the previous evening the students will not be able to deduce from (I_1 and I_2) that the test will take place on day D.

We remark that in the logic that we present in the next section, we have an explicit provability operator. Since consistency is just the dual of provability, it will turn out to be straightforward to express I_3 formally in the logic.

Of course, it is still the case that 'I_1' $\wedge$ 'I_2' is inconsistent, but this makes 'I_3' vacuously true. But I_3 does seem to capture part of the process the average (intelligent) listener goes through on hearing the story. He is told the proof that there can't be a test on any day of the week, and still grants that he is surprised when the test indeed takes place on Tuesday. Of course, he is surprised because he realizes that the "proof" that it can't be Tuesday somehow doesn't count. And I_3 seems to capture exactly why it doesn't count.

If the students interpret the teacher's statement as 'I_1' $\wedge$ 'I_3', then not only is this consistent, but it is consistent with the students getting the test on any day of the week, including the last day!

This seems to violate common sense. How can the teacher be telling the truth and still give the test on Friday? All the students should be

able to figure out on Thursday night that the test must be Friday if it hasn't been held yet. But doesn't similar reasoning let them figure out on Wednesday night that the test will be on Thursday if it hasn't been held yet? After all, they have already eliminated Friday. Such reasoning quickly leads us down the road to inconsistency again. However, there is one thing that distinguishes Friday from the other days: we can conclude on Thursday night that the test will be on Friday (if it hasn't been held yet) just by using the information in I_1. Of course, interpretation I_3 does not allow us to use this information unless I_1 and I_2 together are consistent. But suppose we slightly modify I_3 to allow the students to use I_1 alone, if that helps. This give us

I_4 If I_1 is consistent and the test is held on day D, then on the previous evening the students will not be able to deduce from I_1 that the test will take place on day D; and if (I_1 and I_2) is consistent and the test is held on day D, then on the previous evening the students will not be able to deduce from (I_1 and I_2) that the test will take place on day D.

Now the second clause of I_4 is identical to I_3, and holds vacuously just as before. But it is easy to see that the first clause rules out the possibility of having the test on the last day of the week. Thus 'I_1' $\wedge$ 'I_4' is consistent with the test being held any day of the week but the last one.

There is yet another, perhaps more natural, way of incorporating into our translation the idea that inconsistent information should be discarded. Rather than impose externally on the interpretation of the teacher's statement that we discard inconsistent information (as in I_3), we actually make it part of the translation of the teacher's statement, as in I_5 below:

I_5 If (I_1 and I_5) is consistent and the test is held on day D, then on the previous evening the students will not be able to deduce from (I_1 and I_5) that the test will take place on day D.

Translation I_5 is truly paradoxical. Informally, we can show that 'I_1' $\wedge$ 'I_5' is consistent iff it is inconsistent. For suppose that it is consistent. Then 'I_5' essentially reduces to 'I_2', and the standard argument used by the bright student in the story can be used to show

that 'I_1' $\wedge$ 'I_5' is inconsistent. On the other hand, if 'I_1' $\wedge$ 'I_5' is inconsistent, then 'I_5' is vacuously true, so 'I_1' $\wedge$ 'I_5' is equivalent to 'I_1', which is clearly consistent!

This argument must remain only informal. Due to a syntactic restriction that we place on the application of fixed-point operators in our logic, we cannot even express I_5 (i.e., there is no formula 'I_5' in our logic). Indeed, we can even prove within the logic that if there were any formula φ with the properties we require of I_5, then 'I_1' $\wedge$ φ would be consistent if and only if it were inconsistent.

We remark that the syntactic restriction we place on the application of fixed-point operators is not an *ad hoc* restriction employed expressly for dealing with this paradox, but rather is a standard restriction imposed on logics with fixed-point operators. Attempts to express other paradoxical sentences such as "This sentence is false" in the logic also violate this syntactic restriction. Of course, if we try to enrich the logic so that it can express a sentence like 'I_5', we quickly run into inconsistencies such as the well-known Knower's Paradox (cf. [KM]).

In summary, if we translate the teacher's statement in the naive way, as 'I_1' $\wedge$ 'I_2', then it is false, since the students can indeed deduce the day of the test. But this translation does not seem to capture the intuitive notion of "surprise". The translation 'I_1' $\wedge$ 'I_3' seems to come closer to capturing our intuitive reading of what the teacher said, although it is not a direct translation of his statement. Rather, it seems that it is the translation into logic of our *interpretation* of the statement. If we interpret the teacher's statement as 'I_1' $\wedge$ 'I_3', then no matter which day he gives the test, the statement becomes true. A slight modification, 'I_1' $\wedge$ 'I_4', is true as long as the test is not given on the last day of the week. Finally, we might consider translating the teacher's statement as 'I_1' $\wedge$ 'I_5', but the self-reference in this version is such that we cannot even make semantic sense out of this translation. No wonder the teacher's statement is so baffling!

3. A TRANSLATION INTO A FORMAL LOGIC

As we mentioned above, we translate the teacher's information into a propositional logic with a provability operator and a fixed-point

operator. The syntax is straightforward: starting with a set of primitive propositions p, q, r, . . . and two special propositions **true** and **false**, we get more complicated formulas by closing off under negation, conjunction, disjunction provability, and fixed points. Thus, if φ and ψ are formulas, then so are $\sim\varphi$, $\varphi \wedge \psi$, $\varphi \vee \psi$, and $Pr(\varphi)$ (read "φ is provable"). Furthermore, if p is a primitive proposition such that all free[3] occurrences of p in φ are *positive* and there is no free occurrence of p in a subformula of φ of the form **fix** $q.\psi$, then **fix** $p.\varphi$ (read "fix of φ (with respect to p)") is a formula. An occurrence of p in a formula φ is said to be positive if it is in the scope of an even number of negations (cf. [Ko]). Note the second restriction prevents **fix** $p.$(**fix** $q.(q \wedge p)$) from being a well-formed formula. This restriction is added for technical reasons in order to enable us to give semantics to our language easily. As we shall see when we give the semantics, our interpretation of the Boolean connectives, $\sim$, $\wedge$, and $\vee$ is classical, so we can easily define implication and equivalence in the standard way: $\varphi \Rightarrow \psi$ is an abbreviation for $\sim\varphi \vee \psi$ and $\varphi \equiv \psi$ is an abbreviation for $(\varphi \Rightarrow \psi) \wedge (\psi \Rightarrow \varphi)$.[4] We also take $Con(\varphi)$ (read "φ is consistent") as an abbreviation for $\sim Pr(\sim\varphi)$. Thus, as usual, consistency is the dual of provability.

We now list a few properties (axioms and rules of inference) we would like our logic to have, and show that these properties indeed can be used to completely formalize the discussion of the previous section. The semantics we give for the logic in the next section will indeed have all the required properties.

Since we intend our logic to be an extension of classical propositional logic, we would like to be able to assume all the standard properties of propositional logic. In particular, we should have as axioms:

A0. Every substitution instance of a tautology of propositional logic.

Thus, for example, $Pr(\varphi) \vee \sim Pr(\varphi)$ will be a valid formula. Of course, we also want *modus ponens*:

R0. $$\frac{\varphi,\ \varphi \Rightarrow \psi}{\psi}$$

The notion of provability has received much attention in the philosophical literature (cf. [Bo]). Our notion of provability is a rather unsophisticated one. The main properties we require of it are captured by the two axioms and one rule of inference given below:

A1. $Pr(\varphi) \Rightarrow \varphi$

A2. $Pr(\varphi) \wedge Pr(\varphi \Rightarrow \psi) \Rightarrow Pr(\psi)$

R1. $\dfrac{\varphi}{Pr(\varphi)}$

A1 guarantees that if something is provable then it is true, while A2 says that if both φ and $\varphi \Rightarrow \psi$ are provable, then so is ψ. Finally, R1 says that if φ is provable then so is $Pr(\varphi)$.

In our formal discussion of I_5, we will also require one further property of *Pr*, or rather of its dual *Con*. In order to make it precise, suppose v is a *valuation*; i.e., a mapping from primitive propositions to the truth values T and F such that $v(\textbf{true}) = T$ and $v(\textbf{false}) = F$. We can extend v so that it gives truth values to all propositional formulas (i.e., one with no occurrences of *Pr* or **fix**) in the standard way: $v(\sim\varphi) = T$ iff $v(\varphi) = F$, $v(\varphi \vee \psi) = T$ iff $v(\varphi) = T$ or $v(\psi) = T$, and $v(\varphi \wedge \psi) = T$ iff $v(\varphi) = T$ and $v(\psi) = T$. As usual, we will say that a propositional formula φ is *satisfiable* if $v(\varphi) = T$ for some valuation v. Clearly we want φ to be consistent if it is satisfiable. Thus, the following property should hold:

A3. $Con(\varphi)$ iff φ is a satisfiable propositional formula.

Note that this property assures us that there are some consistent formulas (or, more accurately, that there are formulas φ for which $Con(\varphi)$ is provable). We could, of course, extend this property so that it applies to non-propositional formulas as well; we will not need such an extension here.

The intuitive idea behind the fixed-point operator is that if we view a formula φ with a free propositional variable p as a "function" of p, then $\textbf{fix}\ p.\varphi$ is a fixed point of that function; i.e., $\psi = \textbf{fix}\ p.\varphi$ is a formula such that ψ is true iff $\varphi[\psi/p]$ is true, where we define $\varphi[\psi/p]$ to be the formula that results when we replace all free occurrences of p in φ by ψ (renaming bound variables in φ if necessary to avoid

"capturing" propositions free in ψ). Thus, we want the fixed-point operator to satisfy:

A4. $\quad \mathbf{fix}\, p.\varphi \equiv \varphi[\mathbf{fix}\, p.\varphi/p]$.

Note that it is not possible to find fixed-points of *all* formulas in such a way as to satisfy A4. For example, there is no fixed-point of $\sim p$ with respect to p, since if ψ were such a formula, then from A4 it would follow that ψ would be true iff $\sim\psi$ were true. Fortunately, our syntactic restrictions guarantee that $\mathbf{fix}\, p.\sim p$ is not a well-formed formula (since p does not occur positively in $\sim p$). Indeed, this restriction is there precisely to prevent formulas such as $\mathbf{fix}\, p.\sim p$ from being well-formed.

In Section 4 we give a semantics for this logic so that all instances of A0–4 are valid, and R0 and R1 are sound inference rules. But first we show how to translate I_1–I_5 in this logic and prove (using only A0–A4, R0, and R1) the properties that we claimed for them.

Let the primitive proposition T_D, $D = 1, 2, \ldots$ stand for "the test is held on day D" (where we take day 1 to be Monday, day 2 Tuesday, etc.). To simplify matters, we will assume that a "week" consists of only three days: Monday, Tuesday, Wednesday. Then the formula 'I_1' which represents the fact that the test will be given on exactly one day during the next week is simply

$$(T_1 \wedge \sim T_2 \wedge \sim T_3) \vee (T_2 \wedge \sim T_1 \wedge \sim T_3) \vee$$
$$\vee (T_3 \wedge \sim T_1 \wedge \sim T_2).$$

Using the provability operator, we can now capture in the logic the notion "cannot deduce ψ from φ" by a formula of the form $\sim Pr(\varphi \Rightarrow \psi)$. However, note that if the test is actually on Wednesday, the students, when trying to decide on Tuesday evening whether the test will be Wednesday, have more information than just I_1 and I_2. They also know that the test has not yet occurred; i.e. that $\sim T_1 \wedge \sim T_2$ holds. They certainly can (and do!) use this information in their deduction. Thus, the formula 'I_2' must satisfy

$$(*) \qquad \text{`}I_2\text{'} \equiv [T_1 \Rightarrow \sim Pr(\text{`}I_1\text{'} \wedge \text{`}I_2\text{'} \Rightarrow T_1) \wedge$$
$$T_2 \Rightarrow \sim Pr(\text{`}I_1\text{'} \wedge \text{`}I_2\text{'} \wedge \sim T_1 \Rightarrow T_2) \wedge$$
$$T_3 \Rightarrow \sim Pr(\text{`}I_1\text{'} \wedge \text{`}I_2\text{'} \wedge \sim T_1 \wedge \sim T_2 \Rightarrow T_3)].$$

Of course, all the occurrences of 'I_1' here are just abbreviations for the formula $T_1 \vee T_2 \vee T_3$.

This immediately suggests that the information I_2 can be captured by a fixed-point operator. Indeed, suppose we take 'I_2' to be

$$\textbf{fix}\ q.\varphi_0,$$

where φ_0 is the right-hand side of the equivalence above, with all occurrences of 'I_2' replaced by q. Thus, 'I_2' is

$$\begin{aligned}\textbf{fix}\ q.[&T_1 \Rightarrow \sim Pr(\text{`}I_1\text{'} \wedge q \Rightarrow T_1) \wedge \\ &T_2 \Rightarrow \sim Pr(\text{`}I_1\text{'} \wedge q \wedge \sim T_1 \Rightarrow T_2) \wedge \\ &T_3 \Rightarrow \sim Pr(\text{`}I_1\text{'} \wedge q \wedge \sim T_1 \wedge \sim T_2 \Rightarrow T_3)].\end{aligned}$$

Note that each occurrence of q in φ_0 is positive (once we rewrite subformulas of the form $\varphi \Rightarrow \psi$ as $\sim\varphi \vee \psi$). By A4, this representation of I_2 in our language indeed has property (*), as desired. Now it is a straightforward exercise to show that 'I_2' $\wedge$ 'I_2' is inconsistent using the axioms and inference rules given above. We follow the student's argument as given at the beginning of this paper. First we show that 'I_2' $\Rightarrow \sim T_3$ is provable, and then use that to show that 'I_2' $\Rightarrow \sim T_2$ is provable, and finally show that 'I_2' $\Rightarrow \sim T_1$ is provable. From this we can conclude that 'I_1' $\wedge$ 'I_2' is inconsistent. We sketch a few of the formal details below:

1. 'I_1' $\wedge \sim T_1 \wedge \sim T_2 \Rightarrow T_3$ (propositional reasoning)
2. 'I_1' $\wedge$ 'I_2' $\wedge \sim T_1 \wedge \sim T_2 \Rightarrow T_3$ (1)
3. $Pr(\text{`}I_1\text{'} \wedge \text{`}I_2\text{'} \wedge \sim T_1 \wedge \sim T_2 \Rightarrow T_3)$ (2, R1)
4. 'I_2' $\Rightarrow \sim T_3$ (3, (*))
5. 'I_1' $\wedge$ 'I_2' $\wedge \sim T_1 \Rightarrow T_2$ (4)
6. $Pr(\text{`}I_1\text{'} \wedge \text{`}I_2\text{'} \wedge \sim T_1 \Rightarrow T_2)$ (5, R1)
7. 'I_2' $\Rightarrow \sim T_2$ (6, (*))
8. 'I_1' $\wedge$ 'I_2' $\Rightarrow \sim T_1$ (4, 7)
9. $Pr(\text{`}I_1\text{'} \wedge \text{`}I_2\text{'} \Rightarrow \sim T_1)$ (8, R1)

10. $'I_2' \Rightarrow \sim T_1$ (9, (*))

11. $'I_2' \Rightarrow \sim 'I_1'$ (4, 7, 10)

12. $\sim('I_1' \wedge 'I_2')$ (11)

13. $Pr(\sim('I_1' \wedge 'I_2'))$ (12, R1)

The inconsistency of $'I_1' \wedge 'I_2'$ leads us to consider I_3. We can capture $'I_3'$ easily in our logic. It is just the formula:

$$Con('I_1' \wedge 'I_2') \Rightarrow 'I_2'.$$

We leave it to the reader to check that this formula indeed captures I_3 as described in the previous section. Since we have already shown $Pr(\sim('I_1' \wedge 'I_2'))$ (i.e., $\sim Con('I_1' \wedge 'I_2')$), it follows that $'I_3'$ is vacuously true and thus $'I_1' \wedge 'I_3'$ is equivalent to $'I_1'$. Now using A2 we can show that for any day of the week D, we have $Con('I_1' \wedge 'I_3' \wedge T_D) \equiv Con('I_1' \wedge 'T_D)$. Using A3 we can easily prove $Con('I_1' \wedge T_D)$ for any day of the week D. Thus $Con('I_1' \wedge 'I_3' \wedge T_D)$ is provable for every day D. So if we interpret the teacher's statement as $'I_1' \wedge 'I_3'$, then it is consistent for the teacher to give the test on any day of the week.

$'I_4'$ is like $'I_3'$, but it has one extra clause allowing us to use $'I_1'$ alone in trying to deduce when the test will be. This extra clause is essentially the same as (*), but without the self reference to $'I_2'$. Thus $'I_4'$ is the formula:

$$'I_3' \wedge [Con('I_1') \Rightarrow (T_1 \Rightarrow \sim Pr('I_1' \Rightarrow T_1) \wedge T_2 \Rightarrow \sim Pr('I_1' \wedge \sim T_1 \Rightarrow T_2) \wedge T_3 \Rightarrow \sim Pr('I_1' \wedge \sim T_1 \wedge \sim T_2 \Rightarrow T_3))].$$

Just as before $'I_3'$ is vacuously true, and clearly $'I_1'$ is consistent, so that $Con('I_1')$ is true. It is of course easy to see using propositional reasoning that $'I_1' \wedge \sim T_1 \wedge \sim T_2 \Rightarrow T_3$ is provable, so $Pr('I_1' \wedge \sim T_1 \wedge \sim T_2 \Rightarrow T_3)$ holds. Thus $'I_4'$ implies $\sim T_3$, thereby ruling out the last day. But no other days are ruled out, so that if we interpret the teacher's statement as $'I_1' \wedge 'I_4'$, then it is consistent for the teacher to give the test on any day of the week but the last.

Finally, in order to capture I_5, note that 'I_5' would have to satisfy

$$(**) \qquad \text{'}I_5\text{'} \equiv [Con(\text{'}I_1\text{'} \wedge \text{'}I_5\text{'}) \Rightarrow$$
$$T_1 \Rightarrow \sim Pr(\text{'}I_1\text{'} \wedge \text{'}I_5\text{'} \Rightarrow T_1) \wedge$$
$$T_2 \Rightarrow \sim Pr(\text{'}I_1\text{'} \wedge \text{'}I_5\text{'} \wedge \sim T_1 \Rightarrow T_2) \wedge$$
$$T_3 \Rightarrow \sim Pr(\text{'}I_1\text{'} \wedge \text{'}I_5\text{'} \wedge \sim T_1 \wedge \sim T_2 \Rightarrow T_3))].$$

Note that this is the same as (*) with the extra hypothesis that $Con(\text{'}I_1\text{'} \wedge \text{'}I_5\text{'})$ (and all occurrences of 'I_2' replaced by 'I_5'). It would seem that we could now capture 'I_5' by means of a fixed-point operator, just as we did 'I_2'. But this time there is a problem. When we replace abbreviations such as $\varphi \Rightarrow \psi$ by $\sim\varphi \vee \psi$ in the right-hand side of the identity (**), it is easy to see that the occurrence of 'I_5' in $Con(\text{'}I_1\text{'} \wedge \text{'}I_5\text{'})$ is in the scope of an odd number of negations. Thus we cannot replace 'I_5' with q and take a fixed point with respect to q without violating our syntactic constraints on fixed points.

This syntactic constraint is there for a reason! We discuss its need formally below, but for now we show that there can be no formula 'I_5' in our logic that satisfies the equivalence (**). For suppose there were such a formula. Let φ_1 be the formula

$$T_1 \Rightarrow \sim Pr(\text{'}I_1\text{'} \wedge \text{'}I_5\text{'} \Rightarrow T_1) \wedge$$
$$T_2 \Rightarrow \sim Pr(\text{'}I_1\text{'} \wedge \text{'}I_5\text{'} \wedge \sim T_1 \Rightarrow T_2) \wedge$$
$$T_3 \Rightarrow \sim Pr(\text{'}I_1\text{'} \wedge \text{'}I_5\text{'} \wedge \sim T_1 \wedge \sim T_2 \Rightarrow T_3)$$

Then we would have:

1. $\text{'}I_1\text{'} \wedge \text{'}I_5\text{'} \Rightarrow Con(\text{'}I_1\text{'} \wedge \text{'}I_5\text{'})$ (dual of A1)
2. $(\text{'}I_5\text{'} \wedge Con(\text{'}I_1\text{'} \wedge \text{'}I_5\text{'})) \Rightarrow \varphi_1$ ((**))
3. $\varphi_1 \Rightarrow \sim I_1$ (same as the argument that 'I_2' $\Rightarrow \sim I_1$ sketched above; we omit details here)
4. $(\text{'}I_1\text{'} \wedge \text{'}I_5\text{'}) \Rightarrow \sim(\text{'}I_1\text{'} \wedge \text{'}I_5\text{'})$ (1, 2, 3)
5. $\sim(\text{'}I_1\text{'} \wedge \text{'}I_5\text{'})$ (4)
6. $Pr(\sim(\text{'}I_1\text{'} \wedge \text{'}I_5\text{'}))$ (5, R1)

7. 'I_5' $\equiv$ **true** (6, (**))

8. ('I_1' $\wedge$ 'I_5') $\equiv$ 'I_1' (7)

9. Con('I_1') (A3)

10. Con('I_1' $\wedge$ 'I_5') (9, A2, and some propositional reasoning)

Lines 6 and 10 now give us the desired contradiction.

Thus, attempting to translate the teacher's statement by 'I_1' $\wedge$ 'I_5' is semantically meaningless! We cannot give a truth value to such an assertion (at least, not without running into contradictions).

4. A SEMANTICS FOR THE LOGIC

We now show how to give semantics to the logic introduced in the previous section in such a way as to make all the axioms and rules of inference sound. It turns out to be easy to give semantics to a propositional logic with a provability operator or with a fixed-point operator in such a way as to make all the required properties hold. However, having both a fixed-point operator and a provability operator adds a number of complications.

We begin by giving semantics to a propositional language with a provability operator, but no fixed-point operator. As mentioned above, a valuation v is a function mapping primitive propositions to $\{T, F\}$ (with v(**true**) $= T$ and v(**false**) $= F$). We now show how to extend valuations so they assign truth values to all formulas. We proceed by induction on the structure of formulas:

1. $v(\sim\varphi) = T$ iff $v(\varphi) = F$

2. $v(\varphi \wedge \psi) = T$ iff $v(\varphi) = T$ and $v(\psi) = T$

3. $v(\varphi \vee \psi) = T$ iff $v(\varphi) = T$ or $v(\psi) = T$

4. $v(Pr(\varphi)) = T$ iff $v'(\varphi) = T$ for all valuations v'.

We define a formula to be *valid* if it is true in all valuations; i.e., φ is valid if $v(\varphi) = T$ for all valuations v. Note that we have identified provability with validity, since $v(Pr(\varphi)) = T$ exactly if φ is valid. Similarly, we identify consistency with satisfiability.

It is easy to check that A0–A3 are sound with respect to this semantics, as are both rules of inference. Moreover, since the truth of a formula of the form $Pr(\psi')$ or $\sim Pr(\psi')$ does not depend on the valuation, it is easy to see that the following axiom is also sound:

A5. $Con(\varphi) \wedge \psi \Rightarrow Con(\varphi \wedge \psi)$,
if ψ is of the form $Pr(\psi')$ or $\sim Pr(\psi')$.

In fact, it can be shown that A0–A3, A5, R0, and R1 provide an elegant complete axiomatization for this logic (without the fixed-point operator), although we will not present this proof here.

In order to understand the semantics of the fixed-point operator, let us first consider the logic without the provability operator. In this case, we can extend valuations to formulas with fixed-point operators (but no provability operators) by adding the following clause to the first three clauses in the definition of valuation above:

5. $v(\textbf{fix}\ p.\varphi) = v(\varphi[\textbf{false}/p])$

To see why axiom A4 is sound with respect to this definition (at least as long as we do not have a provability operator in the language), we need two preliminary lemmas.

LEMMA 4.1. For all formulas φ, ψ with no occurrences of the provability operator, and all valuations v, if $v(\psi) = F$ then $v(\varphi[\textbf{false}/p]) = v(\varphi[\psi/p])$, while if $v(\psi) = T$ then $v(\varphi[\textbf{true}/p]) = v(\varphi[\psi/p])$.

Proof. By a straightforward induction on the structure of formulas. ■

Before we can state the next lemma, we need a few definitions. Suppose we put an ordering on $\{T, F\}$ by taking $F < T$. We will say that a formula φ is *monotonic* in p if for all valuations v, $v(\varphi[\textbf{false}/p]) \leq v(\varphi[\textbf{true}/p])$. Intuitively, a formula is monotonic in p if, viewed as a function of p, its truth value increases as the truth value of p increases.

LEMMA 4.2. If each free occurrence of p in φ is positive, then φ is monotonic in p.

Proof. By induction on the structure of φ. Details left to the reader. ■

Now suppose that φ is a formula with no occurrences of the provability operator such that every occurrence of p in φ is positive. If $v(\textbf{fix}\ p.\varphi) = F$, then by the definition above, $v(\textbf{fix}\ p.\varphi) = v(\varphi[\textbf{false}/p])$, and by Lemma 4.1, $v(\varphi[\textbf{false}/p]) = v(\varphi[\textbf{fix}\ p.\varphi/p])$. And if $v(\textbf{fix}\ p.\varphi) = T$, then by definition $v(\varphi[\textbf{false}/p]) = T$. From Lemma 4.2 it follows that $v(\varphi[\textbf{true}/p]) = T$, and then by Lemma 4.1 it follows that $v(\varphi[\textbf{fix}\ p.\varphi/p]) = T$. Thus, no matter what the truth value of $v(\textbf{fix}\ p.\varphi)$, we have that $v(\textbf{fix}\ p.\varphi) = v(\varphi[\textbf{fix}\ p.\varphi/p])$, so A4 is sound.[5]

Unfortunately, these arguments no longer hold if we allow both a provability operator and a fixed-point operator in the language, and give semantics by taking all five clauses in the definition of valuations. For example, let φ be the formula $Con(p) \vee q$ and let v, v' be valuations such that $v(q) = F$ and $v'(q) = T$. It is easy to see that $v(\varphi[\textbf{false}/p]) = F$ and $v'(\varphi[\textbf{false}/p]) = T$. From Clause 5 in the definition of valuations, it follows that we have $v(\textbf{fix}\ p.\varphi) = F$ and $v'(\textbf{fix}\ p.\varphi) = T$. Since $v'(\textbf{fix}\ p.\varphi) = T$, it follows that $v(Con(\textbf{fix}\ p.\varphi)) = T$. Now $\varphi[\textbf{fix}\ p.\varphi/p]$ is exactly the formula $Con(\textbf{fix}\ p.\varphi) \vee q$, so it follows that $v(\varphi[\textbf{fix}\ p.\varphi/p]) = T$. Thus $v(\textbf{fix}\ p.\varphi) \neq v(\varphi[\textbf{fix}\ p.\varphi/p])$. Axiom A4 is not sound with these semantics!

The problem is that although Lemma 4.2 still holds if we allow provability predicates in the language, Lemma 4.1 doesn't. An easy counterexample is provided by taking ψ to be $Pr(p)$, φ to be p, and v to be a valuation such that $v(p) = T$. It is certainly not the case that $v(Pr(p)) = v(Pr(\textbf{true}))$. It might seem that this problem is due to the fact that we consider p to be free even if it is in the scope of Pr. This is indeed the case. However, if we consider it to be bound, then it will not be possible to translate I_2 using the fixed-point operator, since we want the fixed-point operator to bind something in the scope of Pr.

There are in fact two special features of the language without Pr that make Clause 5 work.[6] The first is that the language is *extensional*: the value of $v(\varphi)$ depends only on the value of $v(p)$ for the primitive propositions p that appear in φ. However, Pr is an *intensional* operator. The value of $v(Pr(\varphi))$ depends on the value of φ under other valuations. The second feature is somewhat more subtle. Suppose we want to calculate $v(\textbf{fix}\ p.\varphi)$. As the arguments above show, we can

hypothesize that its value is F, the same as that of the primitive proposition **false**. We can then check if $v(\textbf{fix}\ p.\varphi) = v(\varphi([\textbf{false}/p]))$. If so, then we also have that $v(\textbf{fix}\ p.\varphi) = v(\varphi[\textbf{fix}\ p.\varphi/p])$, and our hypothesis is correct. If not, then we revise our hypothesis, taking $v(\textbf{fix}\ p.\varphi) = T$. As we have shown, this then gives us a value satisfying A4. Once we include Pr in the language, our hypotheses about the value of $\textbf{fix}\ p.\varphi$ must become more complicated. Since Pr is intensional, we must make a hypothesis about the value of $\textbf{fix}\ p.\varphi$ under *all* valuations, not just v. Further, we can no longer calculate the effects of this hypothesis in v by calculating $v(\varphi[\psi/p])$ for some formula ψ. Getting a semantics for the full language where A4 is sound requires more work.

We proceed inductively as follows. Suppose we are given a formula $\textbf{fix}\ p.\varphi$, and we have already defined $v(\psi)$ for every subformula ψ of φ (including φ itself) and every valuation v in such a way that $v(\psi)$ depends only on the primitive propositions free in ψ. Thus if v and v' agree on the primitive propositions free in ψ, then $v(\psi) = v'(\psi)$. (Of course, we assume that Clauses 1–4 in the definition of valuation are used when they apply.)

We define the *φ-formulas* to be those obtained by starting with the subformulas of φ and then closing off under negation, conjunction, disjunction, and provability (but not fixed-points!). It is important to note that in particular $\varphi[\textbf{fix}\ p.\varphi/p]$ is a φ-formula. This observation would not be true without our syntactic restriction that p does not appear free in the scope of a subformula of φ of the form $\textbf{fix}\ q.\psi$.

Let an *hypothesis* about the value of $\textbf{fix}\ p.\varphi$ be a function that yields a truth value for each valuation. Intuitively, $h(v)$ represents our hypothesis for the truth value of $\textbf{fix}\ p.\varphi$ in valuation v. We further assume that $h(v)$ depends only on the primitive propositions free in $\textbf{fix}\ p.\varphi$, so that $h(v) = h(v')$ if v and v' agree on all the primitive propositions free in $\textbf{fix}\ p.\varphi$. Note that this assumption guarantees that there are only finitely many hypotheses. Given a valuation v and an hypothesis h, let the *φ-valuation* vh be the function extending v that assigns truth values to all φ-formulas by taking $vh(\textbf{fix}\ p.\varphi) = h(v)$, and applying Clauses 1–4 in the obvious way (so that, for example, $vh(Pr(\psi)) = T$ iff $v'h(\psi) = T$ for all valuations v'). We

immediately get the following straightforward lemma; note that part (2) of the lemma is an analogue of Lemma 4.1.

LEMMA 4.3.

1. If ψ is a φ-formula which does not have **fix** $p.\varphi$ as a subformula, then for all valuations v and hypotheses h and h', we have $vh(\psi) = vh'(\psi)$.
2. For all valuations v, all hypotheses h, and all subformulas ψ of φ, if $h(v) = F$ then $vh(\psi([\mathbf{false}/p])) = vh(\psi([\mathbf{fix}\ p.\varphi/p]))$, and if $h(v) = T$ then $vh(\psi([\mathbf{true}/p])) = vh(\psi([\mathbf{fix}\ p.\varphi/p]))$.

Proof. By a straightforward induction on the structure of ψ. However, again note that we need to use the syntactic restriction that p does not appear free in formulas of the form **fix** $q.\psi$ to prove Part (2). ■

Now we can define the *revision rule* R (formally, a mapping from hypotheses to hypotheses) by taking $R(h)(v) = vh(\varphi[\mathbf{fix}\ p.\varphi/p])$.

Conceptually, we are working with the same idea that was described above for the language without *Pr*. We hypothesize some values for **fix** $p.\varphi$ and then calculate values for $\varphi[\mathbf{fix}\ p.\varphi/p]$ based on these hypotheses. These become the new hypotheses for the values of **fix** $p.\varphi$. This transition is captured by the revision rule R. Note that without *Pr*, extensionality ensured that there were only two relevant hypothesis to consider: whether the value of **fix** $p.\varphi$ under v is T or F. To find $v(\varphi[\mathbf{fix}\ p.\varphi/p])$ with the former hypothesis, we need only calculate $v(\varphi[\mathbf{true}/p])$, and with the latter $v(\varphi[\mathbf{false}/p])$. But now there are many more hypotheses to consider, and their effect cannot be taken into account as easily. It is this that leads us to the notion of vh.

We put the obvious ordering on hypotheses, taking $h \leqslant h'$ if $h(v) \leqslant h'(v)$ for all valuations v. Let h_0 be the least hypothesis with respect to this ordering. Thus $h_0(v) = F$ for all valuations v. Define $h_1, h_2, \ldots$ inductively via $h_{i+1} = R(h_i)$.

LEMMA 4.4. $h_0, h_1, h_2, \ldots$ is a monotonically increasing sequence.

Proof. We prove by induction on i that $h_i \leqslant h_{i+1}$. The base case is trivial, since h_0 is the least hypothesis. So suppose we have the result

for i and we wish to prove it for $i + 1$. Let v be any valuation. If $h_i(v) = F$, then clearly we have $h_i(v) \leqslant h_{i+1}(v)$. If $h_i(v) = T$, then we must have $i \geqslant 1$ (since $h_0(v) = F$ for all v by definition) and, again by definition, $h_i(v) = R(h_{i-1})(v) = vh_{i-1}(\varphi[\mathbf{fix}\ p.\varphi/p])$ and $h_{i+1}(v) = R(h_i)(v) = vh_i(\varphi[\mathbf{fix}\ p.\varphi/p])$. Since $h_i(v) = T$, by Lemma 4.3(2), $vh_i(\varphi[\mathbf{fix}\ p.\varphi/p]) = vh_i(\varphi[\mathbf{true}/p])$. Since $\varphi[\mathbf{true}/p]$ does not have $\mathbf{fix}\ p.\varphi$ as a subformula, by Lemma 4.3(1) we have $vh_i(\varphi[\mathbf{true}/p]) = vh_{i-1}(\varphi[\mathbf{true}/p])$. Now by Lemma 4.3(2) again, $vh_{i-1}(\varphi[\mathbf{fix}\ p.\varphi/p])$ is either equal to $vh_{i-1}(\varphi[\mathbf{false}/p])$ or $vh_{i-1}(\varphi[\mathbf{true}/p])$, depending on whether $vh_{i-1}(\mathbf{fix}\ p.\varphi)$ equals T or F. Now by our syntactic restrictions on the application of **fix**, we have that φ is monotonic in p, so in either case we have $vh_{i-1}(\varphi[\mathbf{fix}\ p.\varphi/p]) \leqslant vh_i(\varphi[\mathbf{fix}\ p.\varphi/p])$. Thus $h_i(v) \leqslant h_{i+1}(v)$, as desired. ■

Since $h_0, h_1, h_2, \ldots$ is an increasing sequence of hypotheses and there are only finitely many hypothesis, this sequence must have a fixed point; i.e., there must be some n such that $h_n = h_{n+1}$. Define $v(\mathbf{fix}\ p.\varphi) = h_n(v)$. We extend v to all φ-formulas using Clauses 1–4. It is easy to see that we now have $v(\psi) = vh_n(\psi)$ for all φ-formulas ψ and

$$\begin{aligned} v(\mathbf{fix}\ p.\varphi) &= h_n(v) = h_{n+1}(v) = vh_n(\varphi[\mathbf{fix}\ p.\varphi/p]) \\ &= v(\varphi[\mathbf{fix}\ p.\varphi/p]), \end{aligned}$$

so that A4 is sound.[7]

We remark that using the fact that $v(\varphi)$ only depends on $v(p)$ for the primitive propositions p that appear free in φ, it is straightforward to show that the validity problem for this logic is decidable, and can in fact be computed in time exponential in the size of the formula.

5. A COMPARISON WITH THE MODAL LOGIC APPROACH

In this section we compare our approach to that of Binkley [Bi]. We take the liberty of slightly modifying Binkley's notation to make the comparison proceed more smoothly. Binkley has a propositional language enriched with a modal operator K_D, $D = 1, 2, \ldots$ where $K_D\varphi$ can be interpreted as "on the evening before day D the students know φ to be the case".[8] In our notation, $K_D\varphi$ essentially corresponds

to $Pr(S_D \Rightarrow \varphi)$, where S_D is the stock of information that the students have on the evening before day D.

Binkley requires that K_D satisfy the following axioms and rule of inference:

A1′. $K_D\varphi \Rightarrow \sim K_D \sim \varphi$

A2′. $K_D\varphi \wedge K_D(\varphi \Rightarrow \psi) \Rightarrow K_D\psi$

A3′. $K_D\varphi \Rightarrow K_D K_D\varphi$

R1′. $\dfrac{\varphi}{K_D\varphi}$

If, as suggested above, we interpret $K_D\varphi$ to mean $Pr(S_D \Rightarrow \varphi)$, then for any choice of S_D these axioms and inference rule are easily seen to hold, given our semantics for Pr. Indeed, if we had added the following (sound) axiom to our system

A6. $Pr(\varphi) \Rightarrow Pr(Pr(\varphi))$,

then A1′, A2′, A3′, and R1′ would be provable from A0, A1, A2, A6, R0, and R1.

Binkley in addition requires the following two axioms:

A4′. $\sim T_D \Rightarrow K_{D'} \sim T_D$, if $D' > D$

A5′. $K_D\varphi \Rightarrow K_{D'}\varphi$, if $D' > D$

A4′ says that if the test does not happen on day D, then on the evening before any later day, this fact will be known, while A5′ says that the students do not forget facts that were previously known. (Actually, as Binkley points out, it suffices for his arguments that we replace A5′ by the weaker $K_D\varphi \Rightarrow K_D K_{D'}\varphi$ for $D' > D$.)

Provided that S_D is such that $\sim T_D \Rightarrow (S_{D'} \Rightarrow \sim T_D)$ for $D' > D$, then it is easy that the translation of K_D into our logic will satisfy A4′. And if $D' > D$ implies that $S_{D'} \Rightarrow S_D$, it will satisfy A5′ as well.

Binkley considers the case of a two-day week for simplicity and claims that in this case the teacher's announcement amounts to the following four assertions:

1. $T_1 \vee T_2$

2. $\sim(T_1 \wedge T_2)$

3. $T_1 \Rightarrow \sim K_1 T_1$

4. $T_2 \Rightarrow \sim K_2 T_2$

This should of course strike the reader as quite reiminscent of our translation. In particular, the first two clauses together are just a restatement of our 'I_1' in the two-day case, while the last two correspond to 'I_2'.

Binkley then claims that these assertions can all be true together. This amounts to saying that there are some choices of S_D satisfying the axioms above such that the conjunction of these statements (when translated into our logic) is consistent. He then goes on to prove formally that although these statements can be true together, they cannot be known by the students to be true. But in our terms, the fact that the students know these statements to be true is just to say that each S_D implies the conjunction of these four statements (i.e., they are part of the students' stock of information on each of the days). Once we assume this, then in some sense the *minimal* assumptions we can make on S_D is that they satisfy the fixed-point descriptions as in our translation of 'I_2'. Not surprisingly, Binkley's proof that $K_1\varphi$ is inconsistent, where φ is the conjunction of these four statements, has very much the same flavor as our proof of the inconsistency of 'I_1' $\wedge$ 'I_2'.

6. CONCLUSIONS

We have analyzed the Surprise Test Paradox by translating it into a formal logic with fixed-point operators and provability. We have given four possible translations of the teacher's statement. The first is provably false, the second is consistent, and is true no matter which day the teacher gives the test, the third is consistent but rules out the last day, and the fourth is paradoxical in that it cannot be given a truth value in our semantics.

The puzzle gains its force from the interplay between the translations. On first hearing the puzzle, most people seem to take some variant of translation 1, and agree that the teacher cannot give the test on any day of the week. When told that the test is given on Tuesday, they switch to some variant of translation 2 or translation 3, and admit (somewhat doubtfully) that the teacher did indeed seem to be telling

the truth. However, further thought usually leads to a realization of a possible paradox that amounts to some variant of translation 4.

The subtlety of the reasoning here is further reinforced by the apparent difficulty involved in giving semantics to a language with both a provability operator and a fixed-point operator. Both of these operators (or some variant of them) are necessary in order to fully capture the student's understanding of the teacher's statement.

Of course, by doing a translation into formal logic along the lines we have suggested here, it becomes much easier to carefully analyze exactly what is going on. In particular, it becomes clear that the paradoxical translation 4 is paradoxical exactly because of the attempt to take a fixed-point with respect to p of a formula where p does not occur positively. This approach also helps explain much simpler paradoxes like the Liar Paradox. In attempting to translate a statement like "This statement is false" into our logic, we end up with a formula like **fix** $p.\sim p$. Again the p does not occur positively, so this is not a well-formed formula.

ACKNOWLEDGEMENTS

We would like to acknowledge Arnon Avron, Jim des Rivieres, Maya Bar-Hillel, Faith Fich, Yossi Gil, Dexter Kozen, Daniel Lehmann, Steve Mahaney, David McAllester, Michael Rabin, and Ray Strong for fascinating discussions on the surprise test paradox and paradoxes in general. Jim's comments on the non-intuitive nature of translation 2 inspired the inclusion of translation 3. Ron Fagin, Anil Gupta, and Rich Thomason made some useful suggestions for improving the presentation of the paper. Anil's comments were particularly helpful in terms of improving the exposition of Section 4. Much of this work was done while the second author was at Stanford University, supported in part by DARPA contract N00039-82-C-0250 and an IBM Research Student Associateship.

NOTES

[1] A "pragmatically paradoxical" sentence is one of the form "p, but a doesn't know that p" for some fact p and agent a. Although this sentence is consistent, it is inconsistent (with the usual axioms for knowledge) for a to know it.

[2] This line of reasoning essentially appears already in [Au]. However, we carry it a bit further here.

[3] We omit the precise definition of *free* and *bound* propositions here, but intuitively, it is identical to the notion of free and bound variable in first-order logic, with the binding operation here being **fix p** rather than $\forall x$.

[4] We could, of course, also define disjunction in terms of negation and conjunction. For technical reasons, we prefer to have both $\wedge$ and $\vee$ as primitive operations.

[5] Here we have taken fix to be the *least* fixed-point operator. Intuitively, we have taken $v(\mathbf{fix}\ p.\varphi)$ to be false whenever this is consistent with A4. We should have also defined $v(\mathbf{fix}\ p.\varphi) = v(\varphi[\mathbf{true}/p])$. This would have given us the *greatest* fixed point. Essentially the same proof we have given shows that A4 is still sound with this definition.

[6] The exposition in the remainder of this section draws heavily on comments made by Anil Gupta. In particular, the notions of an *hypothesis* and a *revision rule* as described below are due to him.

[7] Using this procedure we are again taking **fix** to be the least fixed-point. We could have instead started with the greatest hypothesis, h^0, defined by $h^0(v) = T$ for all valuations v, and then used the revision rule to get a monotonically decreasing sequence of hypotheses. The fixed point obtained in this way is the greatest fixed point.

[8] Binkley uses J_D rather than K_D to indicate that he is interested in an ideal seeker after knowledge, not necessarily someone who already possesses knowledge. Thus J_D is supposed to represent what the ideal knower judges or believes to be true, not what he knows. These differences do not affect our discussion below.

REFERENCES

[AB] A. R. Anderson and N. D. Belnap, *Entailment, the Logic of Relevance and Necessity*, Princeton University Press, 1975.

[Au] A. K. Austin, 'On the unexpected examination', *Mind* **78**, 1969, p. 137.

[Bi] R. Binkley, 'The surprise test examination in modal logic', *Journal of Philosophy* **65** (5), 1968, p. 127–136.

[Bo] G. Boolos, 'The logic of provability', *American Mathematical Monthly* **91** (8), 1984, pp. 470–480.

[Ga] M. Gardner, 'A new paradox, and variations on it, about a man condemned to be hanged', *Scientific American* **208**, 1963, pp. 144–154.

[KM] D. Kaplan and R. Montague, 'A paradox regained', *Notre Dame Journal of Formal Logic* **1**, 1960, pp. 79–90.

[Ko] D. Kozen, 'Results on the propositional μ-calculus', *Theoretical Computer Science* **27**, 1983, pp. 333–354.

[Kv] I. Kvart, 'The paradox of surprise examination', *Logique et Analyse* **11**, 1976, pp. 66–72.

[Le] D. J. Lehmann, talk given at IBM San Jose, May, 1985.

[MB] A Margalit and M. Bar-Hillel, 'Expecting the unexpected', *Philosophia* **13**, 1984, pp. 263–288.

[OC] D. J. O'Connor, 'Pragmatic paradoxes', *Mind* **57**, 1948, pp. 358–359.

[Pa] D. M. R. Park, 'Fixpoint induction and proof of program semantics', *Machine Intelligence* **5** (ed. A. Meltzer and D. Michie), Edinburgh University Press, 1970, pp. 59–78.

[Pr] V. R. Pratt, 'A decidable μ-calculus' (preliminary report), *Proceedings of the 22nd Annual IEEE Conference on Foundations of Computer Science*, 1981, pp. 421–477.

[SD] D. Scott and J. deBakker, *A Theory of Programs*, unpublished, IBM, Vienna, 1969.

[Sc] M. Scriven 'Paradoxical announcements', *Mind* **60**, 1951, pp. 303–307.
[Sh] R. Shaw, 'The paradox of the unexpected examination', *Mind* **67**, pp. 382–384.

JOSEPH Y. HALPERN
IBM Almaden Research Center,
San Jose, CA 95120,
U.S.A.

YORAM MOSES
MIT,
Cambridge, MA 02139,
U.S.A.