themselves suggest, namely, heat and molecular motion, simply does not work in this case. So the materialist is up against a very stiff challenge. He has to show that these things we think we can see to be possible are in fact not possible. He has to show that these things which we can imagine are not in fact things we can imagine. And that requires some very different philosophical argument from the sort which has been given in the case of heat and molecular motion. And it would have to be a deeper and subtler argument than I can fathom and subtler than has ever appeared in any materialist literature that I have read. So the conclusion of this investigation would be that the analytical tools we are using go against the identity thesis and so go against the general thesis that mental states are just physical states.[19]

The next topic would be my own solution to the mind-body problem, but that I do not have.

[19] All arguments against the identity theory which rely on the necessity of identity, or on the notion of essential property, are, of course, inspired by Descartes' argument for his dualism. The earlier arguments which superficially were rebutted by the analogies of heat and molecular motion, and the bifocals inventor who was also Postmaster General, had such an inspiration: and so does my argument here. R. Albritton and M. Slote have informed me that they independently have attempted to give essentialist arguments against the identity theory, and probably others have done so as well.

The simplest Cartesian argument can perhaps be restated as follows: Let '*A*' be a *name* (rigid designator) of Descartes' body. Then Descartes argues that since he could exist even if *A* did not, ◇(Descartes ≠ A), hence Descartes ≠ *A*. Those who have accused him of a modal fallacy have forgotten that '*A*' is rigid. His argument is valid, and his conclusion is correct, provided its (perhaps dubitable) premise is accepted. On the other hand, provided that Descartes is regarded as having ceased to exist upon his death, "Descartes ≠ *A*" can be established without the use of a modal argument; for if so, no doubt *A* survived Descartes when *A* was a corpse. Thus *A* had a property (existing at a certain time) which Descartes did not. The same argument can establish that a statue is not the hunk of stone, or the congery of molecules, of which it is composed. Mere non-identity, then, may be a weak conclusion. (See D. Wiggins, *Philosophical Review*, LXXVII [1968], 90ff.) The Cartesian modal argument, however, surely can be deployed to maintain relevant stronger conclusions as well.

2

On Two Paradoxes of Knowledge*

I suppose that we have all heard the puzzle about the surprise execution or the surprise examination. I will state the paradox in terms of the examination; the execution, of course, puts the situation in more dramatic colors.[1]

The paradox can be stated in the following way. A teacher announces that he will give an examination within the month. Examinations are always given at noon. He also announces that the exam will be a surprise exam: no student will know on the day before the exam is given that it will be given the next day. A student can then reason as follows:

> The teacher, if he intends to fulfill his announced promise, cannot give the exam on the very last day. If he did, after noon had passed on the previous day, we (the students) would know that only the last day was left and that it had to be the day of the examination. This would be a plain contradiction of the announcement that it was to be a surprise exam, so that day can be crossed off the calendar. But then it cannot be given on the second-to-last day either because, after noon of the day before has passed and the exam still has not been given, we will realize that only two days are left, and that since the

* The present paper is based on a transcript of a recorded lecture given at Cambridge University to the Moral Sciences Club in 1972. The transcript appears to be known at least to B. Phil. students at Oxford, where it has been listed as part of their syllabus. The conversational tone of the paper, as in some other publications of mine, may sometimes reflect its origins. I have made changes and additions and included an appendix, but my solution to the surprise-exam paradox remains as it was presented in 1972 (and probably even earlier elsewhere).

[1] We now have in Sorensen (1988a: ch. 7) a history of the origins of this puzzle. Sorensen goes through much of the philosophical literature on the subject, including early occurrences. But also, basing himself on others whom he cites in the beginning of the chapter, Sorensen traces it to an actual civil defense exercise in Sweden during World War II (sometime in 1943–44) that was to be held within the week on a day not to be known in advance. A Swedish mathematician, Lennart Ekbom, detected a problem. His role in originating the paradox has been lost to the subsequent philosophical literature, but it is mentioned in some sources Sorensen cites.

Although when I gave the talk I appear to have been aware only of the hanging and examination versions, in Quine (1953) a surprise air-raid drill version is explicitly mentioned in the first paragraph, and this may have its origin in the actual event. More important, the date 1943–44 would appear to establish the surprise paradox as the earliest "backward induction" paradox, earlier than the (finitely) iterated prisoner's dilemma and Selten's chain-store paradox, both well-known to game theorists and probably to economists in general. I don't make any claims here as to how to analyze these.

last day is ruled out, the exam must be given on the second-to-last day. But then we would know in advance that that is the day of the exam, which is again a contradiction.

The student can continue the same reasoning backward: as soon as she crosses out one day from the calendar, the last remaining day is as if it were the last day, and so it can be ruled out by the same type of reasoning. There will, finally, be no days left. The student is then in a position to conclude that either the exam will not be given or it will not be a surprise.

As I said before, this problem is sometimes presented in terms of the hangman or judge who announces the execution of a prisoner under the same conditions. I wanted to present the puzzle as concerned with teachers and examinations because it should be realized that this is an everyday occurrence. Teachers *do* announce surprise examinations and no such contradictions seem to arise: it does not seem impossible for a teacher to fulfill his or her promise to give a surprise examination.[2]

It is interesting that this kind of problem is discussed as if it were a philosophical problem at all. How philosophical it actually is depends on whatever philosophical morals we may draw from it. Graham Greene classifies his works into novels, entertainments, and some other works: a novel is supposed to be a more serious work, but the entertainments are often the best. A problem like this might be classified as an entertainment in this sense. But it can have aspects of a "novel" if conclusions concerning our basic concepts of knowledge may be drawn from it. Here, more so than with typical philosophical problems, we are in the kind of "intellectual cramp" that Wittgenstein describes—one in which all the facts seem to be before us, there does not seem to be any new information to be gained, and yet we don't quite know what is going wrong with our picture of the problem.

I once did the following "scientific experiment," which can be a model for the problem. The experimenter announces to the subject that he has a deck of cards (which is finite)—it might be the whole deck or just part of it, but it includes the ace of spades. The cards are going to be turned over in order one by one, and the experimenter further tells the subject that he (the subject) will not know in advance when the ace of spades will turn up.[3]

Now suppose the deck consists of only one card. The experimenter says: "This card is the ace of spades, but you will not know which card it is until it has been turned over." The subject will think that this is obvious nonsense. Many people who have discussed the surprise-exam paradox have assumed that the significant

[2] We all know that in contemporary death penalty jurisprudence people are often not sure when they will be executed—last-minute appeals and the like make the execution date uncertain. But sadistic judges who announce that they have chosen a "surprise" date to execute a prisoner, are, I hope, very rare.

[3] I really think this should be done in a serious psychological laboratory; I did it as a student in college with a fellow student. One can try variations on the number of cards, and also vary whether subjects have heard of the "surprise exam" problem before the card experiment, or whether they have not been told about it before (but may or may not start going through the reasoning themselves during the experiment).

transition comes between the cases of one and two days (or rather, between ruling out the last day and the second-to-last day, as long as there are two or more). Maybe in some sense this is right. But suppose now that there are two cards in the deck, the ace of spades and another one, and the experimenter again announces: "You will not know in advance when the ace of spades will be turned over." When I tried the experiment, the subject,[4] who had heard of the paradox, reacted along the following lines: "There is still something very strange about this announcement. If you have put the ace on the bottom, I will not be surprised after you have turned up the first card. So, if you really mean to do what you say, you can't have put the ace on the bottom. But now I have proved that it must be the top card, and so again will not be surprised. I *do* know in advance."

Consider the case where all fifty-two cards, or at least a large number, are in the deck. Imagine that the experimenter, without telling the subject where it is, assures the subject that he has put the ace somewhere in the deck, and that the subject will not know in advance when the ace will come up if the cards are turned up one by one. Can the experimenter guarantee this? It seems clear that he can, say, by putting the ace somewhere in the middle.[5]

The subject can still go through the same kind of reasoning as in the case of the two cards. It *seems* that the reasoning may be generalized. However, it sounds very *un*convincing in this case. One therefore gets the impression that the reasoning gets weaker and weaker the more cards there are. This in itself is strange because it is the same piece of reasoning applied again and again.

Of course we are familiar with this kind of phenomenon from the paradox of the heap: if someone has only £1 to his or her name, she is poor; and if someone with only £N to his or her name is poor, so is someone with just £(N+1). Therefore, by mathematical induction, no matter how many pounds she has, she

[4] The subject was Richard Speier. If my memory is right, we were both undergraduates (so about 1960). I now (2009) find that Ayer (1973) mentions a card model, but it differs somewhat from the original problem, as in the case of the next footnote (though it is not quite the same as that one either).

[5] In the original version of this lecture I imagined that the experimenter put the ace somewhere in the middle right in front of the subject. But this is not the appropriate model for the examination version. Though no doubt the subject will not know in advance when the ace will turn up, neither will the experimenter. In the original examination problem the teacher has decided on a particular day to give the surprise exam, which makes the situation very different. If the experimenter acts as I described in the earlier version, the last card is excluded all right, but the reasoning of the surprise-exam paradox is superfluous, since everyone sees that the ace has been inserted somewhere in the middle. Similarly for the second-to-the-last card, and so on. (Actually, in the same original version, I eventually mentioned this point of disanalogy. But then I shouldn't have introduced this procedure as if it were an analogue in the first place). Exactly where the exclusion stops is somewhat indeterminate, and in contrast with what I say about the problem below in the text, something like vagueness may be involved. But here, whether or not it applies elsewhere (and I find the view rather dubious in general), it would be vagueness only in the sense of something like the now well-known characterization of Timothy Williamson (1994), since, after all, the card is in some definite place. Its exact position is simply unknown to both people, and the vagueness is merely epistemic (though in this case they could eventually find out the card's location).

is poor. It is a familiar philosophical problem that there must be something wrong with this, but it is hard to say exactly what. The heap/poverty problem involves reasoning with a vague predicate, but it is not clear that the issue in the present problem involves any question of vagueness.

What are the premises of the reasoning in this problem? The student is not to know on the day before the exam that it will be given. Let there be N days in which the exam may be given and let E_i mean that the exam will be given on day i. The teacher announces that the exam will be given on one of the first N days:

(1) E_i for some $i, 1 \leq i \leq N$ (equivalently, $E_1 \vee \ldots \vee E_N$)

The exam is going to be given on exactly one day; that is, it is not the case that it is going to be given on two distinct days:

(2) $\neg(E_i \wedge E_j)$ for any $i \neq j, 1 \leq i, j \leq N$

Then there is the announcement that the examination is to be a surprise. Let $K_i(p)$, for any statement "p," mean that the student knows on day i that p is true. So we can say that it is not the case that the student knows on day i–1 that the exam will be given on day i:

(3) $\neg K_{i-1}(E_i)$ for each $i, 1 \leq i \leq N$

If i is 1, then i–1 is 0, which means that she does not know in advance of the whole series (that is, before the first day) that the exam will be given on the first day.

We now have an additional premise. If the exam has not been given on one of the first i–1 days, then the student knows this on day i–1, as soon as noon has passed:

(4) $(\neg E_1 \wedge \neg E_2 \wedge \ldots \wedge \neg E_{i-1}) \supset K_{i-1}(\neg E_1 \wedge \neg E_2 \wedge \ldots \wedge \neg E_{i-1})$ for each $i, 1 \leq i \leq N$

Given premise (2), we can conclude that if the exam is going to be held on the ith day, it cannot have been given on any previous day. Hence, it follows from (4) and (2) that if the exam is to be given on the ith day, the student will know on the (i–1)th day that the exam has not been given on any of the first i–1 days:

(5) $E_i \supset K_{i-1}(\neg E_1 \wedge \neg E_2 \wedge \ldots \wedge \neg E_{i-1})$ for each $i, 1 \leq i \leq N$

So, these are the premises that are alleged to lead to a paradox. There may be some additional premises about knowledge itself that are required to carry out the reasoning. Obvious premises include: if a student knows any statement on day i, then it is true:

(6) $K_i(p) \supset p$ for each $i, 1 \leq i \leq N$

Also, we may require the "deductive closure of knowledge": if a student knows that p on day i, and knows that if p then q on day i, then she knows that q on day i:

(7) $(K_i(p) \wedge K_i(p \supset q)) \supset K_i(q)$ for each $i, 1 \leq i \leq N$

This premise is, in general, false—people may know all the premises of a deductive argument without knowing the conclusion. Mathematics would be a trivial subject if everyone's knowledge were deductively closed. It would be the easy way out, as the president of the United States would say,[6] to solve this problem by denying such a premise; but we can make the simplifying assumption that these particular students are clever enough to draw all the consequences of the things they know. This is not what is in question.[7]

We must also assume that, on any day, a student knows all principles of logic, including all propositional tautologies. This, again, is not true of students in practice, as anyone who has given a logic course knows. However, we can assume that it is true of these students, so they may do any kind of deductive reasoning. Let us symbolize this schema as follows:

(8) Taut $\supset K_i$(Taut) for each $i, 1 \leq i \leq N$

Can we now deduce a contradiction—that the exam cannot be given by surprise at all, which contradicts premise (1)—from all these premises? We start the reasoning by trying to show that the exam cannot be given on the last day. This reasoning must try to state that if the exam were to be given on the Nth day, the student will know on the (N–1)th day that the exam will be given on the Nth day (the last day), thus showing by *reductio ad absurdum* that the exam cannot be held on the Nth day. By substituting N for i in premise (5), we find that the student knows that the exam has not been given on one of the first N–1 days *on the (N–1)th day*. She knows from premise (1) that it must be given on one of the first N days, and so concludes on day N–1 that it must be given on the Nth day (i.e., $K_{N-1}(E_N)$). But this is an immediate contradiction of the relevant instance of premise (3) (viz. $\neg K_{N-1}(E_N)$). She then concludes that her initial hypothesis, (E_N), has been disproved by *reductio ad absurdum*, and thus that the exam cannot be given on the last day.

[6] The president was Richard Nixon when this talk was given.

[7] By the time I spoke, at least Fred Dretske (see 1970, 1971) had already denied that knowledge is always deductively closed, even for people able to make the deduction. Since then he has been followed by many others. This was supposed to protect against some problems of philosophical skepticism. But presumably these writers (who may or may not give the required restrictions) must think that only in very exceptional cases having to do with skepticism should the relevant principles of deductive closure fail. Otherwise, one person could accuse another of making the well-known fallacy of giving a valid deductive argument (from accepted premises) for his views!

I was probably unaware of Dretske's papers when I gave this lecture. However, I was fortunate enough to have Dretske in the audience when I gave a version of this lecture, and he was not only disinclined to object, but found the paper very convincing. See Chapter 7 for some discussion on this issue.

But this has a fallacy in it: all premise (1) says is that the exam will be given on one of the first N days, not that the student *knows* that fact. To reach the conclusion, we must also have the premise that the student knows on day N–1 that the exam will be given on one of the first N days (i.e., $K_{N-1}(E_i)$ for some i, $1 \leq i \leq N$). But we cannot obtain this from our premises. This is Quine's solution to the paradox in his article "On a So-Called Paradox" (Quine 1953). (He chooses the version in which a prisoner is to be hanged.) The prisoner is supposed to know that the judge's decree that he is to be hanged will be fulfilled. But how does he know this? Maybe the judge is a liar. As Quine puts it:

> It is notable that K [the prisoner] acquiesces in the conclusion (wrong, according to the fable of the Thursday hanging) that the decree will not be fulfilled. If this is a conclusion that he is prepared to accept (though wrongly) in the end as a certainty, it is an alternative which he should have been prepared to take into consideration from the beginning as a possibility.
>
> K's fallacy may be brought into sharper relief by taking n as 1 and restoring the hanging motif. The judge tells K on Sunday afternoon that he, K, will be hanged the following noon and will remain ignorant of the fact till the intervening morning. It would be like K to protest at this point that the judge was contradicting himself. And it would be like the hangman to intrude upon K's complacency at 11.55 next morning, thus showing that the judge had said nothing more self-contradictory than the simple truth. If K had reasoned correctly, Sunday afternoon, he would have reasoned as follows. "We must distinguish four cases: first, that I shall be hanged tomorrow noon and I know it now (but I do not); second, that I shall be unhanged tomorrow noon and know it now (but I do not); third, that I shall be unhanged tomorrow noon and do not know it now; and fourth, that I shall be hanged tomorrow noon and do not know it now. The latter two alternatives are the open possibilities, and the last of all would fulfil the decree. Rather than charging the judge with self-contradiction, therefore, let me suspend judgment and hope for the best." (Quine 1953:20, 21)[8]

Quine's solution to this problem has never seemed to me to be quite satisfactory: consider again the card experiment. It does seem strange, even though not literally contradictory, to take a card (face down) and say, "This is the ace of spades, but you do not know that this is the ace of spades."[9] In the second half of my utterance, am I inviting you to suppose that I cannot be trusted? Wasn't I communicating knowledge to you in the first half? Indeed, in this case (analogous to the one with only one day in the examination or hanging period), the hearer will not know what to believe, given the strangeness of the performance, and therefore would not know, though the strangeness does not persist if there are many cards.

Quine says that the fallacy derives from the fact that the prisoner does not *know* that the judge is telling the truth, or that the student does not *know* that there will be an exam given at all. But often, I think, you do *know* something simply because a good teacher has told you so. If a teacher were to announce a surprise exam to be given within a month, a student who did badly could not excuse herself by saying that she did not *know* that there was going to be an exam. If there is only one day, we have the anomalous situation I have just mentioned. But if there are many days, then it is natural to give the students knowledge on the basis of what the teacher tells them.

Clearly, we are justified in changing premise (1) to allow that a student *knows* from the beginning that an exam will be given on one of the first N days.

(1′) $K_0(E_i)$ for some $i, 1 \leq i \leq N$ (in other words $K_0(E_i \vee \ldots \vee E_N)$)

Similarly, we may allow that she *knows* at the beginning that the exam will not be given on two days and that it will be a surprise:

(2′) $K_0(\neg(E_i \wedge E_j))$ for any $i \neq j, 1 \leq i, j \leq N$

(3′) $K_0(\neg K_{i-1}(E_i))$ for each $i, 1 \leq i \leq N$

Can we now derive the paradox? We need two further premises. First, that if a student knows a statement on day i, she knows it on any later day:

(9) $K_i(p) \supset K_j(p)$ for any i, j such that $0 \leq i \leq j \leq N$

On its face this simply means that we are assuming that the student does not forget anything that she knows. Second, we need (though its use could perhaps be avoided) what has been called the double-K principle: that if a student knows on day i that p, then the student knows on day i that she knows on day i that p:

(10) $K_i(p) \supset K_i(K_i(p))$ for any $i, 0 \leq i \leq N$

[8] Two things to add, summarizing Quine's discussion. First, Quine originally discusses the case of many days and argues that the idea that the decree cannot be fulfilled if the hanging takes place on the last day is wrong; he also discusses it in a more abstract way—hence the phrase "restoring the hanging motif." He then goes on to draw the more extreme conclusion quoted in the text, that is, that even in the case of only one day there is no problem in the judge's announcement.

Second, one might elaborate on Quine's remark that "if this is a conclusion that he is prepared to accept (though wrongly) in the end as a certainty, it is an alternative which he should have been prepared to take into consideration from the beginning as a possibility" (1953:65). He says: "The tendency to be deceived by the puzzle is perhaps traceable to a wrong association of K's argument with *reductio ad absurdum*. It is perhaps supposed that K is quite properly assuming fulfillment of the decree, for the space of his argument, in order to prove that the decree will not be fulfilled" (66). Quine goes on to say that the argument of the puzzle requires not only the supposition that the decree will be fulfilled, but that the prisoner *knows* that it will be. This destroys any idea that this is a valid *reductio ad absurdum* argument, where only the weaker assumption would be allowed.

[9] See Moore's paradox ("p, but it is not the case that I believe that p"). As is well known, statements of this form are not contradictory, and may sometimes even be true, but anyone who utters one has made a strange performance. The case is similar here, with appropriate changes. Suppose someone asked me my name and I said, "It is Saul Kripke, but you still don't know what my name is." This may not literally be a contradiction, but is obviously very odd.

This latter is a controversial principle in the logic of knowledge; so one might think that that is what is going wrong and try to avoid the use of it. I will, however, make a few preliminary remarks about it. Is it the case that if someone knows something, she knows that she knows it? There have been two attitudes about this in recent philosophy. One is that all that is meant by knowing that you know something is that you know it. This sort of extreme attitude is stated, according to Hintikka (1962:108–10), by Schopenhauer, and is further argued by Hintikka himself. At the opposite extreme is the view[10] that, maybe we know many things: that Nixon is president of the United States, that the Russians had a revolution in 1917, that the sun is mostly gas—things which epistemological skeptics alone would deny. We do not, however, really *know* that we know them because that would involve a very high degree of certainty. Maybe our evidence does not constitute knowledge (though in fact it does, I guess, if we are lucky). To know that you know something is to perform a very great epistemological feat, which is not comparable to just knowing it—since you can't distinguish knowledge from mere justified belief in something false. It is very hard to adjudicate between these two positions, or even find a position in between, because "I know that I know that *p*" is not a sentence that we often find on our lips.

I would suggest the following in favor not of the principle being true, but of its being *nearly* true: true enough for all practical purposes. Suppose I know something—for example, I know that Nixon is the president of the United States now. The following is an argument for the double-K principle in this instance. Certainly *you* (in the audience) know that I know that Nixon is the president. For one thing, I have just said this, presumably basing my statement on newspapers, television, and so on. Even if I had not said this, if you knew me, you would presume that I knew the fact on the same basis. Surely, I am not normally in a worse position than *you* to judge this matter. Is there a principle of privileged non-self-access here? I would suppose that normally, if someone else can know that I know something, then I myself can know it at least on the same basis, though perhaps I do not need to use this basis. (I do not myself *need* to argue: well, I have said that Nixon is president, I read the newspapers, and so on; but if you can know that I know this on such a basis, it would be surprising that I am singularly worse off.)

In fact, the argument can be strengthened. For I know that you know that I know that Nixon is president. After all, I just said so. But knowledge implies truth. So if I know that you know something, I must know it myself. Hence, I know that I know that Nixon is president.

[10] The truth is that now (in 2009) I am not sure who in "recent philosophy" (i.e., in 1972) I might have had in mind as holding this opposite extreme. One could certainly imagine the plausibility that someone might hold such a view. However, the claim that knowing implies knowing that one knows can certainly be doubted, and has been doubted by many philosophers.

Yet another variation: I know that everyone who reads the newspapers knows that Nixon is president, and I know that I myself read the newspapers. Therefore, I know that I must know that Nixon is president. Once again, in the case of these variations, it is artificial to suppose that I need to go through the reasoning given, but nevertheless it is valid and implies the appropriate case of the double-K principle.

No doubt there may be exceptions where arguments such as these are not applicable, but the arguments should be acceptable in a very wide range of cases. I would, however, like to distinguish myself from Hintikka and others who advocate such a principle universally and as a near tautology (in Hintikka's case, on the basis of what appears to me to be a circular argument). However, because of the arguments I have given, I believe that one should have no doubt of the principle in the present case. Therefore, I think that the fallacy does not lie here at all. We may, therefore, put any number of Ks before the premises—not only does a student know on a given day that they are true, but she also knows that she knows on that given day that they are true, and so on.

One can now derive the contradiction from these premises plus the principle we have just discussed. Yet it seems very strange that even if the announcement that the exam will be given is true, the student cannot possibly *know* that it will be: for students surely do know that such exams will be given. Let us look again at premise (9), that if a student knows something on a given day, she knows it on any later day. Is this a generally true principle of epistemology? It is true that she can forget, but this is not what is in question: we may suppose that her memory is good enough not to forget any significant detail. Then, will it be true?

Many of you will know that I have written articles on modal logic; suppose I came to one of you and sadly denied that, and claimed ("admitted") they were written by someone named "Schmidt" and that I merely signed them. Suppose I even showed you a manuscript in Schmidt's handwriting. After a certain amount of persuading you might well be convinced that I did not write any articles on modal logic. So you would not, at that later date, even believe it, let alone know it. You may say that this means that you did not *know* it at the earlier time. This may be so if I am now telling the truth and have not written any articles on modal logic. But suppose now that I was lying, as some form of English joke, and really had written the articles. You would then have been correct in your initial belief, and, assuming that you were in a good enough position to support your belief rationally, and so on, it would seem you did know this at the earlier date but were rationally persuaded to change your mind. If you wanted to argue that you did *not* know at the earlier date, then you would have to say that you do not know a certain fact if at some future date someone produces phony evidence to change your mind. Thus, what is true for you now would be vulnerable to what might happen later. I think that I would rather say that one can know something now but *lose* that knowledge at a later date on the basis of

further misleading evidence. (It must be *misleading* evidence; if it were genuine, then your supposed knowledge would in fact have been mistaken belief.)

What happens in the particular case of the paradox? Let us again try to rule out the last day. We say, for the *reductio ad absurdum*, that the exam will be given on the Nth day. It will not have been given on any of the first N–1 days; the student therefore *knows* it will not have been given on the first N–1 days and, according to premise (4), knows this on day N–1. The student knew at the outset that it will be given on one of the first N days. So we can conclude that if she knew it was not given on one of the first N–1 days, then she knew that it must be given on the Nth, which contradicts premise (3). But there is a missing step in the argument, which is that just because she knows on day 0 that it will be given on one of the first N days, she must therefore know this on day N–1. This does not in fact follow without using premise (9), and would be false if the student later doubted that the exam was to be given at all.

Is it plausible to call this "a missing step"? Is it clear, in this case, that what the student knew on day 0 she would still know at any future date? The teacher has announced that the exam is going to be given on one of the N following days and that it is to be a surprise. What then will the students think when all the available days but one have passed? Perhaps that something has gone wrong, since, if the teacher still intends to give an exam, it will not be a surprise: they may therefore fall into doubt and say, "Look, maybe the teacher isn't intending to give us an exam now; maybe he's changed his mind." This would be a case of having had knowledge at one time but losing it at some later time. The student has the knowledge that the exam will be given at the outset, but she no longer has it at the end of the examination period. This seems to me to be fairly straightforward common sense, and so it is a fallacy to assume that the student retains this knowledge. Thus, the step of the argument which says that the last day can be ruled out is, in fact, erroneous, and so the whole argument never gets going. This is the basic fallacy in the argument.

This explanation, however, does not yield one feature of the problem which I have mentioned before—that, somehow, the more days there are, the worse the argument becomes. Perhaps we can in some way reinstate the argument by ruling out the last day. We cannot appeal to premise (9), since it is clearly false in this case. But we could add an extra premise: that the student knows, even on day N–1, that the exam will be given. We can make this plausible by saying that it is a rule of the school that a grade must be given for the course and that grades are always given on the basis of exams. Then the students would think on day N–1, when the exam still had not been given, "Something has gone wrong, but it cannot be that the teacher has decided not to give us an exam: it must be that he has decided not to bother making it a surprise."

Now we really can rule out the last day—rule it out in the sense that the premises contradict the supposition that the examination will be given on the last day, for the premises include that the exam, when given, will be a surprise. How

do we rule out the second-to-last day? The intuitive reasoning is something like this: knowing that the examination will not be given on the last day, the (N–1)th is the last possible day left, and then we go through the very same reasoning for that day as we did for the Nth day; we therefore will not need any more premises to allow us to rule out each day one-by-one going backward. But, as we have explicitly argued so far, this is a fallacy: we have not yet concluded that anyone *knows on any particular day* that the examination will not be given on the last day; we have merely concluded that it will not *in fact* be given on the last day.

What do we need for the student to reason? We have to know that on day N–2 she says "I know the exam cannot be on the last day, so there is only one day left for the exam, the (N–1)th." Then that day may be treated as if it were a new last day, and the supposition that it will be given on that day can be ruled out as contradicting the assumption that the exam will be a surprise. But this is fallacious, as things stand, because all we know is that *in fact* the examination will not be given on the Nth day, not that the student knows it on day N–2 (i.e., that $K_{N-2}(\neg E_N)$). A student whose knowledge is deductively closed will know this provided that she knows (on any given day) all the premises on which that conclusion was based. What premise did we use? We demanded that the student know on day N–1 that the exam was still going to be given. But now we need the stronger premise that she will know on day N–2 that she will know on day N–1 that the exam is still going to be given. This is acceptable in the situation I described, where exams are a rule of the school. But another premise used was that the exam will be a surprise: that the student was not to know on day N–1 that the exam would be given on day N (this is the particular case that we used). For the student to use this on day N–2, she will have to know on day N–2 that she will not know on day N–1 that the examination will be given on day N, that is, that $K_{N-2}(\neg K_{N-1}(E_N))$. She knows this on day 0, from premise (3), but N–2 need not be 0. If we accept premise (9), then since she knows on day 0 that $\neg K_{N-1}(E_N)$, then she must know it on day N–2. Is the principle plausible in this case? Well, what does the student think on day N–2 if the exam has not yet been given? "There is going to be an exam—it's a rule of the school. But is it really going to be a surprise? If I *knew* it were going to be a surprise, then I would know the exam would have to be tomorrow, in which case it would not be a surprise at all. Therefore, I do not *know* that it is going to be a surprise. Maybe the teacher is going to stick to it being a surprise and maybe he is not going to bother and just give the exam on the last day." So, although the student knew at the outset that it was to be a surprise, she may not know it on day N–2; she may still be said to have known this at the outset, provided that it still will be a surprise—that is, that it is given on the (N–1)th day rather than the Nth. Again, it is premise (9) that is causing the fallacy—only now about the surprise element rather than about whether the exam will be given.

Thus we cannot use premise (9) to conclude from what the student knows initially to what she knows on day N–2; we would need an extra premise to say

that the student still knows on day N–2 that the exam will be a surprise. When will this extra premise be plausible? Well, one case is the case where N=2, that is, where the entire class period has only two days. Then premise (9) is not needed, since what the student is supposed to know initially suffices; there is not time for the knowledge to be lost. However, in this case the teacher's announcement has the Moore's paradoxical flavor we already noted in Quine's analysis of the one-day case. This is exactly what happened when I tried the card experiment on a fellow undergraduate, as I have mentioned above. Therefore, as I have already said, the difference between the last day and the penultimate day is not always the crucial one.

What if there are many days? Then to exclude day N–1, one needs another argument: that the student still would know on day N–2 that the exam will be given and will be a surprise (not known in advance). Once again, we could invoke the "rule of the school" device. We can suppose that it is long-settled school policy that the exam must be given on a day when the students do not know that it will be given, even on the day before. And, of course, we also suppose that school policy demands an exam. Given these things, the supposition that the exam will be given on day N–1 will lead to a contradiction of the appropriate premises. This type of idea could be iterated to exclude successive days from the list. The rule of the school will get successively more complicated and involve iterations of knowledge about knowledge, lack of knowledge, and the preservation of the situation.[11]

Let me generalize this argument by describing how the reasoning is iterated. In each case, we conclude that the exam cannot be given on a given day, the Jth day; we then try to rule out the (J–l)th day. To do this we have to assume not only that the previous premises are true but also that the student knows on the previous day, the (J–2)th day, that all these previous premises are true. Only then can the student conclude, on the basis of her knowledge on day J–2, that the exam is not going to be given on days from J onward and so say that it must be given on the (J–l)th day, contradicting the premise that it is a surprise. Thus, we always require not only that the previous premises shall be true and known to be true at the outset but also that they shall remain known to be true on day J–2, whatever day that is, and this is *always* an extra premise, since we have not accepted

[11] One could think vaguely that it is school policy to allow enough of these iterations in each case to generate the paradox. Then there *would* be a vagueness question.

In the original version of this talk, I mentioned the case already discussed in note 5, where the experimenter randomly sticks the card somewhere in the middle of the deck. Then it might seem (at least again thinking vaguely) that enough iterations of the knowledge involved will always be available to rule out successive days from the end. But in the original transcript, I explicitly say that it would be valid but superfluous to use this reasoning to rule out the second-from-the-last card, since the subject can see that the card is not being placed near the bottom. I should have gone on to conclude, as I did in note 5, that this model considerably changes the original problem.

premise (9) as generally true but rather as something that must be argued separately in each case.

Thus where we thought that we were only applying the same reasoning again and again, we were in fact adding tacit extra premises at each stage. The feeling of the heap—that the more days that are involved, the weaker the reasoning becomes—derives from these extra premises piling up, which in fact need special arguments to justify them.

This is all I want to say about this paradox. I do not know if it is really solved; I am sure that there are more things one could say about it.

I would like here to go on to consider the principle that *if you know something now you will know it at any later date.* It was assumed (premise (9)) in the fallacious argument for the paradox above, and it is also assumed by Quine:

> If this [that the decree will not be fulfilled] is a conclusion which he is prepared to accept . . . in the end as a certainty, it is an alternative which he should have been prepared to take into consideration from the beginning as a possibility. (Quine 1953: 20)

That is to say, if she knew in the beginning that the exam will be given, then she cannot, at any later stage, fall into doubt or denial of this: and this is what I am denying. Quine, on the contrary, seems to consider it obvious that if she is willing to accept at a later stage that the exam will not be given, then at no earlier date could she have *known* that the exam will be given.

Again, Hintikka says:

> What exactly is implied in the requirement that the grounds of knowledge in the full sense of the word must be *conclusive*? For our purposes it suffices to point out the following obvious consequence of this requirement: If somebody says "I know that p" in this strong sense of knowledge, he implicitly denies that any further information would have led him to alter his view. He commits himself to the view that he would still persist in saying that he knows that p is true—or at the very least, persists in saying that p is, in fact true—even if he knew more than he now knows. (Hintikka 1962:20–21; emphasis in text)

Of course, in a way, what Hintikka is saying here is obviously true because he says, "He commits himself to the view that he would still persist in saying that he knows that p is true . . . even if he knew more than he now knows." As stated, this is not a substantive principle. "Knows that" could be replaced by any propositional attitude verb, say, "believes that" or even "doubts that," and the resulting principle still would be true—that is, it would still be true that someone would persist in saying that he doubts that p, say, even if he would come to doubt more than he now doubts. But what Hintikka really means here, presumably, is that it is a characteristic of knowledge that even if I have more evidence than I now have, I will still know that p; and this is what I have been denying in giving my counterexamples. You may know something now, but, on the basis of further evidence—without any loss of evidence or forgetfulness—be led to fall into doubt about it later.

Hintikka says that the principle is true only for the "strong sense of knowledge." This implies that there are two senses of the phrase "to know": a strong one and, perhaps, a weak one for which the principle is not really true. There is something in Malcolm (1952) about this too: Malcolm admits that there are cases where you can know something but, later on, on the basis of extra evidence, conclude that you did not know it. He gives the following example: if you *know* that the sun is about 90 million miles from the Earth, you might later, on the basis of learned astronomers saying (perhaps falsely—he is not clear on this point) that an error had been made and that the correct distance was 20 million miles, be persuaded that you were wrong. The astronomers might be saying something false (e.g., if an astronomers' convention had decided to play a trick on the public), but if they were right, then, of course, you did not know it beforehand.[12]

But Malcolm argues that this is not always true, citing another example: suppose there is an ink bottle in front of you, on your desk. Can it be the case that some *later information would lead* you to change your mind? Malcolm writes:

> It could happen that in the next moment the ink-bottle will suddenly vanish from sight; or that I should find myself under a tree in the garden with no ink-bottle about;[13] or that one or more persons should enter this room and declare with apparent sincerity that they see no ink-bottle on this desk. . . . Having admitted that these things *could* happen, am I compelled to admit that if they did happen then it would be proved that there is no ink-bottle here *now*? Not at all! I could say that when my hand seemed to pass through the ink-bottle I should *then* be suffering from hallucination; that if the ink-bottle suddenly vanished it would have miraculously ceased to exist . . .
>
> . . . No future experience or investigation could prove to me that I am mistaken. Therefore, if I were to say "I know that there is an ink-bottle here," I should be using "know" in the strong sense. . . .
>
> In saying that I should regard nothing as evidence that there is no ink-bottle here now, I am not *predicting* what I should do if various astonishing things happened. If other members of my family entered this room and, while looking at the top of this desk, declared with apparent sincerity that they see no ink-bottle, I might fall into a swoon or become mad. I *might* even come to believe that there is not and has not been an ink-bottle here. I cannot foretell with certainty how I should react. But if it is *not* a prediction what is

[12] See Malcolm (1952:184). Since knowledge implies truth, then if Malcolm is really giving an example where knowledge is lost, the announcement by the astronomers must be wrong—after all, you did once know that the sun is about 90 million miles from the Earth. Things would be different if the discussion were one of certainty, justified conviction, or whatever, where these are not construed as implying truth. The same is true of much of my previous discussion in this paper. Everything would be different if knowledge were replaced by a concept that does not imply truth. For example, I wouldn't have had to give an example where I *falsely* convince someone that I never wrote on modal logic. However, other arguments that used the fact that knowledge implies truth would go.

[13] He was in his room at his desk.

> the meaning of my assertion that I should regard nothing as evidence that there is no ink-bottle here?
>
> That assertion describes my *present* attitude towards the statement that here is an ink-bottle. It does not prophesy what my attitude *would* be if various things happened. My present attitude toward that statement is radically different from my present attitude toward those other statements (e.g., that I have a heart).[14] I do *now* admit that certain future occurrences would disprove the latter. Whereas no imaginable future occurrence would be considered by me *now* as proving that there is not an ink-bottle here.
>
> These remarks are not meant to be autobiographical. They are meant to throw light on the common concepts of evidence, proof, and disproof. (Malcolm 1952:185–86; emphasis in text)

He includes the statement "three plus two is five" in the same batch as the ink-bottle case. I am not sure that the ink-bottle case is a good example—a magician might persuade you that you had been tricked.[15]

[14] The statement that he himself has a heart Malcolm supposes to be a statement he knows only in the weak sense since he could be persuaded later of its falsity.

[15] Malcolm is surely right that ordinarily we would regard the presence of an ink bottle in the room as conclusive, not merely probable. Malcolm quotes Ayer to the effect that "no proposition, other than a tautology, can possibly be anything more than a probable hypothesis" (1952:183, note 4). (See also Malcolm's quotations from Descartes and Locke on the same page.) Actually, Hume already states that some empirical statements are not really just probable. He writes: "One would appear ridiculous, who would say that it is only probable the sun will rise to-morrow, or that all men must dye; though it is plain we have no further assurance of these facts, than what experience affords us" (2000: Book 1, Part III, Sec. XI). (However, Hume goes on to reserve the term "knowledge," following previous authors, for a priori knowledge, and uses "proofs" for arguments giving empirically certain knowledge; I take this to be a bit of technical terminology, not really a denial of what Malcolm affirms against Ayer.)

In Malcolm's case of the ink bottle, indeed I may be certain that there is an ink bottle here, but I will be equally certain that the extraordinary future events Malcolm describes will not happen. If I seriously entertain the idea that some of them will or even may happen, then I am entertaining the idea that perhaps a clever magician is deceiving me, or some other even more *outré* case. What Malcolm says does not seem to me to describe my present attitude toward the bizarre possibilities he mentions.

My own intuitions differ from Malcolm's in the opposite direction about some other cases. The statement that I have a heart in the quoted material refers to his claim earlier in the paper that if astonished surgeons told him that when they operated on him they found that he had none, he would believe them, in contrast to the case involving the ink bottle. But I find this belief much harder to give up, even under extraordinary circumstances (I would probably think that the surgeons must be putting me on).

In my discussion of the astronomer's case, I was worried that if the distance from the earth to the sun was *knowledge*, the astronomers must be tricking us (by definition). But let us just speak of what would lead me to give up my *belief* about the distance between the sun and the earth. An extraordinary error such as Malcolm describes would be very hard to swallow without an elaborate explanation. I might wonder whether astronomy is much of a science after all. A slighter error would be better and might be easier to explain. I simply would not believe a committee of astronomers who announced that the earth is flat after all.

Malcolm says at the end of the paper (189) that the ideas in this paper derive from discussion with Ludwig Wittgenstein, something I may not have noticed in 1972. If so, they appear to be based on Wittgenstein's exposition to him of some of the ideas that are now published in *On Certainty* (Wittgenstein 1969). I am not sure, however, that the Wittgenstein of that book would agree with the way Malcolm puts various matters.

I may be in no doubt now as to whether there is an ink-bottle in front of me and yet it seems to me compatible with this to suppose that future evidence could persuade me that there is no ink-bottle. There seem to me to be two different questions here: whether I have the kind of certainty characterized by there being no doubt now, and whether I take the attitude that no future evidence could disprove this. But let this distinction, and the question as to whether this particular example is correct, be set aside—there may be correct examples. The strange—and unargued—thing here is, why does this show that there is a strong sense of "know" for which this is true? Suppose there are some cases of knowledge in which no future evidence will lead me to change my mind, and *other* cases of knowledge in which I would change my mind. That does not show that the word "know" is being used in two *senses* anymore than there being Americans who are rich and Americans who are poor shows that the word "American" is being used in two senses. Any class may, in various interesting ways, divide up into subclasses. Why not instead say that, in general, knowing does not imply that no future evidence would lead me to lose my knowledge, but in some cases, where I do know, it just is in fact the case (and not because of some special sense of "know") that no future evidence would lead me to change my mind?

One would need some additional—say, linguistic—evidence that this shows that "know" is being used in two senses. After all, is it likely that, in Ubangi or Swahili, two different words are used for these two different senses of "know"? There are, of course, different senses of the word "know" in English: those that would be translated as *connaître* as opposed to *savoir*, *kennen* as opposed to *wissen*. These are different senses of knowing: you *know a person* as opposed to *knowing that p is true*. These are indeed different senses of "to know," and this fact is exhibited by the fact that other languages differentiate between them. We all, of course, have heard of the "biblical sense" of "to know," which in English derives from the King James Bible's translation of the corresponding classical Hebrew, and is perhaps similarly distinct in some other languages. But why should there be different kinds of propositional or factual knowledge? Prima facie, it seems to me that the idea that factual-propositional knowledge has two different senses is a red herring. So, what can these people have in mind? Why do they not just say that there are two cases?

I think what they have in mind is this.[16] First, there are obvious principles about knowledge which seem to them to imply that, in general, if you know something, no future evidence could lead you to change your mind—the grounds must be conclusive. But then there are counterexamples; this conclusion does not seem to be correct. So then they argue: well, that must be a weaker sense of "to know." But why not accept counterexamples as counterexamples? Why

[16] Even though there has been a recent resurgence of interest in the question of whether "knows" has different senses, I have chosen not to incorporate the ever-growing recent literature on this subject.

invoke a doctrine of different senses of "know"? Should there really be different dictionary entries? But there must be something behind the idea that "know" has a sense in which knowledge cannot be lost (Hintikka) or at least that our present attitude toward the statement is that the knowledge cannot be lost (Malcolm). This is to be our second paradox. Unlike my treatment of the first paradox, I shall merely state it, and not attempt to solve it—because I discovered it!

I want to try to prove the principle that I earlier declared to be false: that if you know something now, you have got to know it later. One cannot really prove it in such a simple form. You might forget, and so on. But one can try to prove the more careful principle suggested by Malcolm's discussion of the alleged strong sense of "know": that if I know something now, I should, as a rational agent, adopt a resolution not to allow any future evidence to overthrow it. But this does not seem to be our attitude toward statements that we know—nor does it seem to be a rational attitude.

Consider the following. First, the deductive closure of knowledge:

(*i*) If *A* knows that *p* and *A* knows that *p* entails *q*, and, on the basis of such knowledge, *A* concludes that *q*, then *A* knows that *q*.

And then (let "*p*" be any statement):

(*ii*) *p* entails the following hypothetical: any evidence against *p* is misleading (where misleading is to mean *leads to a false conclusion*).

If *p* is true—notice that (*ii*) does not say anything about knowledge—any evidence against it is misleading, that is, leads to the false conclusion that not-*p*. Now, suppose that

(*iii*) The subject *A* knows that *p*, and *A* knows (*ii*).

Then, provided that he carries out the appropriate deduction, it follows from premise (*ii*) that we can conclude:

(*iv*) *A* knows that any evidence against *p* is misleading.

(The statement applies to any evidence, arising now or in the future, but naturally we are most interested in the future.) This already seems very strange: that just by knowing some common or garden-variety statement, which I am calling *p*, one knows a sweeping thing: that any future evidence against *p* will be misleading.

We might have as a general principle something like this (though it is very hard to state in a nice, rigorous way, especially giving it the necessary generality):

(*v*) If *A* knows that taking an action of type *T* leads to consequence *C*, and *A* wishes above all else to avoid *C* (i.e., this is the only relevant issue), then *A* should resolve now not to take any action of type *T*.

This, too, is a very sweeping statement, but we are considering the case where *A* knows, at a certain time, that if he does anything of a certain kind in the future, it will lead to some consequence that he thinks bad, there being no other relevant consequences which would override it. For example, suppose he knows that if he opens the door, someone standing outside is going to shoot him. It would then be a reasonable thing for him to resolve not to open the door.

So, he should resolve not to take any action of type *T*. Let the action of type *T* be accepting evidence against *p*—that is, doubting or denying that *p* on the basis of some future evidence. The consequence *C* is gaining a false belief—or at least losing a true one, if we merely fall into doubt—and this is something that we do not want. Then one may conclude:

(*vi*) *A* should resolve not to be influenced by any evidence against *p*.

To make the argument clearer, notice there are two ways in which one can make this resolution. In the first place, one can resolve not to *look at* any alleged evidence against *p*. For example, I might resolve not to read books of a certain type. I think that, in practice, this is the most important case. It is not possible to keep to such a resolution in the case of the surprise exam (see also note 17), nor does it seem to be what some of the authors I have mentioned have in mind. In the second place, one could conceivably resolve that, if one is faced, regardless of whether one wanted it, with particular evidence against *p*, one should nevertheless ignore it, since one knows that it must be misleading, given that one knows that *p*. Neither of these things seems to be our attitude toward future evidence in cases where we know something. I think it is from such an argument that the idea of a strong sense of "know" may come; that, in some special cases, these conclusions are true. But if you look at the premises and reasoning, there does not seem to be any "super" sense of "know" being supposed, just "know" in the ordinary sense. So there must be something else wrong, and this is the question—what *is* wrong?

Some political or religious leaders have indeed argued along some lines such as those of (*vi*). They have argued on this basis that if their followers or subjects are not strong enough to stick to the resolution themselves, they—the leaders—ought to help them avoid contact with the misleading evidence. For this reason, they have urged or compelled people not to read certain books, writings, and the like. But many people need no compulsion. They avoid reading things, and so on.

If the conclusion of this argument were accepted, our solution to the first paradox would in some sense go by the board, since the student should resolve that, no matter how things appear in the future, she should never lose any of her beliefs in the teacher's announcements. That is a trivial special case of principle (*vi*), but it is the genesis of my considering this second problem and this particular set of premises. And they have their own sort of importance in epistemology. One can be led in two directions by them: first, one can think

that conclusion (*vi*) is correct, and so to know something means that no possible future evidence should lead me to change my mind. But since that is almost never the case, we know almost nothing. This is the skeptical attitude. Alternatively, one might be led to the corresponding dogmatic view—that, since we know all sorts of things, we should now make a resolution not to be swayed by any future evidence.[17]

The commonsense view is, for example, that you *do* know that I have written certain papers on modal logic but that future evidence could lead you to change your mind about this. So, you should rationally leave yourself open to such changings-of-mind, even though it is the case that you *know* that I wrote these papers. The question is, why?[18,19]

APPENDIX I

In a recent class discussion,[20] Fred Michael remarked that Quine's move, at least as to the last day, could be avoided by a simple change in the notion of "surprise." Regard an exam (or hanging) as a surprise if one cannot know in advance of a given day that *if* the event (exam or hanging) will occur at all, the event must occur on that day. Then the question of whether one knows in advance that there will be an exam (a hanging) becomes irrelevant.[21] One could try to complicate the definition of surprise corresponding to the successive elimination of days, but eventually the pileup of extra knowledge assumptions in my main discussion will go over into more and more complicated conditionals with more and more antecedents, and very artificial notions of "surprise."

[17] Although I was in fact led to the second paradox by my consideration of the first (the surprise-exam paradox), in one way taking this case as paradigmatic is misleading. In the case of the surprise exam, the student experiencing the passing of successive days (and perhaps seeing no exam until nearly the end) cannot avoid facing the future evidence, and may not be able to stick to her resolution. The same is not true of what I had in mind for many typical cases of the paradox, where one may avoid contact with the "misleading" counterevidence altogether—for example, by avoiding reading certain books or articles.

The corresponding cases of Malcolm (1952) are at least similar in that it was important to me to phrase the problem in terms of a resolution at the present time, rather than as a prediction as to what I would do. However, in all of Malcolm's cases the idea that one might avoid the misleading counterevidence altogether is not there. (And, as I said above, for some of his cases it is not clear that Malcolm regards the counterevidence as misleading.)

[18] But see also my note 12 above. Since the authors I was concerned with (Hintikka and Malcolm), as well as the previous discussion of the surprise exam, were discussing knowledge, and since knowledge is the subject of the present paper, this second problem has been put in terms of knowledge. But there could be parallel problems for certainty, rationally justified conviction, and so on.

[19] My thanks to the late G. E. M. (Elizabeth) Anscombe and the Cambridge Moral Sciences Club for transcribing this lecture. Thanks to Jeff Buechner, Gary Ostertag, Harold Teichman, and especially to Romina Padró for their help in producing the present version. This paper has been completed with support from the Saul A. Kripke Center at the City University of New York, Graduate Center.

[20] In my spring 2009 seminar at the CUNY Graduate Center.

[21] I now find that a similar remark is made in Ayer (1973:125).

However, we owe to Shaw (1958) the proposal that the teacher's announcement be taken to be *self-referential*: one cannot derive in advance *from the present announcement* when the exam will be given. Done that way, any pileup is avoided, either in knowledge assumptions or in antecedents. The formulation I prefer of this self-referential approach to the surprise exam is in a paper by Frederic B. Fitch (1964).[22, 23] Given enough material to formalize the kind of self-reference involved in Gödel's first incompleteness theorem (say, by quoting syntax into elementary number theory), one can formalize the problem in terms of deducibility and eliminate any notion of knowledge. That is, the announcement (A) can be that the exam will be given on a day such that one cannot deduce from (A) itself, plus the fact that it has not been given on previous days, that it will be given on that very day. Fitch simply accepts that such an announcement, so formulated, leads to a contradiction. The reasoning follows that of the usual surprise exam argument, successively eliminating the days starting from the last. However, he argues that a slightly weaker version is "apparently self-consistent" (1964:163). This modifies the announcement so that "what is intended in practice is not that the surprise event will be a surprise *whenever* it occurs, but only when it occurs on some day *other than the last*" (163). He

[22] Shaw (1958) seems to think that the self-reference makes the problem dubious, like the liar paradox, but Fitch's Gödelian formulation is beyond logical doubt. Shaw is also somewhat sloppy in his formulation. The deduction in advance that is ruled out involves not only the announcement itself, but also that the exam has not been given yet.

[23] See also Kaplan and Montague (1960). Their version is still about knowledge, whereas that of Fitch shows that the notion of knowledge can be eliminated in favor of deduction. Kaplan and Montague have both done justly esteemed and famous work, and the present paper is formally unexceptionable. However, I see some problems in it. For example, is it really one of our "intuitive epistemological principles" that "one cannot know a non-analytic sentence about the future" (81–82)? No wonder one cannot know anything about the exam (hanging) in advance; it will be just as much a surprise as the rising of the sun! (See also note 15 above.)

But the main problem with their paper is this: in order to obtain Gödelian self-reference, they take knowledge as a predicate, rather than an intensional operator (see their excuse at the end of the first paragraph on 80). They then reduce the problem to an analogue of the liar paradox (which they call "the Knower," 88). Basically, it is "my negation is known" (87, formula (1)), from which they derive a contradiction using intuitive principles about knowledge. But then the entire original flavor of the surprise exam (hanging) is lost. One could blame everything on the use of a knowledge predicate. In a subsequent paper, Montague (1963) himself, inspired by this one, argues that if modality is treated as a predicate, customary modal laws cannot be maintained, precisely because of the possibility of Gödelian self-reference.

There is an important difference between the goals of Kaplan–Montague and those of Fitch. Kaplan and Montague are looking for "a genuinely paradoxical decree" (1960:85), that is, one that apparently can be proved to be both true and false. Fitch, on the other hand, avoids the possibility of any paradox almost by the definition of his enterprise. The whole argument can be formalized in first-order arithmetic ("Peano Arithmetic"), or strictly speaking, an extension of that system, by adding finitely many propositional constants corresponding to the $E_1, \ldots, E_N$. By definition, there can be no contradiction as long as Peano Arithmetic is consistent (the added propositional constants change nothing).

Given that we are looking for a self-referential interpretation, I think Fitch's perspective to be better than that of Kaplan and Montague. The problem is not to find a paradox like the liar, but simply a highly counterintuitive conclusion that it is impossible to announce a surprise exam (or decree a surprise hanging) within a certain time limit. One case in Kaplan and Montague (the decrees D_2 and D_3) (82–84), which they consider merely to be a decree that cannot be fulfilled (and thus not genuinely paradoxical), in fact comes closer to capturing the right flavor than the subsequent modification to get a "genuinely paradoxical decree".

includes in the modified announcement that the last day "will be a surprise in the *weak* sense of being not provably implied by the prediction itself" (163).[24]

APPENDIX 2: LETTER TO FITCH

I now append a letter that I wrote to Fitch.[25] Were I to have written it today, I would not have started with Löb's theorem, but with Gödel's second incompleteness theorem, which is the main point.[26] The point is that Fitch's modified announcement implies that various things cannot be derived from it, namely things about when the exam will be given. *A fortiori*, it implies its own consistency. But, by Gödel's second incompleteness theorem, any statement implying its own consistency must itself be inconsistent, contrary to Fitch's guess about his modified announcement.

As I conclude in the letter to Fitch, I do not think that the self-referential interpretation of the problem is a very natural one.

[UCLA]

Los Angeles, California
Department of Philosophy
August 4, 1972

Prof. Frederick Fitch
Department of Philosophy
Yale University
New Haven, Connecticut 06520

Dear Fred:
You may be interested in the following observation related to your article on the Prediction Paradox in the *APQ* for April 1964. The statement in (16) [and the one in (17) also] which you call "apparently self-consistent" is actually refutable. The reason: call the statement "P." P implies ~Bew#~P, since if Bew#~P then, of course, Bew#[P ⊃ Q_1] and Bew#[P ⊃ Q_2]. So P ⊃ ~Bew#~P is provable. So Bew# ~P ⊃ ~P is provable. By the famous theorem of Löb (*viz*, if Bew#A ⊃ A is provable, so is A; *JSL* 1955 pp. 115–8), ~P is provable [Löb 1955].

The situation should really be looked at this way: let P be any statement which implies that it itself is not refutable. Then, Z plus P, where Z is elementary number theory, is a system which can prove its own consistency, and therefore is inconsistent by Gödel's Second Incompleteness Theorem. Hence, P must be refutable in Z. Your statement P is a statement

[24] See his formulae (16) (for two days) and (17) (for three days). (There appears to me to be a subtle error involving the use of exclusive "or" in the formulation of (17) that I hope to discuss elsewhere, not particularly in connection with Fitch's paper.)

[25] Fitch replied to this letter simply saying that he could find nothing wrong with it.

[26] As I mentioned in the letter, I have shown that Löb's theorem, proved by Löb in a different way, is actually a simple corollary of Gödel's second incompleteness theorem. See, for example, Smullyan (1992:110), and Boolos and Jeffrey (1980: ch. 16). In the letter to Fitch, I actually give the argument, and maybe this is why I mention Löb's theorem.

which implies its own irrefutability, since it implies that various things cannot be deduced from it, and is therefore refutable. The same argument easily establishes that Löb's theorem, cited above, is a simple consequence of Gödel's Second Incompleteness Theorem, as I observed in a short unpublished paper. Simply take P in the preceding paragraph to be ~A.

The same observations show that much of what Bennett says in his review (*JSL* 1965) of your paper and others is wrong. Some of the statements which he declares to be self-consistent or even logically true can be shown to be refutable by the same argument, at least if they are interpreted in terms of Gödelian self-reference.

Since any statement whatsoever which implies its own self-consistency (and therefore any statement which implies that something else cannot be deduced from it) is refutable, it seems to me that interpretations of the Prediction Paradox in terms of Gödelian self-reference and deducibility in number theory do not really capture the spirit of the original paradox. For, surely the original paradox was not meant to follow from such general considerations. My own view of the paradox is different, but that's another story.

Best,
Saul Kripke

APPENDIX 3: COMMENTS ON THE SECOND PARADOX

There has been a considerable secondary literature on my second paradox, which has come to be known as the dogmatism paradox. As I said, it can be discussed for other epistemic concepts such as certainty, rational belief, and so on; and indeed there is literature on some other forms. Moreover, one might distinguish between first and third person formulations of the problem, and in the latter case, whether it is a subject who really knows or merely *thinks* he or she knows. Here I am discussing a subject *S* who genuinely knows that *p*.

In the published literature the first discussion of my second paradox is in Gilbert Harman's well-known book *Thought* (Harman 1973:147ff), and I emphasize this version. I hope I have understood Harman correctly. To discuss his treatment of the problem, one should remember that my point is about a resolution *made in advance* to ignore certain types of evidence. Mostly the strategy followed is that of failing to read literature of a certain type, and so on. One might make a resolution to ignore particular evidence even when one is forcibly confronted with it, but this is often more difficult to keep.[27]

People who follow these strategies are, after all, not uncommon at all, as I have mentioned in the paper. Often, however, we think of them for this reason as dogmatists who do not really know. Here, however, the premise was that we are dealing with a subject who really knows. We are arguing that such a subject ought to maintain the dogmatic attitude because any counterevidence really is misleading. (In my own discussion, I imagine trying to convince people, falsely, that I never worked on modal logic after all, so that some prior knowledge would be lost.)

[27] Those who followed the main strategy might be compared to what Ulysses would have done if he had decided to put wax in his ears or follow another route. Tying oneself to the mast is something of an analogue of the second strategy some might follow if forced to confront the evidence (the sirens). Think also of cases of trying to avoid addictive substances (I have heard the comparison made between dangerous—misleading—books and dangerous drugs) and what an addict may warn his friends about his behavior before he tries to withdraw.

Harman's discussion goes like this, where we replace his particular example with the letter "*p*." He says, "Since I now know that *p*, I now know that any evidence that appears to indicate something else is misleading. That does not warrant me in simply disregarding any further evidence, since getting that further evidence can change what I know. In particular, after I get such further evidence I may no longer know that it is misleading. For having the new evidence can make it true that I no longer know that *p*; if I no longer know that, I no longer know that the new evidence is misleading" (1973:148–49).

Well, one need not disagree with what Harman says about the acquisition of the new evidence (at least for typical cases). But remember that I was talking about a resolution to be made in advance. Just because the subject wishes to avoid a loss of knowledge such as Harman describes, so for that reason she or he makes the resolution. Usually, this resolution is to avoid certain types of contact with alleged evidence, such as reading the wrong books (for they can contain nothing but sophistry and illusion), associating with the wrong people, and so on. Moreover, by hypothesis, the books and so on, *are* misleading, and the subject knows they are.

One should certainly construe the resolution to include a more specific form, avoiding contact with some specific counterevidence to *p*, though usually one will not know of it. Harman is right to say that if such contact nevertheless occurs, one may well lose the knowledge that *p*, and hence no longer know that the counterevidence is misleading. But just this is why the subject resolves not to get into such a situation![28]

I should add something I was probably not aware of when I originally gave this lecture, which is that sometimes the dogmatic strategy is a rational one. I myself have not read much defending astrology, necromancy, and the like (I remember Stephen Weinberg making the same remark). Even when confronted with specific alleged evidence, I have sometimes ignored it although I did not know how to refute it. I once read part of a piece of writing by a reasonably well-known person defending astrology.[29] Another time, I saw an advertisement professing to prove that Vincent Foster had been murdered (presumably on orders of the Clintons, though this was not stated explicitly). I was not in a position to refute specific claims but assumed that this was a piece of no value.[30, 31] One epistemological problem is to delineate the cases when the dogmatic attitude is justified.

[28] I am assuming that we are dealing with a subject who wishes to avoid losing knowledge. Sometimes there are people who "don't want to know" or do wish to lose knowledge that they have, sometimes for arguably good reasons. They are not in question here.

[29] What I had glanced at was a piece by Hans Eysenck professing to prove the theories of a particular French astrologer.

[30] My reaction was amply confirmed by several investigations, including even one headed by Kenneth Starr. If someone doesn't like this example, try, say, holocaust denial.

[31] It is a merit of Robert Nozick's discussion of the problem (1981:237–39) that he recognizes that there are cases where alleged evidence for dubious or crackpot views contrary to what we know may be ignored (239). The rest of his discussion, in this respect like Harman's, imagines us confronted with a particular piece of evidence *e*, purporting to undermine, or even contradict, the knowledge that *p*. Nozick bases his rather detailed discussion of this matter on his view that knowledge fails to be closed under Universal Instantiation (UI). Even though *S* may know that all evidence against *p* is misleading, according to Nozick that does not show that *S* knows that some specific evidence *e* against *p* is misleading, and therefore can be disregarded. He then adds a subtle discussion of this particular case. Many (including me) might find Nozick's rejection of the closure of knowledge under UI in and of itself implausible. But I have discussed Nozick's theory in some detail in the present volume and hope to be pardoned for omitting any further discussion of its application to this case. (For more discussion of Nozick's views on deductive closure, see Chapter 7.) David Lewis also developed an epistemological theory stating that whether a subject knows that

REFERENCES

Ayer, A. J. (1973). "On a Supposed Antinomy." *Mind* 82:125–26.
Bennett, J. (1965). "Review of R. Shaw, *The Paradox of the Unexpected Examination.*" *Journal of Symbolic Logic* 30:101–12.
Boolos, G., and R. Jeffrey (1980). *Computability and Logic*, 2nd ed. Cambridge: Cambridge University Press.
Dretske, F. (1970). "Epistemic Operators." *Journal of Philosophy* 67:1007–23.
———. (1971). "Conclusive Reasons." *Australasian Journal of Philosophy* 49:1–22.
Fitch, F. (1964). "A Goedelized Formulation of the Prediction Paradox." *American Philosophical Quarterly* 1:161–64.
Harman, G. (1973). *Thought.* Princeton: Princeton University Press.
Hawthorne, J. (2004). *Knowledge and Lotteries.* Oxford: Oxford University Press.
Hintikka, J. (1962). *Knowledge and Belief: An Introduction to the Logic of the Two Notions.* Ithaca, NY: Cornell University Press.
Hume, D. (2000). *Treatise of Human Nature.* Ed. David Fate Norton. Oxford: Oxford University Press. First published 1740.
Kaplan, D., and R. Montague (1960). "A Paradox Regained." *Notre Dame Journal of Formal Logic* 1:79–90.
Lewis, D. (1996). "Elusive Knowledge." *Australasian Journal of Philosophy* 74:549–67.
Löb, M. H. (1955). "Solution to a Problem of Leon Henkin." *Journal of Symbolic Logic* 20:115–18.
Malcolm, N. (1952). "Knowledge and Belief." *Mind* 61:178–89.
Montague, R. (1963). "Syntactical Treatments of Modality, with Corollaries on Reflexion Principles and Finite Axiomatizability." *Acta Philosophica Fennica* 16:153–67; reprinted in Montague (1974).
———. (1974). *Formal Philosophy: Selected Papers of Richard Montague.* Ed. R. Thomason. New Haven, CT: Yale University Press.
Nozick, R. (1981). *Philosophical Explanations.* Cambridge, MA: Harvard University Press.
Quine, W. V. O. (1953). "On a So-Called Paradox." *Mind* 62:65–67. Reprinted under the title "On a Supposed Antinomy," in *The Ways of Paradox and other Essays.* Cambridge, MA: Harvard University Press, 1966, 19–21; page references are to the reprint.
Shaw, R. (1958). "The Paradox of the Unexpected Examination." *Mind* 67:382–84.
Smullyan, R. (1992). *Gödel's Incompleteness Theorems.* Oxford: Oxford University Press.
Sorensen, R. (1988a). *Blindspots.* Oxford: Clarendon.

---

p can depend on the conversational context, notably whether traditional skeptical philosophical doubts have been brought into the discussion (see Lewis 1996). In the case of the present paradox, he thinks that merely hearing that there is purported counterevidence is enough to create a context in which one no longer knows that *p*, and hence that such counterevidence is misleading. (It is unclear to me whether Lewis means that we are aware of what the purported counterevidence is, or aware merely of its existence.) We have seen that this is in general too strong, but I think Lewis is probably not really unaware of this. I can't go into Lewis's contextual theory here, though my inclination is not to agree. Other philosophers, such as Hawthorne (2004) and Sorensen (1988b), have discussed the problem. I will not discuss their views here.

———. (1988b). "Dogmatism, Junk Knowledge, and Conditionals." *Philosophical Quarterly* 38:433–54.
Williamson, T. (1994). *Vagueness.* London: Routledge.
Wittgenstein, L. (1969). *On Certainty.* Ed. G. E. M. Anscombe and G. H. von Wright. Trans. G. E. M. Anscombe and D. Paul. Oxford: Blackwell.