Doris Olin,
Paradox
Montreal 2003, McGill-Queen's
University Press

3 Believing in surprises: the prediction paradox

I think that this flavour of logic refuted by the world makes the paradox rather fascinating. The logician goes pathetically through the motions that have always worked the spell before, but somehow the monster, Reality, has missed the point and advances still.[1]

The paradox

A teacher announces to her student *S* that she will give him exactly one examination during the next week, and it will be a surprise: *S* will not be able to predict, prior to the day of the examination, on which day it will be held.[2] The student, a star logician, objects that this is impossible. He argues as follows. "If the exam were held on Friday, then on Thursday evening, realizing that no examination had yet been given, I would reasonably expect it on Friday; hence a Friday examination would not be a surprise. But, if the examination were given on Thursday, then on Wednesday evening I would be aware that no examination had yet been given and, recognizing that it cannot be given on Friday, would expect it on Thursday; so a Thursday examination would not be a surprise. Similarly for the remaining days. Consequently, the surprise examination cannot be given – you cannot do what you said you would do."

The teacher, visibly shaken, declines to answer, and cancels the class in order to think. On Tuesday of the next week, she presents the student with an examination whose first question is "Is this test

a surprise?" Grudgingly, *S* recognizes that he must answer in the affirmative.

This is the paradox of the surprise examination: there is an apparently impeccable argument for a conclusion that seems patently false. Clearly the teacher can give a surprise examination on, for instance, Tuesday.

Variations of the paradox abound. One version involves a sadistic judge who sentences a man to be hanged, but adds a twist concerning the date. He is to be hanged, the judge decrees, on one of the three following days at noon, and he will not know beforehand which day it will be. The condemned man reasons as follows. "If I am to be hanged on the third day, then on the evening of the second day, I would expect the hanging on the third day; so the decree cannot be fulfilled on the third day. If the judge has in mind the second day, then on the evening of the first day I would realize the sentence cannot be carried out on the third day, and thus would expect it on the second day. And so on. So the judge's decree cannot be carried out." The prisoner is at peace until the hangman arrives on the morning of the second day.

Another variation has a friendly philosophy student carefully arrange a deck of cards, and then announce that he will turn over the cards one by one, showing you the face, until he arrives at the jack of spades; you will not be able to predict when that card will appear, he says, before actually seeing it. You think to yourself that he cannot leave the jack of spades to the end of the deck, because then its appearance would not be a surprise. Nor can he leave it to the second to last card, since after 50 cards have been turned over, knowing that it cannot be the last card, you would expect it to appear as the second to last card. Before long, you realize that it is possible to continue in this way until all 52 cards have been eliminated.

These are all versions of what has come to be known as "the prediction paradox". One rather delightful feature of the paradox is that it appears to have had its origin in a historical event. Sometime during 1943–44, it was announced on Swedish radio that a civil defence exercise would take place one day of the following week, and that in order to provide a proper test of the civil defence system, no one would be able to predict the day of the test in advance. A Swedish mathematician, Lennart Ekbom, is apparently

to be credited with having first detected the paradox lurking behind the announcement.[3]

How does the prediction paradox fit into the classification scheme developed in the first chapter? Clearly, it is a type I paradox. It is also apparent that the paradox is falsidical, since a surprise examination, or a surprise hanging, is surely possible in the circumstances described. Finally, the paradox is uncontroversial: there is virtually complete agreement that the conclusion of the argument is false. So there is a general consensus that to dissolve the paradox, it is necessary somehow to defuse the paradoxical argument, to reveal the fallacy or error. In particular, it is pointless to construct further arguments that the surprise examination is possible; that is something we already know. It is also worth noting that the paradox-generating argument is one we all find seductive; so we are likely to be off the mark if our solution consists in isolating and refuting a premise that is initially implausible or highly controversial.

Before considering possible solutions, it will be helpful to get more clarity on the paradox itself. Note first that the announcing of the surprise examination is crucial to the problem, for it must be plausible to suppose, at certain points in the argument, that the student has good reason to believe that a surprise examination will be given; and the sole reason to suppose this is that the teacher, who is generally reliable, has said so. Epistemic concepts such as "has good reason to believe" and "is entitled to believe" seem central to the paradox. In fact, the relevant issue, in determining whether the exam will be a surprise, is what the student is *justified* or *warranted* in believing. That is, to say that the examination will be a surprise is to say that the student will not be justified in believing, before the day of the examination, that the examination will occur on that day.

It might be suggested, with some plausibility, that the notion of surprise can be unpacked in terms of knowledge: the student will not *know* the day of the examination in advance. An adequate solution to the paradox should, I believe, be applicable to *either* interpretation of the central concept. However, the interpretation in terms of justified belief is the minimal one, given that knowledge implies justified belief. Further, the *reason* the student apparently does not know the day of the examination in advance is that he is not justified, before that day, in believing that it will be on that day. Thus the interpretation in terms of justified belief enables us to

focus on the key issue, rather than get tangled up in irrelevant questions concerning the more complex concept of knowledge.

The details of the paradoxical situation must also be sharpened, and made more explicit. The teacher, let us suppose, said to *S*, "An examination will be held on exactly one of the days Monday to Friday; and if an examination is held on day D, you will not be justified in believing this before that day." Now, if the student's argument is not to be open to trivial objections, at least the following assumptions concerning *S*'s memory, reasoning powers and evidence are needed.

(A_1) The student is an expert logician: if he is justified in believing $P_1, \ldots, P_n$, which jointly imply (or strongly confirm) Q, then he sees that $P_1, \ldots, P_n$ jointly imply (or strongly confirm) Q.

(A_2) On Sunday evening, and throughout the next week, the student remembers what the teacher said, and also remembers that she is generally reliable and trustworthy.

(A_3) On Sunday evening, and on any evening of the week, the student knows what evening it is and, on any evening of the week, he remembers whether an examination has been held on that or any previous day of the week.

(A_4) Throughout the week, the student has no source of evidence relevant to the teacher's announcement other than that given by (A_2) and (A_3).

With this background understood, the steps of the paradox-generating argument can be stated as follows:

(1) If the only examination of the week is held on Friday, then on Thursday evening the student will be justified in believing that an examination will be held on Friday.

(2) If the only examination of the week is held on Thursday, then on Wednesday evening the student will be justified in believing (1), and therefore also justified in believing that the examination will be on Thursday.

And so on. Eventually, we reach the conclusion that the examination cannot be given. Notice that the argument is expressed in the third person, not the first. This helps us to keep in mind that the argument can be worked through by a bystander who happens to hear the announcement addressed to the student, as well as by the student.

At this point, ideas for a solution may be percolating in the reader's mind. Let me anticipate some initially appealing responses.

Response I

It is often maintained that the solution to the paradox lies in recognizing that the student has presented a compelling argument that no surprise exam can be given, and therefore cannot believe the teacher's announcement. So the student has no reason to expect any examination next week, and a surprise examination can therefore be held on any day of the week.

This, however, is hardly a solution. For one thing, as was noted earlier, what is wanted as a solution is not an argument or proof that the surprise examination can be given; that is something we already know. Even more important, this line of thought requires us to "give in" to the paradox-generating argument – to grant that it is a good argument. But no coherent solution can find the fallacy in the student's argument while granting that the argument is sound.[4]

Response II

Another initially appealing response is that the flaw in the argument has something to do with temporal order, and with the fact that the argument moves backwards in time. True enough, the argument must proceed by first eliminating Friday, then moving back to Thursday, and so on. But is there anything illegitimate about this? The intuition that the temporal direction of the argument is critical, that something akin to time travel is going on, might be defended as follows. In the first step of the argument, the student assumes that he already knows, when working through the argument, that no exam has been held on the first four days of the week. But this is something he couldn't know until Thursday night. Hence, he begins with an assumption of knowledge to which he is not entitled.

Here there is simply a misunderstanding of the argument. The argument begins with "If the examination is held on Friday, then on Thursday evening, . . .". The statement that the examination is held

on Friday is an assumption only in the sense that it is taken as the antecedent of a conditional that is affirmed. But to affirm "If *P*, then ..." you do not need to know that *P*.

That the paradox does not essentially involve reasoning backwards in time can also be demonstrated by considering a rather ingenious variation.[5] Of five students, Art, Bob, Carl, Don and Eric, one is to be given an examination. The students are lined up in alphabetical order, so that each can see the backs of those before him. The teacher has four silver stars and one gold star which will be placed on the students' backs; the gold star designates the student who will be examined. The teacher tells the students this, and also informs them that the designated student will not be entitled to believe that he is the designated student until after the students break formation.

The students then generate the following argument. If Eric is the designated student, then he will see four silver stars ahead of him and will thus be able to infer that he is the designated student; so he cannot be the designated student. But if Don is the designated student, then since he will see three silver stars ahead of him, and will realize that Eric is not the designated student, he will be entitled to believe that he is the designated student; so he cannot be the designated student. And so on. The students conclude that the specified examination cannot be given. They then break formation, and Carl is surprised to learn that he is the designated student.

This variation suffices to show that *any* analysis that focuses on the temporal factor will not be a comprehensive solution.[6]

Response III

It is tempting to view the paradox as resulting from a misunderstanding of the teacher's announcement that the exam will be a surprise. It has been suggested, for instance, that the surprise component of the teacher's assertion should be understood as implicitly qualified, as claiming that the exam will be a surprise *unless it takes place on the last day.*[7] Construed this way, the announcement certainly does not support the paradoxical argument. For in the very first step of the argument, we can say that if the exam is held on Friday, it will not be a surprise; but this now gives us no reason to rule out Friday as a possible day for the exam. This sort of

reinterpretation of the announcement is unsatisfying, however, because although the teacher might have intended the qualified assertion, there seems no reason to suppose she could not have intended the unqualified assertion. And if she is understood as intending the latter, it still seems possible that the exam should be a surprise. So the original paradox remains.

Another reinterpretation has it that the surprise clause of the announcement should be construed as saying that the student will not know the day of the exam *before the week begins.*[8] Again, no paradox will result on this interpretation. But again there seems no reason why the teacher could not intend the stronger claim that the students will not know the day of the examination *at any time before* that day. Nor does understanding the announcement in terms of this stronger claim seem to preclude the possibility of there being a surprise exam.

Since the paradox proves resistant to these attempts at a "quick fix", we turn now to an examination of the more prominent approaches to the paradox found in the philosophical literature. Initially, the paradox may appear to be a mere puzzle, a brainteaser, that can be disposed of in a few pages. But decades of controversy and a voluminous, still growing, literature suggest otherwise. In fact, I shall argue, this is a paradox of some depth, with much to teach us about familiar epistemic principles.

Quine's contribution

An influential early contribution by Quine regards the problem as not particularly taxing; it is remarkable, he says, that the solution to the puzzle is seldom clearly apprehended.[9] The problem occurs, according to Quine, at the very first step of the argument. The student looks ahead to Thursday evening, and discerns just two possibilities: (a) the exam will have already occurred; or (b) the exam will occur on Friday, and the student will be aware of this on Thursday evening (in which case the teacher's announcement will be false). The student rejects (b) on the grounds that it falsifies the announcement, and opts for (a), thus beginning the process of whittling down the possible days till none remain. But, says Quine, the student should have discerned not two possibilities for Thursday evening, but four. Apart from (a) and (b), the student should also

consider: (c) the exam will fail to occur on Friday, thereby falsifying the announcement; and (d) the exam will occur on Friday and the student will not know this before Friday. If each of these is recognized as a possibility, then the student will find the path to eliminating Friday blocked.

What Quine seems to mean, in speaking of each of the four situations as a possibility, is that, for all the student knows, any one of them might obtain; that is, he does not know, of any one of them, that it does not obtain. But only two of them, (a) and (d), are compatible with the teacher's announcement. If the student does not know that the other two do not obtain, this must mean that he does not know the truth of the teacher's announcement. So the core of Quine's analysis is that the student does not know the truth of what the teacher has said.

But the teacher is a reliable and trustworthy person, and she is in a position to know the statement she asserts. Why should the student not believe her? Why does he not know the truth of what she says? Quine has nothing to say on this score. True, the announcement is a statement about the future; true also that the only grounds the student has for it is the testimony of the teacher. But surely we (and Quine) do not want to deny all knowledge or warranted belief about the future (do I not know that my desk will be in my office when I arrive in the morning?); and testimony provides the basis for great quantities of what we normally claim to know. We can readily grant that if Quine is right, the paradoxical argument collapses. However, in the absence of any reason to deny the student knowledge that, on the face of it, he would seem to have, Quine does not appear to have laid the paradox to rest.

The logical approach

The single most popular approach to the paradox, over the decades since it first surfaced, has been to construe it in purely logical terms by interpreting the teacher's statement in terms of deducibility. The origin of the paradox, it is then argued, lies in self-reference, and it thus bears some resemblance to the liar paradox.

The core of this approach was first presented in a seminal paper by R. Shaw.[10] Shaw's version of the paradox has it that the students are told, at the end of term, that it is an unbreakable rule of the

school that an examination will be given on an unexpected day of the next term. To say that the day of the examination will be a surprise, Shaw insists, is to say that it is *not deducible from the rules of the school* (in conjunction with background information concerning whether or not an examination has yet been given).

How exactly should the rules of the school be stated? Suppose we try:

> Rule 1: An examination will take place on one day of the next term.
>
> Rule 2: The examination will be unexpected, in the sense that it will take place on a day such that on the previous evening it will not be possible for the students to deduce *from Rule 1* that the examination will be on the next day.

Given these two rules, the last day of term is eliminated, since it would violate Rule 2. But any other day will satisfy the rules. An attempt to run the paradoxical argument will succeed only at the first step, if this is how the rules are interpreted. But suppose now that we add a third rule:

> Rule 3: The examination will take place on a day such that on the previous evening it will not be possible for the pupils to deduce *from Rules 1 and 2* that the examination will take place on the next day.

On this understanding of the rules, the last two days of the term can be eliminated by the student's argument. Consider the situation the evening before the second to last day. The examination has to be on one of the next two days, by Rule 1. By Rule 2, it can be deduced that it will not be on the last day. Hence, by Rules 1 and 2, it follows that it will be on the second to last day. But this deduction means that Rule 3 would be violated if the examination were on the second to last day. The last two days, then, are not possible given these three rules; but any other day of the term is possible. In general, it will take $n + 1$ rules of this sort to eliminate the last n days of term.

Provided that the teacher's statement is understood in terms of these rules, and that there are at least as many days in the term as

there are rules, the paradoxical argument will be stopped before every day is eliminated. Thus there will be days on which the surprise examination is possible.

The paradox-generating argument, Shaw suggests, is seductive because we interpret the teacher's announcement in terms of Rule 1 and the following:

> Rule 2*: The examination will take place on a day such that on the previous evening the pupils will not be able to deduce from *Rules 1 and 2** that the examination will take place on the next day.

Rule 2* is self-referential (like the sentence of the liar paradox), and clearly does imply a contradiction. If we rely on this interpretation, the paradoxical argument will unquestionably go through; but this should not be troubling, for self-referential sentences are widely regarded as illegitimate or defective. The key to resolving the paradox, according to Shaw, is to recognize that we are interpreting the teacher's statement as the self-referential Rule 2* when we work through the argument; but in judging that the exam is possible, we interpret the teacher as asserting something like Rules 1 and 2.

The difficulty with this general approach, however, is that it does not go to the heart of the paradox. More specifically, the paradox does not arise from construing the announcement as Rules 1 and 2*. Notice that the teacher, as well as the student, can work through the paradoxical argument. But she can do so only if she assumes that the surprise exam has been previously announced to the student. Without this tacit premise, the argument cannot even begin. But from 1 and 2* it can be deduced that the exam cannot be given, without having to make *any other assumptions*. So the teacher's having announced the exam plays a role in the paradox for which this approach has no room. The teacher's announcement is not being interpreted as 1 and 2* in the paradoxical argument.

The (KK) thesis

The argument of the prediction paradox requires, as we have already noted, certain assumptions about the student's cognitive abilities and situation. Call these the factual assumptions. But there

are also philosophical assumptions at work in the background that have not yet been brought to light.

One influential approach to the paradox has it that the paradoxical argument makes essential use of the philosophically controversial (KK) thesis, which says:

(KK) If S knows that p, then S knows that S knows that p.

The (KK) thesis apparently licenses unlimited iterations of knowledge. But according to this approach, (KK) is not viable, and the flaw in the argument is thus its reliance on this principle.

To begin, let us try to understand why (KK) seems necessary as an assumption.[11] As noted earlier, the paradox, and the notion of surprise, may be expressed either in terms of knowledge, or in terms of justified belief. Those who focus on (KK) take knowledge to be the central concept of the paradox. But an adequate solution should apply to both the knowledge and the justified belief versions of the paradox. The analogue to (KK), which would presumably be essential for justified belief versions of the paradox, is:

(JJ) If S is justified in believing that p, then S is justified in believing that S is justified in believing that p.

Most of what follows will apply, with appropriate changes, to (JJ).

The argument, couched in terms of knowledge, begins with:

(1) If the only examination of the week is held on Friday, then on Thursday evening the student will know that an examination will be held on Friday.

The reasoning underlying this step is that the student knows, after the announcement, that there will be an exam during the week and retains this knowledge throughout the week, and also knows at any point in the week whether an exam has yet been given. At this stage, there is no need for the (KK) thesis. But now consider the next two steps:

(2) If the only examination of the week is held on Thursday, then on Wednesday evening the student will know (1), and thus know that the examination will be held on Thursday.

(3) If the only examination of the week is held on Wednesday, then on Tuesday evening, the student will know (1) and (2), and thus know that the examination will be held on Wednesday.

The pattern is clear. Each step of the argument, after the first, requires that the student have knowledge, at the appropriate times, of the preceding steps.

It is here that (KK) is thought to play a role. Step (1) rests, in part, on:

(a) The student knows on Sunday evening that there will be exactly one examination during the week.
(b) The student retains this knowledge throughout the week.
(c) The student knows, on every evening, what day of the week it is and whether an examination has yet been given.

In order to know (1) on Sunday evening, then, the student must know (a). That is, it must be the case that:

(a*) The student knows on Sunday evening that he knows on Sunday evening that there will be exactly one examination during the week.

To ensure the truth of (a*), it is argued, we must appeal to (KK).[12]

Two philosophers whose diagnosis centres on the role of (KK) are James McLelland and Charles Chihara.[13] They grant that in the intuitive, unformalized version of the paradox, no explicit appeal is made to (KK). Rather, we reason from assumptions concerning what the students know in the situation to a statement P (A Friday exam will not be a surprise). Since the students, we think, can deduce whatever we can, we then attribute to them *knowledge* of P. But this will follow only if they also *know* the premises concerning what they know in the situation. Thus, we are in effect reasoning in accord with (KK).

McLelland and Chihara attempt to refute (KK), and thereby show that the paradoxical argument rests on a false premise. One criticism they advance is that (KK) implies that if we know P, then we can disregard any evidence that would indicate that we do not know P. For if we know that we know that P, then we know that any

such evidence is misleading (is evidence for something false), and thus may reasonably be disregarded. But this is rarely true in the ordinary situations in which one claims to know. The (KK) thesis would thus set the standards for knowing very high: it would require something like conclusive evidence.

If successful, however, this sort of reasoning tells not just against (KK). It is essentially the reasoning of Harman's paradox of dogmatism, which makes no reference to (KK).[14] If P is true, then any evidence against P is misleading evidence. If I know that P is true, then I know that any further evidence I may encounter against P is misleading, and I may therefore disregard it. So, once I know that P is true, I am in a position to disregard any future counter-evidence against P.

The fallacy in this sort of reasoning emerges clearly once explicit reference to time is introduced. If I know at time t that P, then I know at t that any evidence against P is misleading. However, if at a later time t_1 I acquire evidence E against P, I may well *not know at* t_1 that E is misleading. For, given my new body of evidence, I may not be justified in believing P at t_1, and thus may not know at t_1 that P. Knowledge and justified belief may shrink, as well as grow, with the acquisition of new evidence.

There are, in any case, several reasons to suppose that refuting (KK) (or (JJ)) will not suffice to resolve the paradox. First, the paradoxical argument does not require (KK) in its full generality; (KK) is a premise far stronger than necessary. In the argument, Friday is first eliminated and then it is assumed that *the student* could also eliminate Friday. This obviously does not follow, given just the factual assumptions specified earlier. There is a gap in reasoning that could be filled by (KK). But much less than this is required.

Consider a three-day version of the paradox. The first step requires no iteration of the student's knowledge (or justified belief). It assumes that the student knows that P (There will be exactly one exam in the three-day period). The second step requires that the student know that he knows that P (one iteration); the third step requires two iterations. All that is necessary, then, is that the student have the kind of epistemic self-awareness that would permit two iterations of his knowledge. And it surely seems *possible* that the student, who is credited with superb memory and logical skills, in general, with all the intellectual assets of an ideal knower, should have this level of epistemic self-knowledge.

In short, to stop the argument in the three-day case, one would have to show that it is *impossible* for the student to have two iterations of knowledge. Merely showing that (KK) is not, in its full generality, true, is inadequate.

The second reason for deeming (KK) irrelevant to the prediction paradox is that there are variations that do not require any iterations of the subject's knowledge (or justified belief). Consider again the designated student variation. The first step of the argument in this variation assumes that Eric knows that exactly one of the five students will be given an exam (*P*). The second step, to eliminate Don, requires that *Don* knows that *Eric* knows that *P*. (KK) has no bearing on this sort of iteration, which involves a change of subject (*D* knows that *E* knows that *P*), or change of cognitive viewpoint.

Clearly, (KK) is not directly relevant to this variation. Still, it has recently been maintained by Timothy Williamson that no sufficiently lengthy iteration of knowledge, including those *that involve a change in cognitive viewpoint*, can be true; and that this impossibility provides the basis of a solution to the prediction paradox.[15] However, although this broader approach is more likely to be relevant to the designated student variation, there is yet another ingenious variation that escapes even Williamson's wider net.

Consider the sacrificial virgin paradox.[16] The inhabitants of a tropical island observe an annual ritual of sacrificing a virgin to the local volcano. A number of virgins are blindfolded and brought before the volcano. They all hold hands in a line and can only communicate the statement: "No one to your right is a sacrificial virgin". This is done by squeezing the hand of the virgin to one's left. The virgins are logically skilled and reliable, and will give the signal if and only if they know the truth of what is communicated. The chief takes the leftmost virgin to the mouth of the volcano and, if the offering is acceptable, sacrifices her and sends the others home. If not, he tries again with the new leftmost virgin. The virgins are informed of all of this, and also told that the sacrificial virgin will not know she is the sacrifice before being tossed in.

A visitor to the island objects that the ceremony cannot take place. Any virgin is either (i) the rightmost, (ii) a middle or (iii) the leftmost virgin. (i) If the sacrificial virgin is the rightmost, then she realizes she is the rightmost since her right hand is free. Thus if she

is offered, she goes to the volcano knowing she is the last alternative, and therefore is able to infer that she is the sacrifice. So she cannot be the sacrificial virgin. Realizing this, the rightmost virgin will signal by squeezing the hand of the virgin on her left, who is either a middle or the leftmost virgin. (ii) If the virgin to the immediate left of the rightmost virgin is a middle virgin, then, if she is offered, she is aware beforehand that no one to her left has been sacrificed. And, since she has received the signal on her right hand, she is entitled to infer that she is the sacrifice. So she cannot be the sacrifice. Realizing this, she squeezes the hand on her left. Similarly for all the middle virgins. (iii) If the sacrificial virgin is the leftmost virgin, then, since she has received the signal, she realizes she is the only remaining virgin and is therefore the sacrifice. So she cannot be the sacrifice and the ceremony is impossible.

What degree of iteration of knowledge is required in the sacrificial virgin paradox? The visitor to the island first argues that the rightmost virgin cannot be the sacrifice. His argument is based, in part, on assumptions that she knows certain facts about the ceremony and also knows that her right hand is free. But he then goes on to argue that *she knows that she is not the sacrifice*. Hence, he has to credit her with *knowing that she knows* the relevant facts. Thus one iteration of knowledge is required for the rightmost virgin. Similarly for any middle virgin. She must know that she cannot be the sacrifice, hence she must know certain premises about what she knows. The leftmost virgin, of course, is not required to know that she cannot be the sacrificial virgin. Thus, one iteration of knowledge is required for every participant other than the last. This is the case no matter how many virgins – how many potential sacrifices – there are.

Surely it cannot plausibly be argued that this single iteration is impossible. Hence, focusing on iterations of knowledge, and versions of (KK) will not provide the key to resolving the many variations of the prediction paradox.

The epistemic approach

The solutions canvassed up to this point have all been found wanting. In this section, I offer my own analysis for critical inspection. It is in the same general tradition as Quine's proposal in that the

central issues are taken to be epistemic, and the student is (in a limited way) denied knowledge of the announcement.[17] But Quine offers no positive explanation of the student's ignorance, and, as a consequence, his account seems open to the charge of scepticism either with regard to the future or with regard to testimony. The present proposal rectifies these deficiencies, and shows that the prediction paradox has considerable philosophical significance in the realm of epistemology.

First, the set-up. "Surprise" is interpreted in terms of "justified belief": to say that the exam will be a surprise is to say that the student S will not be justified in believing, before the day of the exam, that the exam will be on that day. To give the argument its due, we do not want S to fail to have justified belief for *accidental* reasons; we want him to be something akin to an ideal believer. So the four factual premises set out earlier (in the first section) must be assumed. As well as these assumptions concerning the details of the situation, the argument also relies on certain epistemological principles, which may be stated as:

(A_5) If T is justified in believing $P_1, \ldots, P_n, P_1, \ldots, P_n$ jointly imply Q and T sees this, then T is justified in believing Q.

(A_6) If T is justified in believing $P_1, \ldots, P_n, P_1, \ldots, P_n$ strongly confirm Q, T sees this and has no other evidence relevant to Q, then T is justified in believing Q.

Finally, as we saw earlier, the student must be credited with a certain degree of epistemic self-awareness. One way to spell this out is as follows. If there are k other premises required for the argument, then (A_{k+1}) will specify that throughout the week, the student is justified in believing $(A_1), \ldots, (A_k)$; (A_{k+2}) will say that, throughout the week, the student is justified in believing (A_{k+1}); and so on.[18] For a period of n possible test days, after the initial k premises, a further $n - 1$ premises will be necessary.

Even though S has been credited with ideal reasoning skills and memory, it is surely possible for the teacher's announcement to be true. Where, then, is the flaw in the argument? The first stage of the argument is just:

(1) If the only exam of the week is held on Friday, then on Thursday evening the student will be justified in believing that an exam will be held on Friday.

No doubt this first step of the argument looks inescapable. We reason that on Thursday evening the student will be justified in believing that it is now Thursday evening, and that an exam has not been held on this or any previous day of the week; and he will also be justified in believing, based on the teacher's announcement, that an exam will be held on exactly one of the days Monday to Friday. Hence, the student will be justified in concluding, on Thursday evening, that an exam will be given on Friday.

But this reasoning depends on our ignoring part of the student's total available evidence. Grant that on Thursday night the student remembers that the teacher is generally reliable and said:

(A) There will be an exam on exactly one of the days Monday to Friday.

We make use of this fact about S's evidence to conclude that S will be justified in believing (A) on Thursday evening. But, in so doing, *we overlook another part of the student's evidence.* He is also supposed to remember that the teacher, who is generally reliable, asserted:

(B) If an exam is held on day D then you will not be justified in believing this before that day.

Will the student be justified in believing (A) on Thursday evening? I think not. For suppose he is so justified. Now surely he will be justified in believing (A) only if he is also justified in believing (B), for there is no epistemically relevant difference for him between the two statements. He has exactly the same evidence for each: the fact of the teacher's announcement. However, if S is justified in believing both (A) and (B), then, realizing that it is now Thursday night and an exam has not been held on this or any previous day, he is also justified in believing:

There will be an exam on Friday and I am not now justified in believing that there will be an exam on Friday.

But surely this is impossible. It can never be reasonable to believe a statement of the form "*P* and I am not now justified in believing *P*". For if a person *T* is justified in believing a statement, then he is not (epistemically) blameworthy for believing it. But if *T* is justified in believing that he is not justified in believing *P*, then he would be at fault in believing *P*. Hence, if *T* is justified in believing that he is not justified in believing *P*, then he is *not* justified in believing *P*.

The upshot is that the student is not justified in accepting both (A) and (B) on Thursday night. And since his evidence concerning (A) is no better than his evidence concerning (B), he is not entitled to accept just (A).

Let me emphasize. It is not being maintained that the student can *never* accept the teacher's testimony;[19] the claim is only that he cannot believe it *on Thursday evening if no exam has yet been given*. And bystanders who happen to overhear the announcement, but to whom it is not addressed, can believe it even under these circumstances. There is no reason why someone other than *S* may not justifiably believe "There will be an exam on Friday and S is not now justified in believing there will be an exam on Friday."

Thus we see that the seemingly airtight argument can be stopped at the very first step, and a surprise exam is therefore possible on any day of the week. Of the premises needed for the first step of the argument, (A_1)–(A_4) simply specify the student's relevant intellectual abilities and evidence; and (A_5) and (A_6) seem, on the face of it, to be highly plausible principles. This is just as it should be, for an argument that we all find seductive is not likely to be based on obviously false premises. Now, however, it can be seen that (A_6) must be rejected. Even though on Thursday night the student has evidence that strongly confirms (A), sees this, and has no other relevant evidence, he is not justified in believing (A). For he cannot be warranted in believing (A) without also being justified in believing a statement of the form "*P* and I am not now justified in believing *P*".

This analysis rests, of course, on the assumption that the student has equally good evidence for (A) and (B).[20] So one might think the paradox would break out again if we simply made a slight revision. Let us suppose the teacher makes the same announcement, but, as well, the student has strong independent evidence that an exam will be given on exactly one of Monday to Friday. (For example, an

automatic and irreversible process has been set in motion that guarantees that an exam will be held on exactly one of Monday to Friday.) In this revised situation, the objection to the first step of the argument is no longer open to us. The student cannot believe both (A) and (B) on Thursday evening, but now he has reason to retain his belief in (A) and reject (B). But even if we therefore grant that a Friday exam will not be a surprise, we can still stop the argument at a later stage. The second step of the argument reads:

(2) If the only exam of the week is held on Thursday, then on Wednesday evening the student will be justified in believing (1), and therefore also justified in believing that an exam will be held on Thursday.

Now suppose that we grant that *S* will be justified in believing (1) on Wednesday evening. Still, we cannot reach the desired conclusion. The difficulty is that in order to be justified in believing that an exam will be held on Thursday, *S* will have to be justified in believing both (A) and (B); for he can rule out a Friday exam only on the basis of (B). But the student cannot be justified in believing both (A) and (B), since this would result in his being justified in believing:

> An exam will be held on Thursday and I am not now justified in believing that an exam will be held on Thursday.

Thus, under the revised conditions, the surprise exam can be held on any day but the last.

In general, however we revise the situation, we must claim at *each* stage of the argument that the student is justified in using (A) to predict the date of the exam; and at *some* point we also assume that he is justified in using (B) to rule out certain days. But the joint use of (A) and (B) in this way is impossible, and the paradoxical argument must therefore fail.

The starkest form of the paradox is the one-day version, in which the announcement is just "There will be a surprise exam tomorrow". The proposed analysis has it that, even under these circumstances, an exam will be a surprise. For the student cannot be justified in believing:

> There will be an exam tomorrow and I am not now justified in believing there will be an exam tomorrow.

If he has no reason to prefer either conjunct of this conjunction, then he cannot believe either, and a surprise exam is thus possible. On the other hand, if we consider a revised version of this abbreviated form, the result is quite different. If the student has strong independent evidence that an exam will be given tomorrow, then he is justified in believing that an exam will be given tomorrow and a surprise is not possible.

The analysis proposed here also has the virtue that it is able to deal with Sorensen's ingenious variations. To see how it applies, we must construe each situation in terms of a statement analogous to (A), which describes the basic set-up, and one analogous to (B), which states that the outcome will be a surprise. For simplicity, assume that there is no relevant epistemic difference between the subject's evidence for the two propositions. (Where there is such a difference, the analysis will proceed as above.)

Consider the designated student variation. (A*) gives the information concerning the set-up: five students are lined up in alphabetical order, a gold star is placed on the back of one of them and the student thus designated will be examined. (B*) states that the designated student will not be justified in believing that he is the designated student until after the students break formation. The argument for the designated student variation breaks down at the very first step:

(1) If Eric is the designated student, then he will be justified in believing this before the students break formation.

Eric's belief will, presumably, be based on (A*). But (A*) and (B*) are epistemically indistinguishable for him. So he can believe (A*) only if he is also entitled to accept (B*), in which case, he will also be justified in believing:

> I am the designated student and I am not now justified in believing that I am the designated student.

This is surely impossible. The upshot is that Eric cannot believe the teacher's announcement *if* he is in fact the designated student.

Thus, the argument collapses. The reader may verify that the sacrificial virgin paradox is also stopped at the first step: the rightmost virgin cannot be eliminated.[21]

The analysis thus provides a comprehensive solution to the prediction paradox, and the paradox itself appears to have considerable philosophical punch. What we learn from its resolution is that (A_6), although entirely plausible on the face of it, is in fact a flawed epistemic principle. For on Thursday evening, given no prior exam, the student is not entitled to believe the teacher's announcement despite the fact that he has good evidence for it (the testimony of a reliable person). As we shall see in Chapter 5, some would claim that the lottery paradox teaches us essentially the same lesson: roughly put, good evidence is not sufficient for justified belief. If this view of the lottery paradox is correct, then it is a close cousin of the prediction paradox.

This analysis of the paradox, and the rejection of (A_6), rest on two substantive assumptions:

(I) It is impossible to be justified in believing a pair of statements of the form "*P*, I am not now justified in believing *P*".

(II) If it is impossible to be justified in believing each member of the set $P_1, \ldots, P_n$, and there is no proper subset of $P_1, \ldots, P_n$ of which this is true, and you have equally good reason to believe each of $P_1, \ldots, P_n$, then you are not justified in believing any one of these statements.

The reasoning underlying (I) was indicated earlier. Assumption (II), which rests on the non-arbitrary nature of justified belief, is also essential to the analysis of the paradox. Without it, we cannot ensure that *S* is not justified, on Thursday night, in expecting an exam the next day.

The analysis depends, then, on these two epistemological assumptions, and succeeds only if they are correct. If they are, then those of us who are no longer students may stop worrying about surprise exams.

fairly straightforward matter. The truth-value of a conjunction with first conjunct *tb* and second conjunct *fb*, for instance, is presumably *fb*.

Chapter 3: Believing in surprises: the prediction paradox

1. M. Scriven, "Paradoxical Announcements", *Mind* 60 (1951), 403.
2. Parts of this chapter are taken from my "The Prediction Paradox Resolved", *Philosophical Studies* 44 (1983), 225–33, with kind permission of Kluwer Academic Publishers; and from "The Prediction Paradox: Resolving Recalcitrant Variations", *Australasian Journal of Philosophy* 64 (1986), 181–9, with permission of Oxford University Press.
3. See R. Sorensen, *Blindspots* (Oxford: Clarendon Press, 1988), 253.
4. Two "solutions" that first require us to grant that the argument is acceptable are: J. M. Chapman & R. J. Butler, "On Quine's 'So-Called Paradox'", *Mind* 74 (1965), 424–5; and R. L. Kirkham, "Paradoxes and a Surprise Exam", *Philosophia* 21 (1991), 31–52.
5. This variation is due to R. Sorensen, "Recalcitrant Variations of the Prediction Paradox", *Australasian Journal of Philosophy* 60 (1982), 355–62.
6. For instance, the analysis of C. Wright & A. Sudbury, "The Paradox of the Unexpected Examination", *Australasian Journal of Philosophy* 55 (1977), 41–58, which focuses on the retention of beliefs over time, is thus shown to be incomplete.
7. See A. Lyon, "The Prediction Paradox", *Mind* 68 (1959), 510–17.
8. See A. J. Ayer, "On a Supposed Antinomy", *Mind* 82 (1973), 125–6.
9. W. V. Quine, "On A So-Called Paradox", *Mind* 62 (1953), 65–8.
10. R. Shaw, "The Paradox of the Unexpected Examination", *Mind* 67 (1958), 382–4.
11. The (KK) thesis is usually discussed in the context of a formalized version of the argument. I shall try to present the issues informally, both to minimize the logical demands on the reader and to ensure that we remain close to the original intuitive argument.
12. Notice that we are here using a time-specific version of (KK): If *S* knows at *t* that *P*, then *S* knows at *t* that *S* knows at *t* that *P*.
13. J. McLelland & C. Chihara, "The Surprise Examination Paradox", *Journal of Philosophical Logic* 4 (1975), 71–89.
14. G. Harman, *Thought* (Princeton, NJ: Princeton University Press, 1973), 148.
15. See T. Williamson, "Inexact Knowledge", *Mind* 101 (1992), 217–42. The change in cognitive standpoint may involve a change in subject, or a change in time.
16. This variation can be found in Sorensen, "Recalcitrant Variations of the Prediction Paradox".
17. Others whose work falls in this general category include: R. Binkley, "The Surprise Examination in Modal Logic", *Journal of Philosophy* 65 (1968), 127–35; Wright & Sudbury, "The Paradox of the Unexpected Examination"; and R. A. Sorensen, "Conditional Blindspots and the Knowledge Squeeze: A Solution to the Prediction Paradox", *Australasian Journal of Philosophy* 62 (1984), 126–35. The philosophical underpinnings, in each case, however, are quite different, as are the epistemological consequences.

18. Note that if one of the premises *Ai* proves to be false, then the claim that *S* is justified in believing *Ai* should also be regarded as false.
19. Some have misinterpreted my analysis in this way. See, for example, R. Weintraub, "Practical Solutions to the Surprise-Examination Paradox", *Ratio* 8 (1995), 161–9.
20. The analysis has been criticized on this point. C. Janaway, "Knowing about Surprises: A Supposed Antinomy Revisited", *Mind* 98 (1989), 391–409, takes this assumption to be crucial to my analysis, and rejects it on that basis. However, he overlooks the fact that the solution can also apply to cases of unequal evidence. Nothing hinges on whether there is equal evidence for (A) and (B) in the original version of the paradox.
21. Sorensen has one other variation, the paradox of the undiscoverable position, which also falls to this solution in essentially the same way. See my "The Prediction Paradox".
22. This argument has been advanced by C. Chihara, "Olin, Quine, and the Surprise Examination", *Philosophical Studies* 45 (1985), 191–9.
23. Most notably, by J. Cargile, "The Surprise Test Paradox", *Journal of Philosophy* 64 (1967), 550–63 and by E. Sober, "To Give a Surprise Exam, Use Game Theory", *Synthese* 115 (1998), 355–73.