

How to set a surprise exam

Ned Hall

1. Introduction

The story: At the end of class one Friday afternoon, the professor announces to her students that she will give them an exam during one of next week's classes. (Class meets every day during the week.) She adds that the exam will be a surprise, in that the students won't expect, on the morning of exam day, that the exam will be that day. One of her cleverer students pipes up, saying that she cannot possibly fulfill her intention to give such an exam. "For it cannot be held on Friday: if it were, we would expect it on Friday morning (having noted that no exam had yet been given). So Friday is ruled out; the exam must take place on one of Monday through Thursday. But then, for exactly the same reason, it cannot be held on Thursday, else we would know that fact ahead of time (having noted that no exam had yet been given, and having ruled out Friday). And so on: It's really just a simple use of mathematical induction to show that your statement is inconsistent." The professor beams at her bright young student, and says nothing.

Arriving in class next Tuesday, the students discover that they are to take an exam that day. None of them, of course, expect it. The exam consists of one question: "What was wrong with the clever student's reasoning?"

What follows is an answer to the exam question. The question is worth answering, not just because solving puzzles is fun, and not just because it has not yet (in my opinion) received an adequate answer, but also because the solution, in this case, carries several interesting lessons for epistemology. First, a number of remarks about the story and how I will interpret it.

2. Preliminaries

0. So that we can ignore distracting and irrelevant questions about what the clever student is in a position to believe about the *other* students' beliefs, etc., I will pretend that he is the only student, even though he speaks of "we".

1. The story yields different puzzles, depending on how we understand "surprise" (equally, "expect"). I will construe the claim that the student is surprised that p is true to mean that he is not, antecedently, justified in believing p . (Adding the stipulation that he believes a proposition iff he is justified in believing it, I will use "justifiably believe" and "justified in believing" interchangeably.) That's hardly obligatory: I could have taken surprise to consist in lack of knowledge, or lack of (warranted) degree of belief above a certain threshold. Perhaps there are other readings, as well; the latter reading, at any rate, will come in for more discussion later.

2. Notation: Let " E_i " express the proposition that an exam takes place on day i , and let " $J_i(p)$ " express the proposition that the student justifiably believes, on the morning of day i , that p . (Day 1 is Monday, etc.) Then, for the general case of an n -day week, the professor's announcement can be rendered as follows:

$$SE_n: (E_1 \ \& \ \neg J_1(E_1)) \vee (E_2 \ \& \ \neg J_2(E_2)) \vee \dots \vee (E_n \ \& \ \neg J_n(E_n)).$$

To avoid tortured English, I will sometimes use " E_i " as a *name* for the proposition that E_i , etc.; context will make my meaning clear.

3. Diagnosing the flaw in the student's reasoning requires identifying the premise or premises *which the student is not justified in believing*.¹ Finding false premises is not enough, as the student's argument seems mistaken in a way that he himself ought to recognize—whereas one can

¹Clearly, he needs auxiliary premises: for SE_n , the premise for *reductio*, is not a contradiction (after all, it's *true*, in the story).

hardly be criticized, in this sense, merely for reasoning from false premises. So the exam question carries the hasty presupposition that there *is* something wrong with the student's reasoning. Since the exam surprises him, his argument must be unsound; but for all that it may be *cogent*, in the sense that he justifiably believes each of the premises, and they validly yield the conclusion that the announcement is false. We are often, after all, in the unfortunate position of having good reason to believe falsehoods. Needless to say, it will emerge shortly that the argument cannot be cogent.

4. I will ignore interpretations of the story according to which the professor's announcement is (tacitly) self-referential.² On such interpretations, the professor is saying not merely that for some i , an exam will take place on day i , but the student won't justifiably believe this beforehand; rather, she is saying that an exam will take place on some day i , and the student won't justifiably believe this beforehand, *on the basis of her very announcement*.³

How we proceed from here depends on how we understand the expression "on the basis of". Montague and Kaplan (who read the story in terms on knowledge, not justified belief) produce a rendering which makes the professor's announcement not only self-referential, but paradoxically so, in a Liar-like way.⁴ But the paradox has nothing to do with knowledge *per se*; replace "know" by any other modal operator with the right minimal set of formal features, and you'll get the same paradox.⁵ At any rate, self-referential interpretations of the story strike me as quite strained, given

²See for example Montague and Kaplan, "A Paradox Regained", *Notre Dame Journal of Formal Logic* 3 (1960).

³Together, it must be added, with the fact that no exam has yet taken place.

⁴It's easy to get a paradoxical sentence, in the 1-day case. Here's one:

(*) $E_1 \ \& \ \neg K_1((*) \text{ is true})$. ("K₁p" expresses the proposition that the student knows, on the morning of day 1, that p.)

We can now quickly "prove" that $\neg K_1(E_1)$. For suppose that (*) is true. Then the second conjunct is true. On the other hand, if (*) is false, then obviously $\neg K_1((*) \text{ is true})$. So we have proved that $\neg K_1((*) \text{ is true})$. Next, we can assume that what is provable is known by the student; hence we have $K_1(\neg K_1((*) \text{ is true}))$. Finally, we assume that knowledge agglomerates, so that if $K_1(E_1)$, then $K_1(E_1 \ \& \ \neg K_1((*) \text{ is true}))$ —i.e., $K_1((*) \text{ is true})$, a contradiction. Therefore, it must be that $\neg K_1(E_1)$. That's paradoxical, all right: the student may be ignorant, but it is not *a priori* that he is ignorant.

⁵Notice that in the last footnote's "proof" that $\neg K_1(E_1)$, the only features of knowledge used were that knowledge requires truth, that what is provable is known, and that knowledge agglomerates.

that the professor's announcement is perfectly—and therefore, it would seem, unparadoxically—*true*:

5. I will, at times, consider versions of the story according to which the “week” is one day long, or two days long, or even 100 days long. Why? Surely the flaw in the student's reasoning must show up at the very first step. For if he really succeeds in ruling out Friday, then it seems that his grounds for doing so will *therefore* allow him to rule out Thursday, and hence Wednesday, etc. If so, it doesn't matter how many days the week contains...so why bother with variations? (Better: why not focus exclusively on the 1-day case, so as not to be distracted by the irrelevant details of the multi-day cases?) For now, suffice it to say that these assumptions are unfounded. The analysis which follows will reveal systematic differences between all such variations of the story, and the 1-day case in particular will prove radically unlike the multi-day cases.

6. The story lacks detail, and many expansions of it render the task of diagnosing the student's reasoning trivial. For example, suppose he is inattentive and forgetful—so forgetful that he has no good reason to believe that he'll even *remember* the professor's announcement, by Friday. Then the argument falters immediately, since the student cannot conclude that if Friday rolls around without an exam yet having taken place, then he will be in a position to infer that he is about to take the exam. Again, if the student is woefully bad at reasoning—so bad that a simple *modus ponens* is a struggle—then he might remember the announcement, but fail to put two and two together.⁶

Best to forestall such misunderstandings by making explicit certain assumptions about the student's cognitive abilities:⁷

⁶Of course, he might nevertheless be *justified* in believing that an exam will take place: whether he is depends on the extent to which justification is hostage to cognitive ability. It is surely *somewhat* hostage. For example, were I very much smarter than I am I might be justified in believing Goldbach's conjecture (namely, if it's true and follows from the axioms of number theory); but I am not now justified in believing it.

⁷As well as other matters. As my wife, Barbara Popolow, observed, one way the professor could make good on her announcement is by devising an exam so subtle that the student was unaware that he was taking it—for which his cognitive abilities, presumably, would not be to blame. (The Memory condition, below, takes care of this clever wrinkle.)

Let S_i be the set of propositions p such that $J_i p$. Then we will assume that he is logically omniscient, so that the following principles hold (for all i):

Consistency: S_i is consistent.

Closure: Every consequence of S_i is an element of S_i .⁸

Since we do not want the argument to fail because the student is forgetful, or fails to notice whether he has taken an exam, we will also endorse the following principle:

Memory: for all i, k with $k > i$, $E_i \rightarrow J_k(E_i)$, and $\neg E_i \rightarrow J_k(\neg E_i)$.

Next, we will include a principle that signals that the concept of justification we are using is an internalist one, in the sense that one can, in principle, always introspect one's grounds for a belief, or one's lack thereof⁹:

Introspection: If $J_i p$ then $J_i(J_i p)$, and if $\neg J_i p$, then $J_i(\neg J_i p)$.

One more assumption is crucial. Recall that the student's argument breaks down immediately if he is not *justified in believing* that his memory and reasoning skills will remain in good working order; for then he cannot rule out Friday. The foregoing principles secure the *truth* of the needed belief, but not its status as *justified*. So we should add that the student justifiably believes every consequence of the four principles. But that's not enough, for consider the second step in the student's argument, where he purports to rule out Thursday. Suppose that the student's justified beliefs leave open the following possibility: Come Thursday morning, he will still be in good cognitive health—but he will *no longer* justifiably believe that this state will persist through Friday, and so will be unable to cogently argue that the exam cannot be on Friday. If so, he cannot *now* rule out Thursday, since by his own lights it is possible that, come Thursday, he will justifiably believe

⁸“Consequence” is defined in terms of “consistent” in the usual way: p is a consequence of S_i iff the set $S_i \cup \{\neg p\}$ is not consistent. The notion of consistency should be understood to encompass analytic truths, so that, for example, the set S_i counts as inconsistent if it includes both the proposition that Fred is a bachelor and the proposition that Fred is married.

⁹There are concepts of justification which lack this character. For example, you might hold that for a wide range of empirical propositions p , one is justified in believing p iff some suitably reliable process (a perceptual process, perhaps) generated this belief. The principle of Introspection need not follow; for one thing, J_p might not even belong to the given range of propositions. Of course, an externalist conception of justification *could* endorse Introspection; there's nothing about externalism *per se* that prevents this. At any rate, I won't haggle over which concept of justification is the “right” one; I only claim that it is natural and fruitful to read the story in terms of an internalist conception.

nothing stronger than that the exam will be either Thursday or Friday. In other words, his argument for ruling out Thursday presupposes that, come Thursday, the argument by which he now rules out *Friday* will still be available to him. (This shows, by the way, that the argument is no simple induction, and that it would be quite hasty to try to simplify the analysis by examining only the 1-day case.) So it seems that he must not only be justified (on day 1) in believing that he will remain in good cognitive health, but also justified in believing that, for any later day i , he will be justified on day i in believing that he will remain in good cognitive health.

It's apparent that the need for further iterations doesn't stop here (consider the step that purports to rule out Wednesday, etc.). Best to take care of them in one fell swoop, by stipulating that the student's confidence in the persistence of his cognitive health is so unshakable that, for purposes of evaluating *his* beliefs, the four foregoing principles can be treated as *analytic* to the concept of justified belief. I'll call this the Analyticity thesis. Once in place, it allows us to assume that any conclusion that we can draw by means of the four principles can also be drawn within the scope of the J-operators, no matter how deeply nested.

I will, finally, take it for granted as an Iron Law of the School that there can be *at most* one exam in any given week, and that the student is so certain of this Iron Law that it can also be treated as analytic, for him. This will streamline various derivations, by allowing me to freely treat the propositions E_i as pairwise incompatible—a move not licensed by their content.

Hereafter, I will creatively refer to the principles just articulated as “the Principles”.

End preliminaries, back to the exam question: What is wrong with the student's reasoning? The answer will come in five stages: First (§3), I lay some groundwork by considering what the student justifiably believes about the exam, on the assumption that he is *not* justified in believing the professor's announcement. This discussion leads directly to the second stage (§4), which examines

a simple proposal of Quine's, according to which the student's argument commits an elementary fallacy at the first step.¹⁰

Quine's diagnosis assumes that the student is not justified in believing the announcement, even in the multi-day case. The third stage (§§5-6) challenges this claim, and considers a quite different diagnosis due to Wright and Sudbury.¹¹

The fourth stage (§7) argues against the Wright-Sudbury proposal, pointing out that they have hastily dismissed a plausible additional constraint on the notion of justified belief, as it applies to the student. If this constraint—which I call Confidence—is in place, then it quickly follows that the student is *not* justified in believing the announcement, regardless of the number of days in the week. So it seems that Quine's diagnosis is vindicated.

In fact, we merely face a dilemma. For it is crazy to claim that when, say, the “week” is 100 days long, the student nevertheless cannot justifiably believe the announcement; surely he should be able to see that, with so many days at her disposal, the professor is *quite* capable of fulfilling her intention to give a surprise exam. Yet Confidence yields just the opposite result. The fifth stage (§§8-11) argues for a resolution of this dilemma which yields a novel understanding of the story, and of the student's reasoning. The considerations advanced in support of Confidence in fact support a slightly weaker principle, one whose statement requires us to replace the crude category of justified *belief* with the more fine-grained category of justified *degrees of belief*. Representing these degrees of belief as probabilities, and taking “justified belief” to mean “justified degree of belief above a certain threshold”, I show that we can uphold a weaker, probabilistic analog to the Confidence principle, *and* maintain that, provided the “week” is long enough, the student can justifiably believe the announcement. The resulting probabilistic analysis of the story leads to a new diagnosis of the logical flaw in the student's reasoning, and suggests, finally, that even those early

¹⁰See his “On a so-called paradox”, *Mind* 1953, pp. 65-7; reprinted as “On a Supposed Antinomy” in *The Ways of Paradox*, (Cambridge: Harvard University Press 1976), pp. 19-21 (page references are to this latter printing).

¹¹See their “The Paradox of the Unexpected Examination”, *Australasian Journal of Philosophy*, 1977, pp. 41-58; reprinted in Boyer, Grim, and Sanders, eds., *The Philosopher's Annual*, Vol. I, (Totowa, NJ: Rowman & Littlefield 1978), pp. 186-208. Page references are to this latter printing.

stages of it which are logically impeccable exhibit another kind of flaw: circularity. I close the fifth stage by arguing for this claim.

§12 sums up, and highlights several open questions of interest.

3. The student's evidence

Let us agree that aside from the professor's announcement, the student has *no independent evidence* which bears on the question of when or whether an exam will take place. (Let us also agree that he is justifiably certain that the only such evidence he *will* receive is the record, for each passing day, of whether the exam has taken place on that day; this assumption will only come into play much later.) It follows, I claim, that if the student is not justified in believing the announcement, then the announcement is irrelevant to him, and so he must remain agnostic about the exam. Let me explain.

Let T be the closure of the set of propositions $\{E_i\}$ under the truth-functional operations and the operators J_i . Among the elements of T will be propositions that the student is justified in believing, *prior* to the professor's announcement. These will include not just tautologies, but also propositions whose justification derives from the Principles: for example, the student is justified in believing the non-tautologous proposition that $E_2 \rightarrow J_3(J_4(\neg E_3))$.¹² Depending on what other principles are appropriate to the concept of justified belief, there might be other elements of T that he is justified in believing.¹³ But without trying to draw them sharply, we know there will be limits: for example, the student does not start out, prior to the announcement, justifiably believing that E_1 —or even that an exam will take place. That is what is meant by the claim that he is agnostic, and that is part of what follows from the claim that the student lacks any independent evidence about the exam. The other part is captured in the following principle:

¹²Proof: Memory and Analyticity yield both (i) $J_1(E_2 \rightarrow J_3(E_2))$ and (ii) $J_3(E_2 \rightarrow J_4(E_2))$. Closure and the Iron Law yield (iii) $J_4(E_2 \rightarrow J_4(\neg E_3))$. (ii) and Closure yield $J_3(E_2 \rightarrow J_3(J_4(E_2)))$, while (iii), Analyticity, and Closure yield $J_3(J_4(E_2) \rightarrow J_3(J_4(\neg E_3)))$; putting these together yields $J_3(E_2 \rightarrow J_3(J_4(\neg E_3)))$. With Analyticity, Closure and (i), this yields $J_1(E_2 \rightarrow J_3(J_4(\neg E_3)))$, the desired result.

¹³For example, if the Confidence principle to be introduced later were correct, he would also justifiably believe that $J_1(E_4) \rightarrow J_2(E_4)$.

The other part is captured in the following principle:

Irrelevance: If the student is not justified in believing the professor's announcement, then he is justified in believing exactly the same elements of T, before and after the announcement.

We should pause to consider the grounds for Irrelevance, since it is only in virtue of rather subtle features of the story that it holds. That might seem surprising; it might seem that Irrelevance is obvious. For surely the professor's announcement can affect the student's justification for believing elements of T *only* by giving him information about the existence and timing of an exam—and if he is not justified in believing it, how can he treat it as a source of such information?

Quite easily, if the circumstances of the announcement differ slightly. Consider the following variant: The professor announces—not to the students, but to a colleague—that she will set a surprise exam in the following week; unbeknownst to her, the student overhears her. What should the student believe? Clearly, that there will be an exam on one of the five days next week. But he cannot rule out that it is on Friday; after all, he knows full well that the professor (mistakenly) thinks she can fulfill her intention by setting the exam on Friday. Accordingly, he cannot rule out the possibility that he *won't* be surprised—and so is not justified in believing the professor's assertion. But she asserts the very same proposition as in the original story. It follows that (i) the student is not justified in believing her announcement; but (ii) he is justified in believing *additional* elements of T, upon hearing the announcement (e.g., that $E_1 \vee \dots \vee E_5$). Since Irrelevance is straightforwardly false in *this* scenario, more needs to be said about why it is true in the original story.

I won't pursue the matter in detail in this paper, but will rest content with a sketch of what I think is the correct account.

Observe a feature of ordinary conversational contexts, and the reliability of testimony therein. If you and I are conversing (in an ordinary context), and I do something that constitutes a clear violation of one or more of the norms governing such conversations, then—depending on the norms violated—my action may undermine whatever reason you have to believe that I am intending to speak truly. For short: If I openly break a conversational rule, then you cannot consider me

sincere, and so cannot consider me a credible source of testimony. Suppose now that the student is not justified in believing the announcement. Then *given* that the professor is addressing the *student*, it is to be understood that if this is so, the professor knows that it is so, and the student knows that she knows. But in that case, she has done something which renders her no longer a credible source of testimony, and so, with respect to propositions in T, her announcement leaves the student just where he started.

What has she done? Roughly, she has said something which she knows that her audience—the student—cannot, in the context, justifiably believe. *This* action—which, observe, distinguishes the story from the variant—constitutes a violation of conversational norms; the student’s recognition of this violation removes the presumption that the professor is credible.

More needs to be said about the relevant notions of credibility and sincerity, and the nature of the violated norms. Saying it properly is a surprisingly delicate operation, and so best postponed for another occasion. The important point for present purposes is that the thesis of Irrelevance allows us to neatly factor out these tricky questions from our main topic, the diagnosis of the student’s argument. The next section introduces a natural but ultimately unsatisfactory diagnosis due to Quine.

4. Quine’s diagnosis

First, some unfinished business: With Irrelevance in hand, we can quickly see that the student cannot have produced a cogent argument, since he is not justified in believing that the announcement is false. For suppose he is. Then he certainly is not justified in believing it to be *true*; so by Irrelevance, the elements of T that he is justified in believing must remain the same, before and after the announcement. But the proposition in question— $\neg SE_n$ —is not one of them:

- | | | |
|----|--|---------------|
| 1. | $J_1(\neg SE_n)$ | hypothesis |
| 2. | $J_1(E_1) \quad J_1(E_1)$ | 1, Closure |
| 3. | $\neg J_1(E_1) \quad J_1(\neg J_1(E_1))$ | Introspection |
| 4. | $J_1(E_1) \vee J_1(\neg E_1)$ | 2,3, Closure |

4 is unacceptable, since it contradicts the student's agnosticism about the exam. So the student must be pulling a fast one, relying on a premise to which he is not entitled.

To try to find it, let us reconstruct the first step of the argument as a piece of natural deduction, flagging those premises that are suspicious, and making free use of the Principles:

1. SE_5 hypothesis for *reductio*
2. E_5 hypothesis
3. $J_5(\neg E_1 \ \& \ \neg E_2 \ \& \ \neg E_3 \ \& \ \neg E_4)$ 2, Principles
4. $J_5(E_1 \vee E_2 \vee E_3 \vee E_4 \vee E_5)$???
5. $J_5(E_5)$ 3,4, Principles
6. $\neg J_5(E_5)$ 1,2, Principles
7. $\neg E_5$ 5,6

Step 4 is the obvious suspect. It makes explicit an assumption the student relies on when he says that the exam “cannot be held on Friday: if it were, we would expect it on Friday morning (having noted that no exam had yet been given).” Clearly the student needs to assume that, come Friday, he will be justified in believing that an exam is scheduled for the week—else he won't be able to infer, from the fact that no exam has yet been given, that it must be scheduled for Friday. But it is not at all obvious that the student is entitled to 4. It certainly doesn't *follow* from SE_5 , or from the conjunction of SE_5 with E_5 ; to think otherwise is to confuse the claim that a proposition is true with the claim that the student is justified in believing it to be true.

Enter Quine, who finds the crucial flaw in the argument at just this point.¹⁴ The student is *not* entitled to 4. Why not? Because he could only be so entitled if he were justified in believing the professor's announcement, and he is not.¹⁵

¹⁴Quine discusses the “knowledge” version of the puzzle, but the points he makes transfer seamlessly to the version under investigation here.

¹⁵At least, this *seems* to be Quine's ultimate verdict, although he's none too clear on this point. At any rate, his apparent willingness to treat the 1-day and many-day cases on a par (cf. his discussion on p. 21) suggests that he

Is it true that step 4 of the student's argument is legitimate only if he is justified in believing the announcement? Well, suppose it *is* legitimate. Then $J_1(SE_5 \rightarrow (E_5 \rightarrow J_5(E_1 \vee \dots \vee E_5)))$. This proposition, together with the Principles, readily entails that $J_1(SE_5 \rightarrow \neg E_5)$ (just embed the reasoning of steps 1-7 inside the scope of the J_1 operator), which in turn yields $J_1(E_5 \rightarrow J_5(E_5))$. But the student is *not* justified in believing, prior to the announcement, that $E_5 \rightarrow J_5(E_5)$ —for all he knows, the professor might give an exam on Friday with no forewarning. Given Irrelevance, he can come to be justified in believing that $E_5 \rightarrow J_5(E_5)$ *after* the announcement only if he is justified in believing the announcement. So this part of Quine's diagnosis is exactly right: if the student learns nothing relevant from the announcement, then he can't assume that if the exam is given on Friday then he will have good reason to believe so beforehand.

An advocate of this diagnosis can say a bit more to explain why the student's argument is so seductive: (i) If one has produced a cogent argument for p , then of course one is entitled to introduce the further claim that one is justified in believing p —provided, that is, that the argument was not carried out within the scope of an hypothesis. It might be all too easy to overlook this distinction between different contexts of argumentation.¹⁶ (ii) It might be all too easy to blur the distinction between what the student is justified in believing *about* the hypothetical situations he considers and what he is justified in believing *within* them. (iii) It might be all too easy to assume that the student *is* justified in believing the announcement.

Further discussion of (iii) is now in order. To begin, we have of course been assuming that until her announcement, the professor was considered by the student to be perfectly reliable. How might the professor undermine this trust? Well, she might say something that the student antecedently has good reason to believe to be false—good reason that is not undermined (or at any rate not sufficiently undermined) by her testimony to the contrary. (Such a statement could be, but need not

endorses the diagnosis I have attributed to him, since in the 1-day case it is patent that the student cannot justifiably believe the announcement (see below).

¹⁶Overlooking it corresponds to conflating the Introspection principle—which says, in part, that if Jp then JJp —with an absurdly strong variant, which asserts that if $J(p \rightarrow q)$ then $J(p \rightarrow Jq)$. To see how absurdly strong this is, replace 'q' by 'p': this yields, for all p , the claim that $J(p \rightarrow Jp)$. But then by Introspection and Closure we get, for all p , $Jp \vee J(\neg p)$. One's justified beliefs need not manifest such a high degree of opinionation.

be, an outright contradiction.) Or, she might say something the truth of which the student has good reason to believe she's not in a position to judge. But in making her announcement, the professor commits neither of *these* sins.¹⁷ Still, if we look to the 1-day case as a model, we should conclude that her statement renders her unreliable for a quite different and rather unusual reason: she says something that the student, and the student alone, cannot possibly justifiably believe (cf. the discussion in §3, above). For the 1-day case admits of a snappy *reductio*:

- | | | |
|----|---------------------------------|--------------------------------|
| 1. | $J_1(E_1 \ \& \ \neg J_1(E_1))$ | hypothesis for <i>reductio</i> |
| 2. | $J_1(E_1)$ | 1, Closure |
| 3. | $J_1(J_1(E_1))$ | 2, Introspection |
| 4. | $J_1(\neg J_1(E_1))$ | 1, Closure |

But 3 and 4 violate the Consistency requirement. (Of course, it is a curious feature of the announcement that its content does not prevent anyone *else* from justifiably believing it.)

No surprise, then, that the student's reasoning is so tempting. For the professor is clearly an authority on the matters about which she's speaking, so if she *hasn't* contradicted herself, shouldn't the student be justified in believing that come the last day, he will (still) justifiably believe her claim that an exam is scheduled for the week? But he won't justifiably believe it then, if he never justifiably believed it in the first place. And he won't justifiably believe it in the first place, if it is embedded in an announcement that he *can't* justifiably believe.

It's an attractive package: We have, on the one hand, a fairly sharp diagnosis of the flaw in the student's reasoning, and on the other, a set of handy excuses for having been taken in by it. Still, you shouldn't buy it.

¹⁷There might be some question about whether she's in a good position to judge the truth of claims about what the student justifiably believes; this can be settled, I think, by assuming that the stock of information that the student can draw on is in fact common knowledge between him and the professor, so that it is also common knowledge what her announcement justifies him in believing.

5. Quine's diagnosis rejected

We should pause, briefly, to dispense with a bad—though oft-cited—reason for rejecting Quine's diagnosis.¹⁸ Begin with the perfectly sound observation that the story can be told in such a way that the student *is* justified in believing that, come Friday, he will justifiably believe that an exam is scheduled for the week. Just add a second Iron Law of the School: that there must be at least one exam each week. (Take this to fall under the scope of the Analyticity thesis.) Then the first step of the student's argument goes through just fine. So Quine's diagnosis is, evidently, inapplicable.

Perhaps—but in letter only, not in spirit. Observe that with the second Iron Law in place, the last disjunct of the professor's announcement—that $E_5 \ \& \ \neg J_5(E_5)$ —is, from the student's perspective, a *contradiction*. So, from his perspective, the *content* of her announcement is given not by SE_5 but by SE_4 : $(E_1 \ \& \ \neg J_1(E_1)) \vee \dots \vee (E_4 \ \& \ \neg J_4(E_4))$. And now Quine's diagnosis applies straightforwardly: He should simply insist that the student is not justified in believing the announcement, and so, come Thursday morning, not justified in believing that crucial part of it which asserts that if the exam is on Friday then it will come as a surprise—which, from the student's perspective, is tantamount to asserting that the exam *won't* be on Friday, which in turn is tantamount to asserting that the exam is scheduled for one of Monday through Thursday. That is, Quine should insist that the crucial premise that $J_4(E_1 \vee E_2 \vee E_3 \vee E_4)$ is *false*—which is *exactly* the diagnosis he gives to an ordinary 4-day surprise exam scenario. Oddly, it seems to have gone entirely unnoticed by those who press this variant of the story against Quine that its only real effect is to convert an n-day scenario into an n-1-day scenario.

However, legitimate doubts about Quine's diagnosis emerge as soon as we scrutinize the claim that the student is not justified in believing the announcement.¹⁹ No doubt this claim gains

¹⁸ See for example Ayer, "On a Supposed Antinomy", *Mind*, 1973, pp. 125-6; and Janoway, "Knowing About Surprises: A Supposed Antinomy Revisited" *Mind* 1989, pp. 391-410.

¹⁹ Don't say: "Well, if he *is* justified in believing it, then he has no business trying to produce a *reductio* of it!" Of course he doesn't—*whether or not* he is justified in believing it. We established that quickly, clearly, and without need of Quine's diagnosis, at the beginning of the last section. What we are concerned with now is a *different* question: namely, what is the *flaw* in the student's argument?

plausibility from the tempting thought that the 1-day and multi-day cases are, in all relevant respects, really just the same.²⁰ But that's quite mistaken; striking differences separate these cases. I'll focus on a logical difference first.

Consider the two-day case, where the professor announces that $(E_1 \ \& \ \neg J_1(E_1)) \vee (E_2 \ \& \ \neg J_2(E_2))$. Can we construct an argument, analogous to the one presented at the end of the last section, that the student cannot justifiably believe this proposition? That is, can we show—making use only of the Principles—that the claim that $J_1[(E_1 \ \& \ \neg J_1(E_1)) \vee (E_2 \ \& \ \neg J_2(E_2))]$ is inconsistent?

We could, if we were licensed to infer from this claim that $J_1(J_2(E_1 \vee E_2))$. For then we could argue as follows:

- | | | |
|----|---|-----------------|
| 1. | $J_1[(E_1 \ \& \ \neg J_1(E_1)) \vee (E_2 \ \& \ \neg J_2(E_2))]$ | hypothesis |
| 2. | $J_1(J_2(E_1 \vee E_2))$ | 1, ??? |
| 3. | $J_1(E_2 \ \ \ J_2(\neg E_1))$ | Principles |
| 4. | $J_1(E_2 \ \ \ J_2(E_2))$ | 2,3, Principles |
| 5. | $J_1(E_2 \ \ \ \neg J_2(E_2))$ | 1, Principles |
| 6. | $J_1(\neg E_2)$ | 4,5, Principles |
| 7. | $J_1(E_1 \ \& \ \neg J_1(E_1))$ | 1,6, Principles |

As we have already seen, 7 is inconsistent.

But the Principles do not license the move from 1 to 2; they do not permit *any* inferences from the student's present justified beliefs to his present justified beliefs *about* his future justified beliefs. To be sure, there ought to be *some* principle governing how justification persists over time. Perhaps we can find one that will allow us to fill in the above *reductio*; at any rate, we will shortly investigate the matter more closely. My point here is only that *if* the student cannot justifiably

²⁰As Quine seems to think: see in particular p. 21 of "On a Supposed Antinomy".

believe the announcement in the multi-day case, then that is for reasons quite different from those that operate in the 1-day case.

A more intuitive difference between the 1- and many-day cases becomes apparent if we focus on, say, a 100-day case. Here, there is a strong temptation to think that *of course* the student can justifiably believe the professor's announcement: With so many days at her disposal, shouldn't it be obvious to him that she will make good on her stated intention to give a surprise exam? More generally, isn't it intuitively obvious that, if there are sufficiently many days, the student is justified in believing the announcement?

Yes, it is. Still, the intuition bears scrutiny, as it is easy to go astray when seeking arguments with which to back it up. For example, one might observe that *we*—considering the case “from the outside”—do not hesitate at all to conclude that the professor can make good on her announcement (that, after all, is one of the lessons of the story). Why, then, should the student hesitate? Answer: He shouldn't, but that is quite irrelevant to the point at issue. For notice that our observation holds just as much of the 1-day case: here, too, both we and the student should judge that the professor can make good on her announcement. So what? *That* fact is not at issue; what is at issue is whether the professor has said something the student can justifiably believe. If she has not, then while he must recognize that she *can* make good on her announcement, he—and we—should not conclude that she *will*—after all, she has violated a crucial conversation norm, and in so doing rendered herself no longer credible, etc.

Again, one might insist, with Wright and Sudbury, that an adequate treatment of the surprise exam “should make it possible for the pupils to be *informed* by the announcement”²¹. Their subsequent discussion makes it clear that they think that one has been “informed” only if one's justified beliefs have changed²²; if so, then it follows that in those multi-day cases in which the

²¹*op. cit.*, p. 187; italics in the original.

²²cf. p. 196: “Let us entertain as an interpretation of ‘ $D_t^x p$ ’ [the epistemic operator Wright and Sudbury introduce]: *x* has *good reason to believe p* at *t*, where this is taken to involve that *x*'s total state of information at *t* justifies his belief in *p* and assertion that *p*.... It is (at least) good reason to believe the announcement, in this weaker-than-knowledge sense, that we want the headmaster to be able to communicate to the pupils.”

professor's announcement is "informative", the student is justified in believing it (for if he is not, then by Irrelevance his (relevant) justified beliefs will remain unchanged). But of course one can be informed in other ways; for example, one's *degrees of belief* can change. Indeed, the professor's announcement is informative, in this sense, *even in the 1-day case*: for surely the student is justified in considering it *more* likely that there will be an exam, after hearing the announcement. (This point will come in for more discussion in §§9 and 10.)

Well, maybe the initial intuition is untrustworthy, and some subtle argument will show that even in the most multi-day cases, the student is not justified in believing the announcement. As we'll see, there are such arguments. But they should be resisted. For we all, routinely, believe propositions exactly analogous to the professor's announcement—and find, on reflection, that their curious logical structure erodes our justification for believing them not one bit.

Example: As I write this, it is mid-March, and rather cool in the Boston area. Yet I know—or at any rate, am quite justified in believing—that before the year is out we will have at least one day where temperatures exceed 80°. But the weather is fickle, none too easy to predict: so I *also* know—or at any rate, am quite justified in believing—that come midnight before the first 80°-day, I *won't* be justified in believing that temperatures will exceed 80° in the coming day. No doubt I will consider this to be possible, perhaps even likely—but I am quite sure that my evidence will *not* warrant a degree of confidence so high as to be called "belief". So I justifiably believe that there will be a first 80°-day this year, and that it will come as a surprise.

Example: My infant son possesses a sweet and cheerful disposition, and smiles at us quite readily. But as with all new babies, it took a while for that to happen—and we knew full well that it would, while also knowing that it would certainly take no longer than six months. We also knew that, come midnight before the day in which he first smiled at us, we would not be justified in believing that we would see him smile that day; the first ones come far too tentatively for that. So we knew—or at any rate, justifiably believed—that his first smile would come as a surprise. And we were right.

established the converse. In order to do so, we need to introduce some principle describing how justification persists over time. Wright and Sudbury suggest, in effect, the following principle:²³

Persistence: For all i, k with $k > i$, $J_i(p) \rightarrow J_k(p)$.

By itself, Persistence allows *us* to conclude that if the student justifiably believes the announcement—and the announcement is true—then the exam cannot be on the last day:

- | | |
|--|--------------------------------|
| 1. SE_5 | hypothesis |
| 2. $J_1(SE_5)$ | hypothesis |
| 3. E_5 | hypothesis for <i>reductio</i> |
| 4. $J_1(E_1 \vee E_2 \vee E_3 \vee E_4 \vee E_5)$ | 2, Principles |
| 5. $J_5(E_1 \vee E_2 \vee E_3 \vee E_4 \vee E_5)$ | 4, Persistence |
| 6. $J_5(\neg E_1 \ \& \ \neg E_2 \ \& \ \neg E_3 \ \& \ \neg E_4)$ | 3, Principles |
| 7. $J_5(E_5)$ | 5,6, Principles |
| 8. $\neg J_5(E_5)$ | 1,3 |

But the *student* can employ a version of this argument only if he is justified in making the step from 4 to 5—i.e., only if he is *justified in believing* Persistence. It is not enough, for his purposes, merely that Persistence be *true* (even though that guarantees that the exam cannot be on Friday). So, following Wright and Sudbury, let us take Persistence to fall under the scope of the Analyticity thesis, so that the student is, at the outset, justified in believing it.²⁴

Then we can quickly derive another principle, which will be the focus of much of our attention in what follows: Suppose that $J_1(p)$. By Introspection, $J_1(J_1(p))$. Since the student justifiably believes Persistence, it follows by Closure that $J_1(J_i(p))$, for any $i > 1$. In short, we have established

Confidence: For all $i > 1$, if $J_1(p)$, then $J_1(J_i(p))$.

²³*op. cit.*, pp. 189-90. The qualification is necessary only because they introduce the principles governing justification as rules of inference.

²⁴That's not quite what they say: they achieve the same effect by introducing their analog of Persistence as a rule of inference (see the last footnote).

With Confidence in place, and with the additional assumption that the student is justified in believing the announcement, it now looks as if at least the first stage of his argument will go through—so that even though he cannot, of course, *refute* the announcement, he can at least establish that the exam cannot take place on Friday. For consider this stage again:

1. SE_5 given as justifiably believed
2. E_5 hypothesis
3. $J_5(\neg E_1 \ \& \ \neg E_2 \ \& \ \neg E_3 \ \& \ \neg E_4)$ 2, Principles
4. $J_5(E_1 \vee E_2 \vee E_3 \vee E_4 \vee E_5)$ Confidence
5. $J_5(E_5)$ 3,4, Principles
6. $\neg J_5(E_5)$ 1,2
7. $\neg E_5$ 5,6

The introduction of 4 is now legitimate: Since the student is justified in believing the announcement, he is justified in believing that an exam will take place, hence justified in believing that he *will* be justified in believing this claim, come Friday.

But look again: In fact, what we now have are tools for constructing a new *reductio* of the claim that the student is justified in believing the announcement, *no matter how many days there are*:

1. $J_1(SE_n)$ hypothesis
2. $J_1(E_1 \vee \dots \vee E_n)$ 1, Principles
3. $J_1(J_n(E_1 \vee \dots \vee E_n))$ 2, Confidence
4. $J_1(E_n \ \ J_n(\neg E_1 \ \& \ \dots \ \& \ \neg E_{n-1}))$ Principles
5. $J_1(E_n \ \ J_n(E_n))$ 3,4, Principles
6. $J_1(E_n \ \ \neg J_n(E_n))$ 1, Principles
7. $J_1(\neg E_n)$ 5,6, Principles
8. $J_1(SE_{n-1})$ 1,7, Principles

It follows, by repeated applications of this argument, that if $J_1(SE_n)$, then $J_1(SE_1)$ —hence, that $\neg J_1(SE_n)$.

We should consider a bit more closely the role Confidence plays in this *reductio*. Here is one way to bring out that role: Let S_0 be the set of propositions p such that $J_1(p)$ follows from $J_1(SE_n)$ together with the Principles. Given S_i , let S_{i+1} be the set of propositions p such that $J_1(p)$ follows from S_i together with at most *one* application of Confidence. Thus, S_0 will contain $(E_1 \vee \dots \vee E_n)$ but not, e.g., $J_n(E_1 \vee \dots \vee E_n)$; S_1 will contain this proposition but not, e.g., $J_{n-1}(J_n(E_1 \vee \dots \vee E_n))$; etc. For $i = n - 2$, S_i is consistent, and indeed contains all of $\neg E_{n-i+1}, \neg E_{n-i+2}, \dots, \neg E_n$. But S_{n-1} is not consistent, for it contains all of $\neg E_2, \neg E_{n-i+2}, \dots, \neg E_n$, and so also E_1 , and so also SE_1 . In short, were we to “hamstring” Confidence by allowing at most $(n-2)$ -fold iterated applications of it, no contradiction would follow from the assumption that the student justifiably believes the announcement. This observation will become important in the sections which follow, where I develop my preferred diagnosis of our puzzle.

But first we must return to Wright and Sudbury, for whom the lesson of the puzzle is different: The rule of inference that is their analog of Persistence (more accurately: of Persistence when taken under the scope of the Analyticity thesis) is *false*. After all, justification manifestly does not always persist over time: “Good reason to believe p may lapse as more information becomes available; or stronger reason to believe the contrary may emerge.” (p. 201) This verdict, they think, safeguards the intuition that the student can justifiably believe the announcement, while exposing the flaw in his argument as occurring at the very first stage, where he introduces the premise that $J_5(E_1 \vee \dots \vee E_5)$.²⁵ As they put it, “What the pupils ought to realize from the start is that *one* way for the headmaster to carry out the announcement will be to let a situation develop in which their reasons

²⁵Actually, their picture is slightly more complicated, since they also consider several variants in which the student has independent evidence about the professor’s intentions. For example, the student might possess evidence that an exam will take place so firm that he is justified in believing that it won’t be undermined by the passage of four examless days. In that case, they say, the flaw appears at the second stage, where the student assumes that come Thursday, he will still be justified in believing that the exam will come *as a surprise*. Of course, keeping in mind the remarks at the beginning of the last section, we can see that such variants don’t introduce anything fundamentally new.

for believing [that an exam will take place] and [that if it does, it will come as a surprise] are mutually discredited; or in which, if those reasons are of differing strengths, the weaker are discredited.” (p. 204)

Clearly, this diagnosis improves on Quine’s considerably. But it’s still not good enough.

7. The Wright-Sudbury diagnosis rejected

It is certainly true that justification does not always persist through time, even for those in the best of cognitive health—indeed, even for the student: For granting that he is justified in believing the announcement at the outset, suppose the professor postpones the exam until Friday. Then, come Friday morning, the student will no longer be justified in believing the announcement (since, given the other information at his disposal, to do so would be tantamount to believing it in the 1-day case).

But these points are not relevant. What matters is whether *Confidence* holds; if it does, then we get the unhappy conclusion that the student cannot justifiably believe the announcement. And in the context of the Principles, Confidence is logically weaker than Persistence, since Persistence (taken under the scope of Analyticity) states not only that the student is justified in believing that his justification will persist, but that this justified belief is *true*.

Wright and Sudbury raise the right question: “Might there not be some *other* way of representing [the students’] reasoning, involving no use of *d(iv)* [their analog of Persistence], by means of which the initial assumptions would be validly shown to be inconsistent even when the *D*-operators were interpreted in terms of reasonable belief?” (p. 207) And they recognize that the answer is “yes”—in fact, that a rule of inference proposed by Binkley, and exactly analogous to Confidence, will, together with their other inference rules, entail that the student cannot justifiably believe the announcement.²⁶ But, curiously, they seem *not* to recognize that, in the presence of their other inference rules (which they do not seriously question), this rule is *derivable* from

²⁶“The Surprise Examination in Modal Logic”, *Journal of Philosophy* 65 (1968).

$d(iv)$ —though not vice versa—and should therefore have been the focus of their attention all along.²⁷ What little they do say suggests that they think that Confidence can be easily dismissed:

Binkley argues that $d(vii)$ [their analog of Confidence] is valid for an ‘ideal knower’, ideal in respects which he details. But his argument for the principle takes account only of the special case, for which it certainly is valid, where x ’s experience at t will constitute a decision of the truth-value of p . The case relevant to present concerns is rather where r and t both occur too soon for us to be sure whether p is true or false, and where reasonable belief, not knowledge, is required. We are asking, then, whether if one has good reason to believe p at r , one *thereby* has good reason to believe that one will always be similarly placed at any subsequent t which is still too early for a verification or falsification of p . If we waive irrelevant complications to do with mortality *etc.*, there seems to be nothing to be said for such a principle. Certain sorts of reasons to believe may be inherently lasting, but the natural view is that the principle is counter-exemplified precisely by this kind of case: for, to repeat, the pupils ought to recognize that one way for the headmaster to do what he has announced that he will do is to leave matters so late that their reasons to believe the announcement lapse; *they have no reason to suppose* that this will not be his strategy. (pp. 207-8, last italics mine)

Wright and Sudbury err, as there is much to be said for Confidence. Even though I will ultimately recommend that Confidence be rejected, it is important to reject it for the right reasons—which Wright and Sudbury certainly do not provide. I will therefore spend some time speaking in its favor.

Begin by observing something odd about their claim that the students “have no reason to suppose” that the exam won’t take place on the last day. If this is really the case—if the students should recognize that one strategy that the professor might well employ is to wait until the last day to give the exam—then one wonders why, come the morning of the last day (and with no exam having yet taken place), it is suddenly rational for them to *overlook* this, and judge instead that they no longer have reason to believe that an exam will take place.

²⁷Here is the derivation, with their rules expressed in our notation: First, their rule $d(ii)$ states that if p entails q , and if r entails $J_i(p)$, then r entails $J_i(q)$; this is equivalent to our Closure condition. Second, their rule $d(v)$ states that if r entails $J_i(q)$, then r entails $J_i(J_i(q))$; this follows from our Introspection condition. Third, their rule $d(iv)$ states that if r entails $J_i(q)$, then r entails $J_k(q)$, where $i < k$; this is equivalent to our Persistence condition. Finally, the “other way of representing [the students’] reasoning” that they consider is labeled $d(vii)$, and states that if r entails $J_i(q)$, then r entails $J_i(J_k(q))$; this is equivalent to Confidence. (All of our conditions, here, are taken to fall under the scope of Analyticity.)

$J_i(q)$ entails $J_i(q)$, so by $d(iv)$ $J_i(q)$ entails $J_k(q)$. Then, by $d(ii)$, if r entails $J_i(J_i(q))$, then r entails $J_i(J_k(q))$. Suppose now that r entails $J_i(q)$. By $d(v)$, r entails $J_i(J_i(q))$, and so r entails $J_i(J_k(q))$. This shows that $d(vii)$ follows from $d(ii)$, $d(iv)$, and $d(v)$. Notice, however, that $d(iv)$ does *not* follow from $d(vii)$, together with the other rules; for there is no way to show, even if the students are justified in believing that their justification will persist, that this justified belief is *true*.

The problem is brought into sharper relief by considering the 2-day case. Apparently, we should consider it unproblematic that the student (i) justifiably believes that there will be an exam on either day 1 or day 2; (ii) judges it equally likely that the exam will be on day 1 as on day 2 (for if he “has no reason to suppose” that it won’t be on day 2, then this is surely an appropriate way for him to set his odds); and (iii) recognizes that if the exam does not take place on day 1, then the rationally appropriate response will be to withdraw the belief that an exam will take place at all. But an attitude like this bespeaks so little respect for the epistemic authority of one’s future self that it’s doubtful that we are really talking about *justified* belief.

Consider a more extreme example: the student starts out justifiably believing that an exam will take place on day 1 or day 2—and *also* justifiably believing that come the next morning, he will no longer be justified in believing this. That’s possible, but only if the student has good reason to believe that something strange will happen to him before the morning of day 2 (say, that he will suffer amnesia, and so have forgotten that an exam was scheduled). If, on the other hand, he is justified in believing that his day-2 epistemic state will simply be the result of incorporating—in a way that he now deems rational—evidence that accrues between now and then, then in general he has no right to believe a proposition *p* *now* if he is justified in believing that his upcoming evidence will warrant *withdrawing* this belief.

An important lesson emerges. Your justified beliefs leave open various possibilities for the evolution of your epistemic state up until some future time *t*. In some of these, perhaps, something goes awry, such that the *t*-state has not evolved in the right way from your present state. (Maybe you die, or suffer some less serious cognitive defect.) But suppose none of them are like this. Then you are justified in believing that the opinions of your future *t*-self are like those of an *expert* whom you consider wholly trustworthy, since by your lights the beliefs of that future self will not only be justified, but will have an evidential basis *at least as good as your present one*. So, if the *t*-states you consider possible exhibit a consensus—if, say, they are all states in which you justifiably believe *p*—then you are justified in believing *p* *now*: It’s as if your expert told you that *p* was true,

while only withholding, perhaps, the exact details of the evidence upon which he bases this judgment.

The same applies if you justifiably believe that, at t , you will *not* justifiably believe p . If so, then you must now have good reason to believe that the evidence you will have accrued by t will warrant agnosticism (at least) about p ; to continue with the “expert” metaphor, you must now have good reason to believe that your expert possesses evidence which warrants such agnosticism. In that case, you cannot *now* justifiably believe p .

These lessons do not yet apply to the 2-day version of our story, as Wright and Sudbury would tell it. For in that scenario, the student does not justifiably *believe* that come day 2, he will no longer justifiably believe that an exam is scheduled; it’s just that he (justifiably) considers it not too unlikely that this will happen. But the qualification only makes the inconsistency less blatant. Consider a parallel case: I believe that it will rain tomorrow, but I also believe that there is a good chance—say, at or close to 50%—that a certain meteorologist will shortly tell me that this belief is not justified. That I nevertheless maintain it can only show that I do not take the meteorologist’s opinion too seriously. For if I considered him an expert, whose opinion was to be wholly trusted, then I should not hold with such confidence an opinion which I believe his testimony may well undermine. Similarly, if the student is justified in believing that he will suffer from no cognitive defects before the morning of day 2, then he should view his day-2 self as an expert, and not hold with such confidence an opinion which he believes the ‘testimony’, as it were, of this expert may well undermine.

These considerations do not establish that the student’s justified beliefs should obey Confidence, for that principle is rather more conservative: contraposed, it states that if the student merely considers it possible (i.e., does not justifiably believe to be false) that at some later point, he will not be justified in believing p , then he is not now (i.e., on the morning of day 1) justified in believing p . That’s a stronger claim than is necessary to rule out the Wright/Sudbury description of the 2-day case, since for that purpose we can make do with a variant on Confidence in which “possible” is replaced by “not too unlikely”. Still, our discussion shows that Wright and

Sudbury's pronouncement that there is "nothing to be said" in favor of their analog of Confidence is far too glib—and their insinuation that the students can have justified beliefs about the exam without this placing any constraints on their justified beliefs about their *future* justified beliefs is simply off the wall. If the term "justified" is apt, then there must be *some* connection between one's present judgments concerning the facts and one's assessment of how these judgments will rationally evolve.

There is no dearth of arguments that Confidence provides the needed connection, at least for the student; indeed, in developing these arguments and considering objections to them, we'll see that they support the view that a slightly qualified version of Confidence applies to *any* agent.

The first argument is implicit in the foregoing discussion: Suppose there is someone you consider an expert, at least with respect to the evaluation of a certain proposition, *p*. In particular, if this expert informs you that *p* is not credible, then that will give you rationally compelling reason to be agnostic with respect to *p*. Suppose you have some reason to believe that your expert will so inform you—enough that you can't justifiably believe that she won't. Then it seems that you thereby have some reason to be at least somewhat skeptical of *p*—skeptical enough that you cannot be said to be justified in believing it. But the student must treat his future self as such an expert, since he justifiably believes that the opinions of this future self will simply result from correctly incorporating *more* evidence than he now possesses.

A different argument for Confidence derives from a simple model of belief revision. Suppose that when you learn some evidence-proposition *E*—and *E* is compatible with what you were antecedently justified in believing—then your new justified beliefs simply result from conjoining your old ones with *E*; and suppose that you are well aware that this is the way you incorporate new evidence. (It won't matter what you do, when your evidence is incompatible with what you were antecedently justified in believing.) Suppose further that you are justified in believing that between now and time *t*, you will not be misled into treating as evidence anything that is false. It follows that if you are presently justified in believing *p*, then you are also justified in believing that, at *t*, you will be justified in believing *p*. For let *w* be a world compatible with what you justifiably believe. Let *E*

be the proposition which, according to w , constitutes your total evidence between now and t . By assumption, E is true in w ; hence, since w is compatible with what you justifiably believe, so is E . But you justifiably believe that, in any world in which your evidence is compatible with what you (antecedently) justifiably believe, that evidence gets incorporated via the simple updating rule. It follows that, in w , your justified beliefs at t are the result of conjoining E with your presently justified beliefs. Hence, in w at t , you justifiably believe p . Since this holds for any w compatible with what you justifiably believe, it follows that you justifiably believe *now* that, at t , you will justifiably believe p .

Finally, we can argue for Confidence by observing that if we are looking for a principle that can be expressed using no other epistemic notion besides that of justified belief, then there really isn't any other option. For the only other non-trivial candidates are these:

(1) For $i < k$, $J_i(p) \quad J_i(\neg J_k(\neg p))$

(2) For $i < k$, $J_i(p) \quad \neg J_i(\neg J_k(p))$

(3) For $i < k$, $J_i(p) \quad \neg J_i(J_k(\neg p))$

Confidence (together with the Principles) entails each of these, though not conversely; (1) and (2) are logically independent (even given the Principles); (1) and (2) (with the Principles) entail (3), though not conversely.²⁸

Intuitively, these principles say that if you justifiably believe p , then (1) you justifiably believe that you will continue to consider p to at least be possible; (2) you consider it at least possible that you will continue to justifiably believe p ; (3) you consider it at least possible that you will continue to consider p to at least be possible. While these are all substantive claims about how justified belief constrains the rational assessment of one's future epistemic state, each is too weak to serve as an

²⁸ To see that (1) follows from Confidence, observe that $J_k(p)$ entails $\neg J_k(\neg p)$; to see that (2) follows, observe that $J_i(q)$ entails $\neg J_i(\neg q)$, and substitute $J_k(p)$ for q . To see that (2) does not follow from (1), observe that the set $\{J_i(p), J_i(\neg J_k(\neg p)), J_i(\neg J_k(p))\}$ is consistent with (1) but not (2) (intuitively, this describes a situation in which, at i , the student justifiably believes p , but also justifiably believes that, at k , he will consider both p and $\neg p$ to be possible). To see that (1) does not follow from (2), observe that the set $\{J_i(p), \neg J_i(\neg J_k(p)), \neg J_i(\neg J_k(\neg p))\}$ is consistent with (2) but not (1) (intuitively, this describes a situation in which, at i , the student justifiably believes p , but also considers both $J_k(p)$ and $J_k(\neg p)$ to be possible). To see that (3) follows from (1), observe that $J_i(\neg q)$ entails $\neg J_i(q)$, and substitute $J_k(\neg p)$ for q . To see that (3) follows from (2), observe that $J_i(J_k(\neg p))$ entails $J_i(\neg J_k(p))$, and contrapose.

adequate substitute for Confidence. To see why, consider an alternative scenario to the story: The student knows that exams take place only on Fridays; he further knows that while the professor *sometimes* announces earlier in the week that there will be an exam, she does not invariably do so, preferring to let some of her exams come as a surprise. Finally, he knows that she likes to keep her students on their toes, and so *never* lets them know beforehand if there *won't* be an exam.

The professor announces on Monday that an exam will be held this week. Clearly, the student's epistemic state changes: beforehand, he was not justified in believing that an exam would take place on Friday, whereas afterwards he is. Assuming that Confidence does not hold, but that the three weaker principles do, must his justified beliefs about his *future* justified beliefs concerning the exam have changed?

No. Let the "future" in question be, say, Thursday. Then on Monday morning, before hearing the announcement, the relevant aspect of the student's epistemic state can be captured as follows:

- (i) $\neg J_1(E_5)$
- (ii) $\neg J_1(J_4(E_5))$
- (iii) $J_1(\neg J_4(\neg E_5))$
- (iv) $\neg J_1(\neg J_4(E_5))$
- (v) $\neg J_1(J_4(\neg E_5))$

(i) holds because the student has not yet heard the announcement; (ii) holds because he cannot be sure there will be an exam—or that even if there is, the professor will announce this beforehand; (iii) holds because he knows that he *won't* be informed if there is to be no exam; (iv) holds because he considers it possible that the professor will announce an exam; and (v) is overdetermined by these last two reasons.

After hearing the announcement, by contrast, we have

- (i') $J_1(E_5)$

But since (we are supposing) Confidence does not hold, it is consistent to add

- (ii') $\neg J_1(J_4(E_5))$,

while the three weaker principles merely give us

(iii') $J_1(\neg J_4(\neg E_5))$

(iv') $\neg J_1(\neg J_4(E_5))$

(v') $\neg J_1(J_4(\neg E_5))$

—i.e., what we had to begin with.

The upshot is that as far as the weaker substitutes are concerned, the announcement need make no difference to the student's rational assessment of his future justified beliefs. That is unacceptable, and consequently something stronger than these substitutes is required. If we confine ourselves to the language of justified belief, Confidence itself appears to be the only option.

I'll close this section by responding to a number of objections to Confidence.

To begin, you might think that I have failed to appreciate the force of Wright and Sudbury's objection. Aren't they just pointing out that the student should recognize that the professor is quite capable of giving him *misleading evidence* (namely, by postponing the exam until the last day, and thereby making it seem possible that there will be no exam at all)? But if he has some reason to think that he will receive misleading evidence, in light of which he will withdraw his belief in p , then even though he is justified in believing p , he need not be justified in believing that this state will persist.

The suggestion cannot be that he has some reason to believe that he will take as evidence a proposition which is false; for the only evidence-propositions in question are those that record, for each day, whether an exam takes place on that day. If the suggestion is that he will *improperly respond* to this evidence (should he receive it), then of course we should agree that he can justifiably believe p while justifiably doubting that he will continue to believe p . But Confidence is perfectly compatible with this claim, since it concerns what he will be *justified* in believing, not merely what he will *believe*—and opinions that derive from an improper response to evidence are not obviously justified. At any rate, such a suggestion cannot apply here, since by assumption he will never suffer from such cognitive shortcomings, and knows this ahead of time. But if, finally, the suggestion is that he recognizes that he might receive evidence to which the *proper* response will be

to withdraw his belief in p , then, to repeat, this recognition should lead him to view p with at least some skepticism *beforehand*; that, after all, was the lesson of the arguments just canvassed.

Still, isn't it intuitively clear that we can have every reason to believe that down the line, our opinions will be quite different from what they are now—but no less justified, for all that? But the force of the intuition diminishes considerably, once we clear away certain misconceptions. For example, I may recognize that I will come to be justified in believing propositions which I do not now justifiably believe (without, presumably, being able to single out which ones). But that kind of foreseeable change in my opinion is quite obviously compatible with Confidence. Again, I may reasonably suspect that I will suffer from some cognitive defect which, though it breaks the connection between present and future justification, does not at all prevent me from continuing to *have* justified beliefs. For instance, if I know that I will suffer amnesia tonight, then I know that my epistemic state *tomorrow* will be evidentially impoverished compared to my *current* epistemic state—and so there will be many propositions, perhaps even ones I can identify, which I justifiably believe now but know I will not justifiably believe then.

While this kind of example does nothing to show that Confidence fails to apply to the student, it shows that Confidence needs to be qualified, if it is understood as a constraint on the justified beliefs of *any* agent. But it's also fairly obvious how to amend it. The idea is that Confidence should hold for an agent at t_1 (and with respect to $t_2 > t_1$) provided she is justified in believing at t_1 that she will suffer no “cognitive mishaps” between t_1 and t_2 . But instead of imposing a *general* “no mishaps” condition (which would be so hopelessly optimistic as to render the amended principle vacuous), we should relativize the condition to individual propositions; and instead of trying to present an exhaustive list of possible “mishaps”, we should specify them in terms of what they have in common. Thus:

Confidence, revised (and generalized): If an agent is, at t_1 , justified in believing p , then for any $t_2 > t_1$, she is justified in believing at t_1 that any veridical evidence she receives before t_2 will, together with the evidence she now possesses, rationally warrant maintaining her belief in p .

Suppose that the agent is, like the student, justified in believing that she will receive only veridical evidence between t_1 and t_2 , and that her state of opinion at t_2 will be the result of rationally incorporating this evidence (she won't suffer from any lapses of reasoning, attention, or memory). Then, by the revised version of Confidence, she justifiably believes, at t_1 , that the evidence she will receive will rationally warrant maintaining her belief in p , and hence that she will justifiably believe p , at t_2 . Since this holds for any proposition p that she justifiably believes, her justified beliefs at t_1 obey Confidence, as originally presented.

Clearly, the qualifications in the revised version of Confidence are unnecessary in the case of the student. But they are necessary for the rest of us. And once in place, Confidence seems not very intuitively suspect at all. In fact, for reasons that will emerge in the next section, I very much doubt whether any clear-cut counterexample exists.

Still, there are two other objections to consider that deny the need to exhibit a counterexample. The first stresses how plausible it is to suppose that at least one of your justified beliefs will lapse in the face of future evidence. Surely it is sheer hubris to hold that the evidence I accrue over, say, the next 10 years will warrant maintaining *every one* of the justified beliefs I currently have? At the very least, it seems clear that I am not now justified in believing that I will be justified in believing every one of them—even on the assumption that I will suffer no cognitive defects, etc.

Indeed. But to see this argument as directed against *Confidence* is hasty, since it is nothing more than a variant of the so-called “paradox of the preface”. Pick a large enough set of propositions, each justifiably believed, and you can easily pump the intuition that it is reasonable to suspect that at least one of them is false—i.e., that their conjunction is *not* justifiably believed. The set of propositions of the form $J_i p$ —where i refers to a time, say, ten years from now, and p is a proposition I am currently justified in believing—is certainly large enough; we should agree, I suppose, that I am not justified in believing their conjunction, even if I am justifiably certain that I will suffer no cognitive mishaps between now and then. But since we must deal with many other versions of the paradox of the preface for which Confidence is obviously not a suspect, it would be foolish to pin the blame on that principle, in the present case.

A second objection tries to leverage the failure of Persistence into an argument against Confidence. Persistence is false, since there are plenty of actual situations (let alone possible ones) in which justification does not persist—indeed, in which an agent justifiably believes a proposition at one time, even though she subsequently receives veridical evidence which warrants withdrawing this belief. Suppose an agent knows this. Then perhaps, in light of her knowledge, she should harbor at least *some* doubt about the longevity of any of her justified beliefs—enough that Confidence cannot be true of her.

Then again, perhaps not. Never mind that agents will almost always possess plenty of extra information relevant to assessing whether their justification will persist, information that will render the bare fact that Persistence sometimes fails irrelevant. (In general, it doesn't follow from the fact that I know that some A's are not B's that I am not justified in believing that *this* A is a B.) Suppose that somehow, the only information our agent possesses that bears on whether she will receive veridical evidence that will undermine her justified belief in p is that this sort of thing *sometimes* happens. Even so, shouldn't the *frequency* with which it happens make a difference? If she justifiably believes that this frequency is extremely low, then why doubt that she is justified in believing that her justified belief in p will persist? After all, almost every proposition that we are justified in believing has *some* low-but-non-zero probability of coming out false.

On the other hand, if the frequency is higher, it's not clear that she is justified in believing p, in the first place. Here we should distinguish two cases. First, the agent might be aware that it happens with non-negligible frequency that justified beliefs are *wrongly* undermined by subsequent evidence—that is, undermined (and rationally so) even though the propositions in question are true. Well, if this is the sort of epistemic environment she justifiably believes herself to inhabit, then she should be prepared to greet with *extra skepticism* any evidence she encounters against her justified belief in p. But in that case, she is not recognizing as a possibility (i.e., something not ruled out by her justified beliefs) that she will receive veridical evidence that will *undermine* her justification for believing p; she is only recognizing as a possibility that she will receive veridical evidence that, were she unaware of the sort of epistemic environment she inhabits, *would* undermine her justification

for believing *p*. On the other hand, if she is justified in believing that it happens with non-negligible frequency that justified beliefs are *rightly* undermined by subsequent evidence, then her standards for what counts as sufficient evidence to justify belief in *p* should adjust accordingly—in which case she is not, after all, justified in believing *p* (or if she is, then she has *very* good reason to believe *p*—good *enough*, she should judge, to withstand any countervailing evidence that might come along).

While not exhaustive, these two possibilities show that the failure of Persistence—even if it is widespread enough to be significant—cannot automatically be taken as evidence (indeed, even defeasible evidence) against the claim that a particular justified belief will remain so. For absent further argument, we have at least as much right to insist that the rational response to recognizing that Persistence fails is to endorse more stringent epistemic standards—either for what counts as evidence enough to justify belief in *p*, or for what counts as evidence enough to undermine such justification.

8. Clues towards a new diagnosis

Confidence is eminently defensible. All the same it is false, and clearly so. For if Confidence is true then I do not justifiably believe that I will be surprised by the first 80° day this year, you do not justifiably believe that you will be surprised by the first coin toss to land heads, the student does not justifiably believe the announcement, no matter how many days in the “week”, etc.—all of which is absurd. At the same time, I insist that there can be no clear-cut counterexamples to Confidence. What is going on?

An important clue emerges from the revised Confidence principle, restated. First some terminology: say that a proposition is, for an agent, evidence that *p* might be true (false) just in case, given that proposition as evidence, the agent is not justified in believing *p* to be false (true). So if I have evidence that *p* might be false, then I am not justified in believing *p*. Confidence says that evidence that one might receive evidence that *p* might be false *is itself* evidence that *p* might be false.

Perhaps so. But surely such “evidence once removed” is *weaker*—and the problem is that Confidence makes no provision for variation in evidential strength. More exactly, suppose I have as

evidence the proposition E' that I might receive evidence E, which itself is evidence that p might be false. Granting that E' is itself evidence that p might be false, it is surely not as *strong* evidence against p as is E. And surely the proposition E'' that I might receive as evidence the proposition E' is even weaker evidence against p. So too for E''', E'''', etc. Yet Confidence implies that all of the propositions in this series have something in common: they all count as evidence that p might be false—evidence strong enough, that is, to remove justification for believing p.

That's wrong. As one proceeds along the series E, E', E'', ..., the strength of the evidence one finds against p clearly approaches a limit (after all, these strengths decrease and are, as it were, bounded from below). But it's quite implausible that this limit is anything but zero, or complete irrelevance. At any rate, it is surely not above the threshold at which evidence against p becomes so weak that one is justified in believing p, even in the face of it.

Sound familiar? Remove a grain of sand from a heap, and what remains is still a heap. But something must be wrong with this Heap Principle, else we get the absurd conclusion that one grain of sand—for that matter, *no* grains—constitutes a heap. Confidence, it appears, shares exactly the same problem.

I do not know how to solve this problem; certainly, the large and disparate literature on vagueness suggests that doing so is no easy matter. But at the very least, a more cautious statement of Confidence is called for. The next section will introduce greater precision; until then, let us make do with an understanding of Confidence as saying that if one is justified in believing p, then one is also justified in having a high degree of conviction—although not *quite* as high—that one will not receive veridical evidence that warrants withdrawing belief in p. This understanding helps clarify why (if I am right) there cannot be a clear-cut counterexample to the *less* cautious statement of Confidence: just as the erosion in a heap that results when we remove a single grain is slight, so too the erosion in justified conviction that results when we shift focus from p to the claim that belief in p will continue to be justified is slight. In each case, the change is too slight to take us from a determinate instance of one category to a determinate instance of its opposite.

This assimilation of the surprise exam puzzle to the sorites paradox has further payoffs. Significantly, we can use it to arrive at a much better diagnosis of the student's reasoning than either Quine or Wright and Sudbury provide. Recall that in order to complete the first step of his argument, the student needed the premise that, come Friday morning, he will still be justified in believing that an exam is scheduled. Granted that he is justified in maintaining a high degree of belief in this claim about his Friday state, it may not be high enough to count as belief, *simpliciter*. And even if it is, so that he *is* justified in ruling out Friday, there must come some point in his argument at which an application of Confidence yields a proposition the grounds for which are too weak to qualify as justification for *belief*—for as we saw earlier, successive steps in the argument require successively iterated applications of Confidence. It would be a serious mistake to suppose that, somehow, *all* of the applications of Confidence needed in the argument might succeed—for that would be to suppose that the student justifiably believes the announcement, but at the same time is able to produce a cogent argument that it is false. No; if the student justifiably believes the announcement, then his *reductio* must fail at some stage. But it need not fail at the first stage.

Still, too many questions remain open at this point. For example, exactly how close *is* the analogy between Confidence and the Heap Principle? Not incredibly close, it would seem: for while it surely requires many applications of the Heap Principle to move from a clear case of a heap to a clear case of a non-heap, it requires at most a four-fold application of Confidence to move from a clear case of a proposition justifiably believed to a clear case of one that is not (this, on the plausible assumption that the student justifiably believes the announcement). Again, *is* the student justified in believing that the exam won't take place on Friday? On Thursday? The Confidence principle, even when carefully stated, provides only vague advice on how to answer such questions.

What's needed is a statement of Confidence careful enough to be true, while precise enough to be useful.

9. The diagnosis completed: Confidence probabilified

The key is to represent, explicitly, the fact that opinion comes in degrees. I will, to begin, do this in the usual way, taking the student's state of opinion at any time to be represented by a *probability*

distribution over the propositions he can entertain, where a probability of 1 represents utter certainty in the truth of the given proposition, a probability of 0 utter certainty in its falsehood, and intermediate values correspondingly more moderate degrees of conviction. His *conditional* degrees of belief—opinions he would express by such locutions as, “On the supposition that A is true, it seems so-and-so likely to me that B is true”—I will take to be represented by the corresponding conditional probabilities $\Pr(B | A)$, understood to equal the ratio $\Pr(A \& B)/\Pr(A)$, at least when both terms are defined, and $\Pr(A) > 0$. I will assume that the student’s opinions change over time by conditionalizing on the accrued evidence: where $t' > t$, and E is the evidence the student acquires between t and t', $\Pr_{t'}(A) = \Pr_t(A | E)$, for all propositions A. I will assume that the student is certain that his opinions will change in this way. Finally, I will assume that the student’s degrees of belief are all justified, and will take the category of justified *belief*, in his case, to correspond to degree of belief above a certain “belief threshold” > 0 .

(Having chosen to interpret “justified belief” in this way, we must pause to reevaluate the Principles, and their use in the earlier sections. Do they follow from suitable probabilistic analogs? Alas, no: since a set of propositions all of which receive probability greater than θ need not be consistent, both Consistency and Closure fail. But the other Principles can be preserved, and the failure of Consistency and Closure turns out not to matter: suitable replacements can be found which legitimate the use of the Principles in all of the arguments we have examined so far, save only the argument that from $J_1(SE_n)$ and Confidence, $J_1(SE_{n-1})$ follows. But even in this case, a slight modification makes the argument amenable to probabilistic treatment. The details are sequestered in Appendix A.)

It would be silly to pretend that this model of opinion isn’t massively idealized. But it would be almost as silly to fuss over this point, insisting that the model cannot be explanatory because, e.g., real agents rarely have perfectly sharp degrees of belief, or never become *certain* of their evidence, etc. (Compare: It would be foolish to argue that the ideal gas model does *nothing* to explain the behavior of real gases, because real gas particles exert forces on each other, never engage in perfectly elastic collisions, etc.) What is important is simply not to be taken in by features of the

model that are, as it were, mere artifacts of the idealization. In the present case, we will eventually relax the assumptions that the student's "subjective probabilities" are perfectly sharp, and that there is a perfectly sharp belief threshold; the other assumptions built into the model are harmless. First, however, we need to consider whether we can find an analog of Confidence within this probabilistic framework.

As we can, by again exploiting an analogy with expert opinion. Begin with a simple case: Suppose I take someone to be an expert, at least with respect to the evaluation of a certain proposition p . Suppose, further, that I am *certain*—my subjective probability is 1—that *her* subjective probability for p is x . Then it seems that my *own* subjective probability for p should be x .²⁹

What about the general case, where I am not sure of my expert's opinion, but entertain various hypotheses about it, assigning each a probability? Here, the natural quantity to focus on is my *conditional* degree of belief in p , on the supposition that one of these hypotheses—say, that the expert's subjective probability for p is x —is true; and the natural value to assign to this conditional degree of belief is, of course, x . (This choice receives some confirmation from a useful heuristic, which states that an agent's conditional probability $P(B | A)$ is equal to the probability she would assign to B , were she to become certain of A .³⁰) If, for example, I have a subjective probability of 0.4 that her subjective probability for p is 0.9, and a subjective probability of 0.6 that it is 0.5, then my own subjective probability for p is fixed at $(0.9)(0.4) + (0.5)(0.6) = 0.66$. Uncertainty about my

²⁹That's close, but not quite right. For surely I can treat this person as an expert, without having such *complete* confidence in her opinion; put another way, my subjective probability that she is *perfectly reliable* can be very high but still less than 1. If so, my subjective probability for p need not be exactly x . (It will be quite close. Specifically, if my subjective probability that she is perfectly reliable is $1 - \epsilon$, then my subjective probability for p must be in the interval $[x - \epsilon, x + \epsilon]$.) But for now let us simplify, and add the assumption that I am utterly certain of her reliability. Shortly, we will see that for our present purposes, at least, this assumption loses us no generality.

³⁰The heuristic can't be taken too seriously, though. For one thing, the counterfactual supposition should really be that the agent becomes certain of A , *and of nothing stronger*; good luck finding, for typical choices of A , realistic counterfactual scenarios in which *that* will hold. For another thing, there appear to be straightforward counterexamples. For example, let A be the proposition that there is life on Mars, and let B be the proposition that, come tomorrow morning, I will assign a high probability to A . Clearly, were I to become certain of A , I would also become certain, or nearly so, of B . But, even on the supposition that there *is* life on Mars, I still consider it highly unlikely that I will *think* so, come tomorrow morning. So my conditional degree of belief $P(B | A)$ is quite low. (Still, this may just be the first problem in disguise: for in becoming certain of A , I will also—barring cluelessness about my own cognitive states—become certain that I am certain of A .)

expert's opinion, then, does not prevent my judgments about her opinion from constraining my own.³¹

Consider, now, some future time t . Clearly, *one* sort of expert opinion I should recognize is my *own* opinion, as it is at time t —at least, on the assumption that that opinion has evolved in the right way from my current opinion, in response to veridical evidence I accrue between now and t . Putting this observation together with the foregoing discussion, we arrive at the following close cousin of Confidence:

Probabilistic Confidence: For any agent S , times t_1 and t_2 ($t_2 > t_1$), and proposition p , S 's conditional probability for p at t_1 , on the supposition that she will accrue, between t_1 and t_2 , veridical evidence which (given her t_1 -epistemic state) will rationally warrant assigning a probability of x to p , should be x .³²

To bring out the close relationship between our original statement of Confidence and this new, probabilistic version, let us apply the latter to the case of the student, where we can make certain simplifying assumptions. Specifically, we can take the student to be (justifiably) certain, on the morning of day 1, that any evidence he accrues during the week will be veridical, and that his future

³¹ Further complications arise when we consider that I might possess evidence that is relevant to the evaluation of p , but which I believe my expert *doesn't* possess. One reaction is that, in that case, she is no longer an expert for me; after all, I cannot think that she is epistemically better placed than I am, with respect to the evaluation of p . But I think that's hasty. For my expert may have that status not because I consider her better informed, but because I consider her better able to assess the impact of evidence on the likelihood of p . (Compare: Sherlock Holmes is an expert detective not because he possesses every clue to every crime, but because he is particularly adept at evaluating such clues.) Supposing then that I possess evidence E relevant to p , what is important is not my assessment of her *unconditional* degree of belief for p , but rather her *conditional* degree of belief for p , given E . Still further wrinkles then arise if, for example, the hypothesis that she is a reliable expert is itself relevant to p . (For related discussion, see my "Correcting the Guide to Objective Chance", *Mind* 103, Oct. 1994, pp. 505-517.)

Fortunately, we can safely ignore these complications in the present case. For the "experts" in question will be future (that is, post-day-1) selves of the student, and we can safely assume that the student is certain that he possesses no evidence that his future selves won't also possess.

³² This principle is, in essence, an appropriately qualified version of van Fraassen's "reflection" principle (see his "Belief and the Will", *Journal of Philosophy* 81 (1984), pp. 235-56, and also Gaifman's "A Theory of Higher Order Probabilities", in Skyrms and Harper eds., *Causation, Chance, and Credence* (Dordrecht: Kluwer Academic Publishers 1988), pp. 191-219), which states that if S is to be rational, then her conditional probability for p , on the supposition that her future probability for p will be x , must be x . This overly bold statement is subject to decisive counterexamples (see for example Christensen's "Clever Bookies and Coherent Beliefs", *Philosophical Review* 1991, pp. 229-47), van Fraassen's recent defenses notwithstanding (see his "Belief and the Problem of Ulysses and the Sirens", *Philosophical Studies* 77 (1995), pp. 7-37). Hence the qualifications that appear in the principle as I've stated it are indispensable.

states of opinion will be just those that are rationally warranted by such evidence (given his day-1 state of opinion). Then, letting the functions $\text{Pr}_i(-)$ represent his degrees of belief for each day i , it follows from Probabilistic Confidence that for each $i > 1$ and proposition p , $\text{Pr}_1(p \mid \text{Pr}_i(p) = x) = x$.³³

Suppose now that the student justifiably believes p —that is, $\text{Pr}_1(p) = > .$ Let $= 1 -$ and $= 1 -$; so $< .$ Then for any $i > 1$, $J_1(J_i(p))$ iff $\text{Pr}_1(\text{Pr}_i(p) >) > .$ Given Probabilistic Confidence, an easy calculation shows that $\text{Pr}_1(\text{Pr}_i(p) >) = 1 - / ,$ hence that $J_1(J_i(p))$ if $< ^2.$ (See Appendix B for the derivation.) By a second application (with $/$ in for $,$ and $k < i$), it follows that $\text{Pr}_1(\text{Pr}_k(\text{Pr}_i(p) >) >) = 1 - / ^2.$ And so on: In general, Confidence “iterates” n times—that is, $J_1(J_{i_1}(J_{i_2}(\dots J_{i_n}(p))\dots))$ holds, where $1 < i_1 < i_2 < \dots < i_n$ —if $< ^{n+1}.$ Confidence *may* iterate even if this condition is not met; but this is the weakest condition that *guarantees* the iteration. On the reasonable assumption that typical cases of belief correspond to probabilities high enough above the belief threshold for Confidence to correctly apply, it is therefore no surprise that that principle should seem so appealing. (Shortly, we’ll modify our model of belief in ways that will make its appeal even less surprising.)

What of the arguments for Confidence I advanced earlier? Since I consider that principle mistaken—and Probabilistic Confidence the appropriate, weaker replacement—I need to rebut them. As follows:

- The first argument drew lessons from an analogy with expert advice. I reply that the correct lesson is that Probabilistic Confidence is true, and not the stronger non-probabilistic Confidence principle.
- The second argument relied on a model of belief revision whose key feature was that when your evidence E is compatible with what you were antecedently justified in believing, your new

³³ Note that what is substituted for ‘ x ’ here must rigidly designate a number, else absurd conclusions too quickly follow (the same goes for Probabilistic Confidence). For example, substituting ‘ $\text{Pr}_i(p)$ ’ yields (because ‘ $\text{Pr}_i(p) = \text{Pr}_i(p)$ ’ is a tautology) $\text{Pr}_1(p) = \text{Pr}_i(p)$ —i.e., it is a constraint on rationality that if you are certain that you will receive only veridical evidence and that you will incorporate it in the right way, then your opinions cannot change. Which, of course, is silly.

justified beliefs are simply the result of conjoining your old ones with E. I reply that the model is inappropriate, once we recognize that belief comes in degrees. If, for example, evidence gets incorporated by conditionalizing, then counterexamples are easy to come by, as there are plenty of compatible propositions p and E such that $\Pr(p) > \theta$ and $\Pr(E) > \theta$, but $\Pr(p | E) < \theta$.³⁴

- The third argument noted that no principle weaker than Confidence will do, if we restrict ourselves to the notion of justified belief. I reply that the argument weighs much more heavily against the restriction than it does in favor of Confidence.

Return now to the student. With Probabilistic Confidence in hand, it follows that he cannot be *certain* of the professor's announcement. For if he is, then—since $\theta = 0$ —Confidence iterates indefinitely, which as we saw earlier it cannot. This observation helps explain why the professor's announcement sounds a bit odd: while she hasn't asserted something the student cannot justifiably believe (except in the 1-day case), she *has* asserted something his confidence in which can only be so high. In fact, for an n -day week we know that, consistent with the student justifiably believing the announcement, Confidence can iterate at most $n - 2$ times; it follows that we must have $\theta > 1 - 2^{-(n-2)}$. Is the maximum such level of confidence ($\theta = 1 - 2^{-n}$) attainable? Yes: if the student assigns $\Pr_1(E_1) = \theta$ (just below the belief threshold), and in general $\Pr_1(E_i) = \theta (1 - \theta)^{i-1}$, then $\Pr_1(SE_n) = 1 - \theta^n$. What this shows is that as the "week" gets longer, the maximum confidence the student can attach to the announcement (on day 1) rapidly approaches 1 (assuming, plausibly, that θ is small). And this in turn helps explain why the announcement sounds *less* odd, the longer the week.

As an illustration, suppose that the belief threshold $\theta = 0.99$. Then on one extreme we have the probabilities that maximize the student's degree of belief that the announcement is true. We can display these most perspicuously by listing, for each day, two numbers: the student's probability, on the morning of day 1, that the exam will take place that day (call this the "initial probability"); and his probability on the morning of the day itself, on the supposition that the exam has not yet taken place (call this the "surprise probability"):

³⁴It's worth noting, however, that a close analog of this argument can be used to argue for Probabilistic Confidence; see my "On deriving Reflection from orthodox Bayesianism" (ms.) for details.

Distribution 1

<u>day</u>	<u>initial probability</u>	<u>surprise probability</u>
Monday	0.99	0.99
Tuesday	0.0099	0.99
Wednesday	0.000099	0.99
Thursday	0.00000099	0.99
Friday	0.0000000099	0.99
no exam	0.0000000001	

This leaves a probability that there will be no exam quite low enough that the student qualifies as justifiably believing that a surprise exam will take place. Note that if these are his degrees of belief—and they are warranted—then he also justifiably believes that the exam will take place either Monday or Tuesday.

The table helps explain, informally, why this distribution maximizes the student's probability that what the professor says is true. Consider an arbitrary disjunct of her announcement, $(E_i \ \& \ \neg J_i(E_i))$. The student's day-1 probability for this disjunct is given by $\Pr_1(E_i \ \& \ \neg J_i(E_i)) = \Pr_1(\neg J_i(E_i) \mid E_i)\Pr_1(E_i)$. But $\Pr_1(\neg J_i(E_i) \mid E_i) = 1$ if the surprise probability for day i is at or below $\frac{1}{10^i}$, and $= 0$ otherwise.³⁵ Correspondingly, $\Pr_1(E_i \ \& \ \neg J_i(E_i)) = \Pr_1(E_i)$ if the surprise probability for day i is at or below $\frac{1}{10^i}$, and $= 0$ otherwise. But if, in the table, the initial probability for any of the days is increased, its surprise probability will be bumped over the belief threshold—and so that initial probability will not contribute to the student's degree of belief in the professor's announcement. These initial probabilities are, therefore, all as large as they can be, consistent with them all contributing. Put another way, given the probabilities in the table, it is possible for the exam

³⁵To see this, recall that the student is certain that, for each i , his probability distribution on the morning of day i will be the result of conditionalizing his day-1 distribution on the accrued evidence, and that the only relevant evidence is the record of whether the exam has yet taken place.

to take place on *any* day, consistent with the truth of the announcement; no distribution which increases any of the initial probabilities has this feature.

Contrast this distribution with one which would have been appropriate, had the professor merely announced that an exam was scheduled for the week, without claiming that it would come as a surprise:

Distribution 2

<u>day</u>	<u>initial probability</u>	<u>surprise probability</u>
Monday	0.2	0.2
Tuesday	0.2	0.25
Wednesday	0.2	0.33
Thursday	0.2	0.5
Friday	0.2	1.0
no exam	0.0	

Given distribution 2, the student is certain, on day 1, that there will be an exam—but considers it only 80% likely that there will be a *surprise* exam (for only the first four days contribute to his degree of belief in this claim). That is well below any reasonable choice of belief threshold; so if we grant that the student *is* justified in believing the announcement, this distribution cannot be appropriate.

This obvious point is, for all that, quite important. Consider, for example, that it uncovers a serious deficiency in Wright and Sudbury's discussion. Recall that they are at pains to insist that an adequate account of the surprise exam paradox must show how the professor's announcement can be *informative*. But their own account does nothing whatsoever to explain what is distinctively informative about her claim that the exam will *come as a surprise*. Surely, there should be a significant difference between the student's opinions after hearing merely that there will be an exam, and his opinions in the present case, after hearing that there will be a *surprise* exam; and

surely this difference should be reflected in the student's assignment of probabilities. But Wright and Sudbury's blithe pronouncement that the students have no reason to suppose that the exam won't be on the last day suggests that they would consider distribution 2 to be perfectly appropriate, even when it is a *surprise* exam that has been announced.

That is clearly an untenable position. Indeed, it might seem that comparison of distributions 1 and 2 yields the following lessons: By announcing that the exam will be a surprise, the professor gives the student rational warrant (i) to assign some small but non-zero probability to the hypothesis that there will be no exam at all; and (ii) to heavily skew his probabilities towards the beginning of the week.

In fact, though, this is hasty. As to (i), it is quite possible for the student to be *certain* that there will be an exam, while still assigning a very high probability to the professor's announcement (one well above the belief threshold): for example, modify distribution 1 by changing the probability of a Friday exam from 0.0000000099 to 0.00000001; this merely lowers the probability of the announcement from $1 - 10^{-10}$ to $1 - 10^{-8}$. And at any rate, *utter* certainty that there will be an exam is not called for, even when the professor doesn't claim that it will be a surprise. (After all, there is some small probability that the world will come to an end before Friday, etc.) As to (ii), it would be madness to think that the student's probabilities should be skewed as much as those of distribution 1. It's not just that that distribution depicts massive overconfidence that the exam will be on the first day; it's that no matter how many days pass without an exam, this overconfidence remains undiminished, in the sense that each morning, the student remains all but certain that the exam will take place that day. Surely the announcement that the exam will come as a surprise does not warrant skewing the probabilities *this* much. In fact, the student can assign a high probability to the announcement, while distributing probability *evenly* over the five available days, as distribution 3 shows:

Distribution 3

day initial probability surprise probability

Monday	$\frac{99}{496}$	0.1996
Tuesday	$\frac{99}{496}$	0.2494
Wednesday	$\frac{99}{496}$	0.3322
Thursday	$\frac{99}{496}$	0.4975
Friday	$\frac{99}{496}$	0.99
no exam	$\frac{1}{496}$	

Distribution 3 leaves a residual probability of just over 0.002 that no exam will take place, and so a probability of just under 0.998 that the announcement is true—still high enough that the student counts as believing it. There is, however, no way to increase this probability without skewing the distribution toward the beginning of the week. So a more cautious statement of the lesson is called for: Any probability for the announcement high enough to qualify as belief must result from a distribution which either skews the probabilities towards the beginning of the week, or assigns some (small) probability to the hypothesis that no exam will take place; the probability for the announcement will be higher, to the extent that both of these features are present.

It is an unwelcome artifact of the model that it assigns a perfectly sharp belief threshold; in reality (and even when we relativize to a context) things are surely fuzzier. Better to relax this aspect of the idealization, and say that there is a *range* of probabilities (no doubt varying with context), bounded by an *upper belief threshold* above which one finds determinate cases of belief, and a *lower belief threshold* below which one finds determinate cases of non-belief; in between one finds cases that are indeterminate as between belief and non-belief. (Of course, it's also an artifact of this amended model that the range of indeterminacy itself has perfectly sharp boundaries. But we won't gain any further explanatory power by trying to relax this assumption.) Observe that if the upper threshold θ^+ and the lower threshold θ^- are such that $(1 - \theta^+) < (1 - \theta^-)^2$, then, given Probabilistic Confidence, no single application of the original Confidence principle can take us from a determinate case of belief to a determinate case of non-belief—a vindication of my earlier claim that

that principle has no clear-cut counterexample, and hence a further explanation of why Confidence seems (like the Heap Principle) so appealing.

Furthermore, we no longer need to say—implausibly—that once the probability the student assigns to the announcement is fixed, it is perfectly determinate on which of the days he is justified in believing the exam will take place. For example, suppose that the upper belief threshold $\beta^+ = 0.99$, and the lower belief threshold $\beta^- = 0.9$. And suppose the student assigns his probabilities as follows:

Distribution 4

<u>day</u>	<u>initial probability</u>	<u>surprise probability</u>
Monday	0.31	0.31
Tuesday	0.31	0.449
Wednesday	0.31	0.816
Thursday	0.062	0.886
Friday	0.007	0.875
no exam	0.001	

According to distribution 4, it is determinately true that the student believes a surprise exam will take place during the week, since the probability that it will is 0.999. It is also determinately true that he believes it will take place before Friday, since the probability that it will is 0.992. And, finally, it is determinately true that he *doesn't* believe it will take place before Wednesday, for the probability that it will is only 0.62. But it is indeterminate whether he believes it will take place before Thursday, as the probability of 0.93 is in the 'penumbral' region.

Consider, now, versions of the story in which the week is shorter. Understanding " $\neg J_i(p)$ " to mean that it is *determinate* that the student does not justifiably believe that p on the morning of day i , it quite obviously follows, in the 1-day case, that it is determinate that the student does not justifiably believe the announcement, since the probability he can give it must be less than β^- . The 2-

day case is a bit more interesting, at least given the assumption that the range of indeterminacy is such that $(1 - p) > (1 - p)^2$. For in this case the student must give the announcement a probability less than $1 - (1 - p)^2$; while this can easily be high enough not to count as a determinate case of non-belief, it will be too low to count as a determinate case of belief. In other words, it is at best indeterminate whether the student is justified in believing the announcement. Intuitively, this seems right: when asked to judge the matter, we feel uncertain *what* to say about the 2-day case.

Not so the 3-day case, where I think it is clearer that the student can justifiably believe the announcement. Given our choice of upper and lower belief threshold, distribution 5 vindicates this judgment:

Distribution 5

<u>day</u>	<u>initial probability</u>	<u>surprise probability</u>
Monday	0.55	0.55
Tuesday	0.40	0.89
Wednesday	0.044	0.88
no exam	0.006	

The importance of these remarks will become clear when we focus, in the next section, on a question discussion of which must surely seem long overdue: Exactly how *should* the student distribute his probabilities, in response to the professor's announcement? We have seen various possible ways that he *can* distribute them, but no argument for why any of these distributions is rationally appropriate—and so no argument for why the student *is* justified in believing the announcement. For it surely won't do for him to say, "Well, I should distribute my probabilities in a way which maximizes the probability that the announcement is true, without skewing things too much towards the beginning of the week." It's not that this claim is false; in fact, I think that is (roughly) how he should distribute his probabilities. It's that it simply overlooks the need to provide *reasons* for this imperative. After all, it's not that the student is, e.g., justified in assigning

some small-but-non-zero probability to there being no exam merely because doing so is necessary, if he is to assign a high enough probability to the announcement without excessive “skewing”; for why is he obligated to assign his probabilities in this way? No, the professor’s behavior must give him some more direct reasons for this assignment, and for his other assignments. Let us consider what such reasons might be.

10. Justifying the probabilities

In fact, it is not difficult to argue that distribution 4 describes a rationally appropriate response the student can have to the announcement—at least, on the assumption that the professor has heretofore given the student no reason to consider her untruthful, deceptive, or devious. (The force of this assumption will become clear as we proceed.) I’ll begin with a quick sketch, and then trace through the argument more carefully.

The student needs to consider two hypotheses. The far more plausible one—given, as we are assuming, that he has always known the professor to be honest, straightforward, etc.—is that she is not only intending to speak truthfully in making the announcement, but also intends that the student shall never have reason seriously to doubt her sincerity, in the ensuing days. The student should therefore reason that if this is the case, then she will set the exam early enough that it will, on every morning, remain determinately true that he is justified in believing her announcement. I will argue below that this rules out both Friday and Thursday. But it would be a serious mistake to conclude that the student should, on Monday morning, consider it *certain* that the exam will take place before Thursday. No, even though this is extremely likely, he must reserve some probability for the alternative hypothesis that the professor is *not*, in this instance, being honest, straightforward, etc. This hypothesis admits of subcases: (i) she might be lying, and not planning to give an exam at all; (ii) she might be telling the truth, but also be planning to play a trick on the student by waiting until Friday, and so making it the case that, determinately, he can no longer justifiably believe her; (iii) she might be planning to be just a little bit deceptive, by waiting until Thursday, and so making it that case that the student can no longer determinately justifiably believe her. Of these, (i) redounds

most to her discredit, and (iii) least; accordingly, the student should consider (iii) likelier than (ii), and (ii) likelier than (i).

Now let's run through it again, more slowly.

To begin, the student should recognize that *if* there is an exam, then it will—with certainty—come as a surprise. For if not, then there must be some i such that $\Pr_1(J_i(E_i)) > 0$. How, by the student's lights, could $J_i(E_i)$ come out true? Not if the exam takes place *before* day i , since the student is certain that if this happens, he will remember it, and so $J_i(\neg E_i)$ will be true. But if it *doesn't*, then—since $J_i(E_i)$ entails $J_i(\neg SE_5)$ ³⁶—it must somehow be the case that even though an exam has not yet taken place, and even though the student justifiably believes that the professor's announcement was false, he is still justified in believing that an exam is scheduled. This could easily be the case if he receives extra information about the exam—say, by sneaking a look at the professor's notes, and discovering that she has set it for day i . But not otherwise. And since one of our assumptions is that the student is certain, at the outset, that he will receive no such extra information, it follows that he is certain, at the outset, that the exam will come as a surprise, on the supposition that it is indeed scheduled. In other words (or rather symbols), $\Pr_1(SE_5 \mid E_1 \vee \dots \vee E_5) = 1$.

It doesn't yet follow that the student is justified in believing that SE_5 . Is he? Yes. For in the first place, he is certain that it is entirely within the professor's power to make it the case that SE_5 comes out true; and in the second place, nothing she has done casts sufficient doubt on her credibility to undermine his presumption that she has made the announcement sincerely, and so will exercise this power.

Contrast the 1-day case: There, she makes an assertion which, the student knows, *she* knows he cannot possibly justifiably believe; in so doing she violates a central conversational norm, and so renders herself—at least with respect to her announcement—no longer credible. That is why, in the 1-day case, the student is rationally required to suspend belief about whether there will be an exam.

³⁶For by Introspection, $J_i(E_i)$ entails $J_i(J_i(E_i))$, hence by Closure $J_i(E_i \ \& \ J_i(E_i))$, hence $J_i(\neg SE_5)$.

This does not mean that he shouldn't consider an exam *likely*; for plausibly, even in the 1-day case, he should consider it much more likely that the professor is being truthful but mischievous than that she is lying. And, of course, if his background knowledge were sufficiently different—if, for example, he knew that she loved to play tricks on her students, whenever possible—he might be justified in believing that an exam would take place, and so justified in believing that she had spoken falsely. But if his background knowledge is that she has always been truthful, non-deceptive, etc., then that knowledge must leave him in a state of doubt about her intentions.

Not so in the 5-day case. Perhaps she has violated a conversational norm, even in this case: for even though she has not said something the student cannot possibly justifiably believe, she *has* said something to which he cannot possibly assign a probability greater than $1 - (1 - \epsilon)^5$. And this infraction (if such it be) might give him some reason to doubt her credibility. But not enough to prevent him from (determinately) justifiably believing that she will give an exam. So $\Pr_1(E_1 \vee \dots \vee E_5) > \epsilon$. Therefore, since $\Pr_1(SE_5) = \Pr_1(SE_5 \mid E_1 \vee \dots \vee E_5)\Pr_1(E_1 \vee \dots \vee E_5)$, $\Pr_1(SE_5) > \epsilon$. In other words, the student is justified in believing the announcement.

All the same, he must assign *some* non-zero probability to the hypothesis that the exam will not take place. For we have already seen that $\Pr_1(SE_5 \mid E_1 \vee \dots \vee E_5) = 1$. It follows at once that if $\Pr_1(E_1 \vee \dots \vee E_5) = 1$, then $\Pr_1(SE_5) = 1$, which is impossible. So even though the student justifiably believes the announcement, he must reserve some probability $< 1 - \epsilon$ for the hypothesis that the exam won't take place.

Consider next the probability that the student should attach to the hypothesis that the exam is on Friday. Now, if the professor waits until Friday to give the exam, then she will have acted in such a way that the student can no longer justifiably believe that she has spoken truly. Worse: Her announcement will come true precisely *because* she has waited so long that the student can no longer trust it, and so cannot take it as providing him with reason enough to believe that there will be an exam. That's utterly sneaky—in contrast with, say, a Wednesday exam, which would warrant no such accusation. Furthermore, the professor is perfectly aware that a Friday exam will have this effect; so unless she wishes to be quite mischievously deceptive, she will not wait that long. The

student knows all this—and has, moreover, no reason aside from the announcement to consider the professor deceptive in this way. So β should also be small—small enough, plausibly, that the student is (determinately) justified in believing not only that an exam will take place, but that it will take place before Friday.

What about the ratio of β to α ? To reckon this, the student needs to consider how probable he will take an exam to be, should Friday morning arrive with no exam having yet taken place. Plausibly, the student should, in such a situation, consider it much more likely that the professor is being mischievously deceptive than downright untruthful; the values for α and β in distribution 4 reflect this judgment.

The student also has *some* reason—though not nearly so decisive—to think that the professor will not wait until Thursday to give the exam. For if she does, then even though it will not (as with Friday) be determinately true that the student can no longer justifiably believe the announcement, it will *also* not be determinately true that he *can*. So, while letting such a situation develop is not as deceptive as waiting until Friday, it is still not what one would expect from a fully cooperative conversational partner; the student should therefore consider a Thursday exam much more likely than a Friday exam, but somewhat less likely than a Monday, Tuesday, or Wednesday exam. Again, distribution 4 captures these constraints nicely.

To complete the argument, we need merely note that since the student can, on Wednesday, continue to (determinately) justifiably believe the announcement, even if it has not yet taken place, he has no reasons of the sort that tell against Thursday and Friday to doubt that the exam will be on Wednesday, or on an earlier day. Since he clearly has no other reasons which should lead him to judge one of the first three days a likelier choice than any of the others, his probabilities should—as in distribution 4—be flat over these options.

Still, it would be ridiculous to pretend that the argument just given shows that the exact numbers that appear in distribution 4 are uniquely justified. Rather, there is a large family of such permissible distributions, each with roughly the same “shape”. As for the student, we could say that he can, permissibly, adopt any of these distributions as his opinion; or better, we could say that

his opinion is *vague*, and is represented by the family as a whole, in the sense that the probabilistic judgments he determinately endorses are exactly those made true by every member of the family.³⁷ These details are not important, so long as it is recognized (i) that it is determinate that the student justifiably believes the announcement, justifiably believes that the exam will take place before Friday, and does not justifiably believe that it will take place before Wednesday; and (ii) that it is indeterminate whether the student justifiably believes that the exam will take place before Thursday.

It's worth emphasizing, once again, that the student's justification for these opinions depends crucially on his background knowledge of the professor's character. Of course, this knowledge need not be as we are supposing. If, for example, he knows her to have a wicked sense of humor, he might reasonably judge a Friday exam to be *more* likely than an exam on any of the earlier days. Still, all that matters for our purposes is that it be *possible* that the student's background knowledge justify him in reasoning in the way just outlined, and so justify him in assigning his probabilities after the manner of distribution 4. It's not that I've chosen to focus on distribution 4 because of its great intrinsic interest; it's that the reasoning which supports it neatly illustrates two important points.

The first is that the professor's addition that the exam will come *as a surprise* makes a great deal of difference (to be exact, it makes the difference between distributions 2 and 4), but for a quite unusual reason. Ordinarily, the informational value of an assertion is a simple function of its content: e.g., the professor announces merely that there will be an exam, and the student, taking her to be wholly trustworthy, simply incorporates this news; having no reason to favor one day over another, he adopts distribution 2 accordingly. But the additional claim that the exam will come as a surprise changes things dramatically, both by raising some small doubt as to whether there will

³⁷Notice that if (justified) subjective probabilities are typically vague, then there is reason to think that there will be no clear-cut counterexamples to the non-probabilistic Confidence principle, even on the assumption that the belief threshold is perfectly sharp. For suppose that there is a sharp belief threshold β , and that an agent is, determinately, justified in believing p . Then her vague probability for p will span some interval (α_1, α_2) , with both $\alpha_1, \alpha_2 > \beta$. Then unless the upper bound α_2 is so low that $(1 - \alpha_2) < (1 - \beta)^2$, the agent's vague probability for $J_i(p)$ will span an interval which, at worst, will properly contain β --and so, at worst, it will be indeterminate whether the agent justifiably believes that $J_i(p)$.

indeed be an exam, and—more interestingly—by making it the case that a Friday exam (and to some extent a Thursday exam) will constitute a blatant violation of the trust that the student has heretofore placed in the professor. So the way in which the student extracts information from the announcement is unusually subtle, involving explicit consideration of hypotheses about the professor's intentions (i.e., whether she is lying, or being mischievous, or neither) that is ordinarily quite unnecessary. At any rate, it patently does *not* involve the straightforward incorporation of the “news” that he will be surprised.

The second and perhaps more surprising point is that the reasoning which leads to distribution 4 illustrates that even if the student is justified in ruling out Friday (or earlier days), his *argument* for this conclusion is circular. Explaining why merits a section of its own.

11. The circularity in the student's argument

First, a few words about the notion of circularity are in order.

As far as I can see, this notion presupposes two other, more basic epistemological notions. The first is a distinction between directions of justification. I don't know how to define this distinction, but it is so utterly commonplace that it is readily fixed by examples. Thus, I might justifiably believe both that the ambient air temperature is 80° and that my very accurate thermometer registers 80°; but there is a crucial difference between these beliefs, in that the latter serves as my *reason* for the former, and not vice versa. Justification is asymmetric: if it flows in one direction, it does not flow in the other.

Of course, the asymmetry in this particular example could have been reversed: I might have consulted a second thermometer to figure out the temperature, and from this determined what the first thermometer's reading must have been. That shows that what justifies what is a context-sensitive matter, and so too whether an argument is circular: In the first example, to argue against someone who doubts the accuracy of my thermometer by appeal to the fact that the ambient air temperature is 80° is blatantly circular, whereas in the second example it is not circular at all.

The second example also illustrates the second important notion: For while my belief about the one thermometer justifies my belief about the other, it does not do so “directly”, as it were, but

only *by way* of my belief about the ambient air temperature. Justification of one claim by another is often, in other words, *mediated*.

Equipped with these two notions, we can give a fairly straightforward definition of “circular argument”. Observe first that all circular arguments have the feature that one or more premises do not, in the context, serve to justify the conclusion. But this is not a defining feature, since it is shared in common by arguments whose premises and conclusion are wholly irrelevant to each other. Nor should we say that what distinguishes circular arguments is that the conclusion serves, in the context, to justify one or more of the premises; that’s often so, but misses the simplest case of “p, therefore p”. The intuitive idea, rather, is that one or more of the premises are such that any justification that is available for them will *already* justify the conclusion. The sense of “already” is obviously not temporal; it is rather that this justification of the conclusion fails to proceed *by way of* the premises. Bearing in mind that we do not wish to count an argument as circular simply because some premise is *brute*—in the sense that there is nothing, in the context, that could serve to justify it—we arrive at the following definition: An argument is circular just in case at least one of its premises (i) has some justification available to it (in the context); and (ii) is such that any available justification of it also justifies the conclusion—but not by way of the given premise.³⁸

It remains to show that the student’s argument is circular.

Begin with a much simpler case than our story. Suppose that the professor explains that she will choose the exam day as follows: Before the first class, she will (in secret) draw a card at random from a well-shuffled deck. If it is a spade, she will return it, shuffle, and draw again. She will continue in this way until she draws a non-spade, or until she has drawn five times, whichever comes first. If she draws five spades in a row, there will be no exam; otherwise, its day will be given by the number of draws it took her to get a non-spade.

The student is adept at calculating the relevant probabilities, and so adopts

³⁸There are annoying cases which this definition doesn’t quite fit. For example, it might be that no single premise has these features, but some collection of them does. Or it might be that the argument is “partly” circular, in the sense that some of the premises have the given features, while others straightforwardly justify the conclusion. I’m going to ignore these complications, since the definition is, as stated, good enough to go on with.

Distribution 6

<u>day</u>	<u>initial probability</u>	<u>surprise probability</u>
Monday	0.75	0.75
Tuesday	0.1875	0.75
Wednesday	0.0469	0.75
Thursday	0.0117	0.75
Friday	0.0029	0.75
no exam	0.00098	

Given our earlier choices of upper and lower belief thresholds, it is, as before, determinate that the student justifiably believes that SE_5 . It is *also* determinate that he justifiably believes that $J_5(E_1 \vee \dots \vee E_5)$, since the probability that he assigns to this proposition is greater than 0.996. And that means that the first stage of the student's argument is logically impeccable:

1. SE_5
2. E_5 hypothesis
3. $J_5(\neg E_1 \ \& \ \neg E_2 \ \& \ \neg E_3 \ \& \ \neg E_4)$ 2, Principles
4. $J_5(E_1 \vee E_2 \vee E_3 \vee E_4 \vee E_5)$
5. $J_5(E_5)$ 3,4, Principles
6. $\neg J_5(E_5)$ 1,2, Principles
7. $\neg E_5$ 5,6

Since the student is justified in believing that $J_5(E_1 \vee E_2 \vee E_3 \vee E_4 \vee E_5)$, he is certainly entitled to introduce it as a premise at step 4. Furthermore, he is also entitled to the conclusion—that since (as he justifiably believes) SE_5 is true, the exam will not take place on Friday—for his (justified) probability that it will not is greater than 0.997.

All the same, his argument for this conclusion substantially misrepresents his reasons for it, in the way that circular arguments typically do: it presents as a *reason* for its conclusion something which, in the context, receives its very justification *from* that conclusion.

For consider *why* the student is justified in believing that, come Friday, he will justifiably believe that an exam is scheduled: his reason for this is simply that he justifiably believes that, come Friday, *the exam will already have taken place* (and that he will remember this, etc.). After all, on the assumption that the exam *doesn't* take place before Friday, he is *certain* that he won't, then, justifiably believe that one is scheduled, since he will assign this hypothesis a probability of only 0.75. And he justifiably believes that the exam will take place before Friday simply because he justifiably believes that the professor will decide the exam day in the way she has described. It is *not* that he reasons, "Well, come Friday I will justifiably believe that an exam is scheduled; but since that requires that the exam not take place on Friday, I can conclude that the probability that it will is extremely small." That's quite backwards: The student *starts out* knowing what the probabilities are, and concludes that since a Friday exam is so unlikely, he will almost certainly retain through Friday his justified belief that an exam is scheduled. In short, the only justification available for the "premise" that $J_5(E_1 \vee E_2 \vee E_3 \vee E_4 \vee E_5)$ traces to the student's knowledge of how the exam day will be determined; while this does indeed justify the "conclusion" that $\neg E_5$, it manifestly does not do so by way of the "premise".

Exactly the same points hold of our story; the only difference is that the student's basis for distributing his probabilities is not so crystal clear. Instead of relying on his knowledge of an explicit probabilistic mechanism, he must rely on his knowledge of the professor's character, and judge the likelihood that she is lying, or being mischievous. Perhaps these hypotheses should, in light of his background knowledge, receive such a low probability that he counts as justifiably believing that the exam will take place before Friday. If so, he can conclude that come Friday, he will still justifiably believe that an exam is scheduled, *because it will already have happened*. Note well the direction of justification: His knowledge of the professor's character justifies his

distribution of probabilities, which in turn justifies his belief that $J_5(E_1 \vee E_2 \vee E_3 \vee E_4 \vee E_5)$. Once again, the argument is circular: the justification for the premise already justifies the conclusion.

12. Recapitulation and open questions

Even if it is correct in every particular, the discussion to this point leaves several questions unsettled, and raises others. I'll highlight those I consider most important:

- It remains unclear how the story is to be understood if “surprise” is defined as prior lack of *knowledge*, instead of prior lack of justified belief. If we could take knowledge to be justified true belief, the analysis presented here would transfer over pretty much intact. But we can't—at least, not in general, and I don't see any decisive reason why our story is an exception.

- It remains unclear whether, in making her announcement, the professor has violated any conversational norms—and if so, which ones. It's quite clear, I think, that she has committed an infraction in the 1-day case. But even here, one might dispute my claim as to its nature: namely, that she has said something she knows her audience cannot justifiably believe. (After all, as my colleague Michael Glanzberg has pointed out, situations can easily arise in which one is obliged to make assertions one knows can only be treated with incredulity.) So more work needs to be done on these “pragmatic” aspects of the surprise exam.

- It remains less clear than I have, perhaps, made it appear what the relation is between *belief* and *degrees of belief*. The model I presented is beguilingly simple: “Belief” is a vague term; determinate cases of belief are degrees of belief above a contextually determined upper threshold; determinate cases of non-belief are degrees of belief below a contextually determined lower threshold; degrees of belief in between these thresholds are opinions whose status as belief is indeterminate. While I am convinced that this model provides the right framework for developing an adequate diagnosis of the surprise exam, I am also fully aware that it stands in need of more justification and elaboration than I have been able to give it.

Still, my analysis (if correct) goes a long way towards dissipating the air of paradox that has surrounded the surprise exam—and in the process, reveals several interesting facets to the concept of justified belief. Let's review:

Superficially, the surprise exam appears to present us with a genuine paradox: for the student has produced an apparently sound argument for a transparently false conclusion. But that appearance is easily dispelled, once we see that the first steps of the *reductio* require that the student justifiably believe the very statement that is its target. Still, a more substantial problem lurks beneath the surface, since it appears that *we* can construct a sound argument that the student cannot justifiably believe the announcement, no matter the number of days—and this is surely false. But the argument's crucial premise—Confidence—is, even though extremely plausible, false. What explains its plausibility? Answer: the truth of its slightly weaker cousin, Probabilistic Confidence. This principle entails not only that typical instances of Confidence will be determinately true, but also, plausibly, that *no* instance of Confidence will be determinately false (at least when that principle is applied to a proposition determinately justifiably believed).

Endorsing Probabilistic Confidence allows us to safeguard the student's justified belief in the exam in a natural and well-motivated fashion. Perhaps more importantly, the refinement that comes from describing the student's opinions within a probabilistic framework enables us to draw a clear picture of the way in which announcing a *surprise* exam gives the student extra information. For by adding the "surprise" clause to her announcement, the professor has rendered salient hypotheses that ordinarily wouldn't matter to the student: namely, that she is lying—or, more interestingly, telling the truth, but also mischievously planning to act in such a way that the student will, eventually, no longer be able to believe her. It's not that the student should consider these hypotheses at all likely; given his background knowledge of the professor's character, he shouldn't. But *because* he shouldn't, he should consider it very likely that the exam will take place before Friday, and should even consider it fairly likely that it will take place before Thursday.

Notice that the probabilistic character of these judgments blocks our *reductio* of the claim that the student justifiably believes the announcement. For while he is, indeed, determinately justified in

believing that the exam will take place before Friday, hence justified in believing that a surprise exam will take place in the first *four* days, his warranted *degree* of belief in this claim must be lower than his degree of belief in the claim that a surprise exam will take place in the first five days. Moreover, the most that can be expected at the next step is that the student is *not determinately not* justified in believing that a surprise exam will take place in the first three days. And he is—determinately—not justified in believing that a surprise exam will take place in the first two days.

Finally, we can see that even though at least the first step of the student's argument manifests no logical flaw—in that it is valid, and he is justified in believing its premises—its force as a *reason* for ruling out Friday is vitiated by its circularity. For in order to argue from the premise that he will, come Friday, justifiably believe that an exam is scheduled, he must *already* be justified in believing that the exam will take place before then.

It is not that one can never use the fact that one will have some justified belief as a basis for drawing further conclusions. In fact, one reason the circularity in the student's argument is so easily missed is that the first steps closely mimic an argument that—had the announcement been slightly different—would have been perfectly acceptable. For suppose the professor merely announces that an exam will take place during the week. Then, as in our story, the student is justified in believing that, come Friday, he will justifiably believe that an exam is scheduled. But—unlike our story—he is also justified in *concluding* from this, without a trace of circularity, that if the exam takes place on Friday then it will not be a surprise. To be sure, he is justified in believing this conditional in our story, as well: but only because he is justified in believing its antecedent to be false—and this for reasons that his stated argument entirely misrepresents.

Looked at another way, the student's argument involves some deft doublethink: He first pretends that the professor has merely announced an exam, and from this premise correctly concludes that if it takes place on Friday, it won't be a surprise. He then lifts the pretense, and appeals to the (incompatible) premise that she has announced a *surprise* exam; maintaining the

illicitly drawn conclusion that if the exam takes place on Friday, it won't be a surprise, he concludes that if she has spoken truly then the exam cannot take place on Friday.

As the extensive literature on the surprise exam attests, a cleverer piece of misdirection would be hard to find.

Appendix A: Probabilistic versions of the Principles

Begin with those of the Principles that have ready probabilistic analogs:

For Memory we can substitute the following:

P-Memory: for all i, k with $k > i$, $E_i \quad \Pr_k(E_i) = 1$, and $\neg E_i \quad \Pr_k(\neg E_i) = 1$.

For Introspection we can substitute the following:

P-Introspection: if $\Pr_i(p) = x$, then $\Pr_i(\Pr_i(p) = x) = 1$.

For Analyticity we substitute P-Analyticity: the claim that, as far as the opinions of the student are concerned, the probabilistic version of the Principles count as analytic to the concept of warranted degree of belief. (Hence the student is certain of every consequence of them.)

For the Iron Law we make no substitution (but observe that in the presence of Analyticity, the student is certain that there can be at most one exam).

Consistency and Closure are the sticking points. First we should reconceive them, thus:

Consistency*: If $J_i p$, then p is not a contradiction.

Closure₁: If $J_i p$, and q is a consequence of p , then $J_i q$.

Closure₂: If for all $p \in S$, $J_i p$, and q is the conjunction of the members of S , then $J_i q$.

Observe that nothing has been lost, in that the conjunction of Consistency and Closure is equivalent to the conjunction of Consistency*, Closure₁, and Closure₂. For Consistency* follows from Consistency, and likewise Closure₁, and Closure₂ each follow from Closure. To see the converse, let q be the conjunction of those proposition p such that $J_i p$. Any consequence of these propositions is a consequence of q . By Closure₂, $J_i q$ holds; hence by Closure₁, Closure holds, and by Consistency*, q is not a contradiction, and so Consistency holds.

Our representation of justified belief as justified degree of belief at or above the threshold automatically yields Consistency* and Closure₁; so it is Closure₂ that is the sticking point. It quite clearly cannot be recovered. But happily, we do not need it, as all of the arguments in the text—save one—can be reconstructed relying only upon the following corollary:

Closure₃: If $\Pr_i(p_1) = \Pr_i(p_2) = \dots = \Pr_i(p_n) = 1$ and $J_i q$, then $J_i(p_1 \ \& \ \dots \ \& \ p_n \ \& \ q)$.

Since $\text{Pr}_i(p_1) = \text{Pr}_i(p_2) = \dots = \text{Pr}_i(p_n) = 1$ and $\text{Pr}_i(q)$ implies that $\text{Pr}_i(p_1 \& \dots \& p_n \& q)$, Closure₃ also follows from our account of justified belief.

I'll now rehearse two representative arguments from the text. The trick will be to make explicit those points at which Closure₃ is used, and then check to see that its use is indeed legitimate.

I argued, for example, that if Persistence holds, the announcement is true, and the student justifiably believe it, then the exam cannot be on Friday:

- | | | |
|----|---|--------------------------------|
| 1. | SE_5 | hypothesis |
| 2. | $J_1(SE_5)$ | hypothesis |
| 3. | E_5 | hypothesis for <i>reductio</i> |
| 4. | $J_1(E_1 \vee E_2 \vee E_3 \vee E_4 \vee E_5)$ | 2, Principles |
| 5. | $J_5(E_1 \vee E_2 \vee E_3 \vee E_4 \vee E_5)$ | 4, Persistence |
| 6. | $J_5(\neg E_1 \& \neg E_2 \& \neg E_3 \& \neg E_4)$ | 3, Principles |
| 7. | $J_5(E_5)$ | 5,6, Principles |
| 8. | $\neg J_5(E_5)$ | 1,3 |

For which the appropriate reconstruction is this:

- | | | |
|-----|---|--------------------------------|
| 1. | SE_5 | hypothesis |
| 2. | $J_1(SE_5)$ | hypothesis |
| 3. | E_5 | hypothesis for <i>reductio</i> |
| 4. | $J_1(E_1 \vee E_2 \vee E_3 \vee E_4 \vee E_5)$ | 2, Closure ₁ |
| 5. | $J_5(E_1 \vee E_2 \vee E_3 \vee E_4 \vee E_5)$ | 4, Persistence |
| 5.1 | $\neg E_1 \& \neg E_2 \& \neg E_3 \& \neg E_4$ | 3, Iron Law |
| 5.2 | $J_5(\neg E_1) \& \dots \& J_5(\neg E_4)$ | 5.1, Memory |
| 6. | $J_5(\neg E_1 \& \neg E_2 \& \neg E_3 \& \neg E_4)$ | 5.2, Closure ₃ |
| 6.1 | $J_5((\neg E_1 \& \dots \& \neg E_4) \& (E_1 \vee \dots \vee E_5))$ | 5,6, Closure ₃ |

- | | |
|--------------------|---------------------------|
| 7. $J_5(E_5)$ | 6.1, Closure ₁ |
| 8. $\neg J_5(E_5)$ | 1,3, Iron Law |

Step 6 is legitimate, since what Memory gives us in 5.2 is in fact $\text{Pr}_5(\neg E_1) = \dots = \text{Pr}_5(\neg E_4) = 1$; it follows that in 6 we in fact have $\text{Pr}_5(\neg E_1 \ \& \ \neg E_2 \ \& \ \neg E_3 \ \& \ \neg E_4) = 1$, hence 6.1 is legitimate as well.

Let us next examine the student's argument that the exam cannot take place on Friday. It will be easiest if we do not express this as a piece of natural deduction, but make explicit those places where conditionals are introduced. Thus:

- | | | |
|---|---|-----|
| 1. E_5 | $J_5(\neg E_1 \ \& \ \neg E_2 \ \& \ \neg E_3 \ \& \ \neg E_4)$ | |
| 2. $J_5(E_1 \vee E_2 \vee E_3 \vee E_4 \vee E_5)$ | | |
| 3. E_5 | $J_5(E_5)$ | 1,2 |
| 4. E_5 | $\neg J_5(E_5)$ | |
| 5. $\neg E_5$ | | 3,4 |

We are assuming that the student is justified in believing the announcement, and so (thanks to the Iron Law, Closure₁, and an application of Closure₃ whose legitimacy is easily checked) justified in believing 4. We are taking for granted that Confidence yields his justified belief in 2. It is, finally, easily verified that he is justified in believing 1, and in fact justified in believing the stronger conditional $E_5 \quad \text{Pr}_5(\neg E_1 \ \& \ \neg E_2 \ \& \ \neg E_3 \ \& \ \neg E_4) = 1$. It follows that the tacit use of Closure₃ to derive 3 is legitimate, and hence that the argument is valid.

But even though it is valid, and the student is justified in believing the premises 1, 2, and 4, it does not follow that he is justified in believing the conclusion 5. This *will* follow, provided he is not only justified in believing but (justifiably) *certain* of at least two of these three premises. It is easily checked that our principles entail that he is justifiably certain of 1. But they do not entail that he is justifiably certain of 2, or that he is justifiably certain of 4.

Happily, this does not matter, since our principles do entail that if he is *not* justifiably certain of 4, then he *is* justifiably certain of 2. As follows: Suppose he is not justifiably certain of 4. It follows that $\Pr_1(\neg J_5(E_5) \mid \neg E_1 \ \& \ \dots \ \& \ \neg E_4) = 1$, hence that $\Pr_1(J_5(E_5) \mid \neg E_1 \ \& \ \dots \ \& \ \neg E_4) = 0$. But the student is certain about what his Friday state of opinion will be, on the supposition that the exam has not taken place by then. So if $\Pr_1(J_5(E_5) \mid \neg E_1 \ \& \ \dots \ \& \ \neg E_4) = 0$, then it must be that $\Pr_1(J_5(E_5) \mid \neg E_1 \ \& \ \dots \ \& \ \neg E_4) = 1$. It follows that $\Pr_1(J_5(E_1 \vee \dots \vee E_5) \mid \neg E_1 \ \& \ \dots \ \& \ \neg E_4) = 1$; since our principles entail that $\Pr_1(J_5(E_1 \vee \dots \vee E_5) \mid E_1 \vee \dots \vee E_4) = 1$, it likewise follows that $\Pr_1(J_5(E_1 \vee \dots \vee E_5)) = 1$.

So even when we recognize that the picture of belief that underlies the original Principles is too crude, and that the probabilistic replacement entails that there will be valid arguments whose premises are justifiably believed, but whose conclusion is not, we can also see that the student's argument is not one of these. (Of course that is small comfort, given its circularity.)

Finally, we must examine the argument that from Confidence and the Principles we can conclude that $J_1(SE_n)$ entails $J_1(SE_{n-1})$:

- | | |
|---|-----------------|
| 1. $J_1(SE_n)$ | hypothesis |
| 2. $J_1(E_1 \vee \dots \vee E_n)$ | 1, Principles |
| 3. $J_1(J_n(E_1 \vee \dots \vee E_n))$ | 2, Confidence |
| 4. $J_1(E_n \ \& \ J_n(\neg E_1 \ \& \ \dots \ \& \ \neg E_{n-1}))$ | Principles |
| 5. $J_1(E_n \ \& \ J_n(E_n))$ | 3,4, Principles |
| 6. $J_1(E_n \ \& \ \neg J_n(E_n))$ | 1, Principles |
| 7. $J_1(\neg E_n)$ | 5,6, Principles |
| 8. $J_1(SE_{n-1})$ | 1,7, Principles |

This argument *cannot* be successfully reconstructed as it stands. It is fine up through step 7. But the tacit appeal to Closure₃ that yields 8 is not legitimate, as the following model demonstrates. For suppose there is a sharp belief threshold $\theta = 0.9$, and consider the following probabilities:

Distribution 7

<u>day</u>	<u>initial probability</u>	<u>surprise probability</u>
Monday	0.3	0.3
Tuesday	0.3	0.43
Wednesday	0.3	0.75
Thursday	0.098	0.98
Friday	0.001	0.5
no exam	0.001	

Since the student is certain that the exam will not be a surprise, on the assumption that it takes place on Thursday, only the other four days contribute to $\Pr_1(SE_5)$. Still, since $\Pr_1(SE_5) = 0.901 > \dots$, $J_1(SE_5)$ holds. And, since the student is certain that $J_5(E_1 \vee \dots \vee E_5)$ will hold, on the supposition that the exam takes place before Friday (for he is certain that if so, he will remember it, etc.), $\Pr_1(J_5(E_1 \vee \dots \vee E_5)) = 0.998$; hence $J_1(J_5(E_1 \vee \dots \vee E_5))$.

This legitimates premise 1 in the argument, as well as the use of Confidence to yield 3. And, since $\Pr_1(\neg E_5) = 0.999$, the conclusion at 7 is correct. But since Thursday does not contribute, $\Pr_1(SE_4) = 0.9$ —too low to count as belief.

But for our purposes, all that is important is that from $J_1(SE_n)$ and a single application of Confidence, $J_1(SE_{n-1})$ follows. The trick is to target that application not at a consequence derived from $J_1(SE_n)$ (viz., $J_1(E_1 \vee \dots \vee E_n)$), but at $J_1(SE_n)$ itself. That is, we argue that from $J_1(SE_n)$ and $J_1(J_n(SE_n))$, $J_1(SE_{n-1})$ follows. This will work, provided we add a further memory condition:

Memory*: for all i, k with $k > i$, $J_i(p) \rightarrow J_k(J_i(p))$,

for which the following is the probabilistic analog:

P-Memory*: for all $x \in [0,1]$ and i, k with $k > i$, $\Pr_i(p) = x \rightarrow \Pr_k(\Pr_i(p) = x) = 1$.

In other words, we take the student to be infallible with respect to his past probabilities. Observe that P-Memory* validates Memory*.

An argument can now be constructed using the J-Principles—and in particular, only legitimate applications of Closure₃—that shows that $J_1(SE_n)$ and $J_1(J_n(SE_n))$ entail $J_1(SE_{n-1})$. As it's a bit tedious, and leaves the central ideas obscure, I'll leave it as an exercise, and present the reasoning more informally, and directly in terms of probabilities.

Suppose, then, that $J_1(SE_n)$ and $J_1(J_n(SE_n))$ —i.e., that $\Pr_1(SE_n) > \frac{1}{2}$ and that $\Pr_1(\Pr_n(SE_n) > \frac{1}{2}) > \frac{1}{2}$. First, the student is certain that if a surprise exam takes place in the first $n - 1$ days, then he will, come the morning of day n , remember that this is so; since SE_{n-1} entails SE_n , it follows that $\Pr_1(\Pr_n(SE_n) = 1 \mid SE_{n-1}) = 1$, and so of course that $\Pr_1(\Pr_n(SE_n) > \frac{1}{2} \mid SE_{n-1}) = 1$. Next, consider what the student should believe about his day- n probabilities, on the supposition that SE_{n-1} is false. This could come about in either of two ways. On the one hand, it might be that no exam takes place in the first $n - 1$ days; the student is certain that if so, he will remember, and so $\Pr_n(SE_n)$ will equal $\Pr_n(E_n \ \& \ \neg J_n(E_n)) = \Pr_n(E_n \ \& \ \Pr_n(E_n) > \frac{1}{2}) =$ at most $\frac{1}{2}$. On the other hand, it might be that an exam takes place but does not come as a surprise, rendering not only SE_{n-1} but SE_n false; the student is likewise certain that if so, he will remember (here is where P-Memory gets used), and so $\Pr_n(SE_n)$ will equal 0. Either way, he is certain that $\Pr_n(SE_n) \leq \frac{1}{2}$; it follows that $\Pr_1(\Pr_n(SE_n) > \frac{1}{2} \mid \neg SE_{n-1}) = 0$.

But

$$\begin{aligned} \Pr_1(\Pr_n(SE_n) > \frac{1}{2}) &= \Pr_1(\Pr_n(SE_n) > \frac{1}{2} \mid SE_{n-1})\Pr_1(SE_{n-1}) + \Pr_1(\Pr_n(SE_n) > \frac{1}{2} \mid \neg SE_{n-1})\Pr_1(\neg SE_{n-1}) \\ &= 1 \cdot \Pr_1(SE_{n-1}) + 0 \cdot \Pr_1(\neg SE_{n-1}) \\ &= \Pr_1(SE_{n-1}). \end{aligned}$$

Since $\Pr_1(\Pr_n(SE_n) > \frac{1}{2}) > \frac{1}{2}$, it follows at once that $\Pr_1(SE_{n-1}) > \frac{1}{2}$ —i.e., that $J_1(SE_{n-1})$.

Appendix B: Probabilistic Confidence used to analyze Confidence

Choose $\frac{1}{2} < p < 1$, $q \in [0,1]$, with $q > p$. Suppose that $\Pr_1(p) = \frac{1}{2}$, and that Probabilistic Confidence holds, together with the assumption that all future evidence will be veridical, and properly incorporated. Then for all q and times $t > 1$, $\Pr_1(q \mid \Pr_t(q) = x) = x$. We will now investigate the

following question: What is the maximum value of x such that it must be the case that $\Pr_1(\Pr_t(p) > \frac{1}{2}) > x$? Or, interpreting $\frac{1}{2}$ as the belief threshold, what minimum degree of belief must the agent have in the proposition that, come time t , he will believe q ?

Suppose then that $\Pr_1(\Pr_t(p) > \frac{1}{2}) = x$; our question then becomes, For what values of x is this condition inconsistent with our assumptions that $\Pr_1(p) = \frac{1}{2}$, and that Probabilistic Confidence holds? To find out, we first write

$$\begin{aligned} \Pr_1(p) &= \Pr_1(p \mid \Pr_t(p) > \frac{1}{2})\Pr_1(\Pr_t(p) > \frac{1}{2}) + \Pr_1(p \mid \Pr_t(p) \leq \frac{1}{2})\Pr_1(\Pr_t(p) \leq \frac{1}{2}) \\ &= \Pr_1(p \mid \Pr_t(p) > \frac{1}{2}) \cdot x + \Pr_1(p \mid \Pr_t(p) \leq \frac{1}{2}) \cdot (1 - x). \end{aligned}$$

Observe that $\Pr_1(p)$ takes its maximum value when $\Pr_1(p \mid \Pr_t(p) > \frac{1}{2}) = \Pr_1(p \mid \Pr_t(p) = 1) = 1$ and $\Pr_1(p \mid \Pr_t(p) \leq \frac{1}{2}) = \Pr_1(p \mid \Pr_t(p) = 0) = \frac{1}{2}$; this maximum value is $x + \frac{1}{2} - x$. Similarly, since we must have $\Pr_1(p \mid \Pr_t(p) > \frac{1}{2}) > \frac{1}{2}$, the possible values of $\Pr_1(p)$ have a greatest lower bound of $\frac{1}{2}$. So we conclude that from $\Pr_1(\Pr_t(p) > \frac{1}{2}) = x$ it follows that $\Pr_1(p) \in (\frac{1}{2}, x + \frac{1}{2} - x]$.³⁹ Finally, since $\Pr_1(p \mid \Pr_t(p) > \frac{1}{2})$ can take on any value in $(\frac{1}{2}, 1]$ and $\Pr_1(p \mid \Pr_t(p) \leq \frac{1}{2})$ can take on any value in $[0, \frac{1}{2}]$, it follows that $\Pr_1(p)$ can take on any value in $(\frac{1}{2}, x + \frac{1}{2} - x]$. So consistency requires only that $\frac{1}{2} \in (\frac{1}{2}, x + \frac{1}{2} - x]$. Since it is guaranteed that $\frac{1}{2} > \frac{1}{2}$, the requirement becomes: $\frac{1}{2} \leq x + \frac{1}{2} - x$.

The right-hand side is monotonically increasing in x . We get an inconsistency, therefore, exactly if $x < a$, where a is the solution to $\frac{1}{2} = a + \frac{1}{2} - a$. Writing $\frac{1}{2} = 1 - \frac{1}{2}$ and $\frac{1}{2} = 1 - \frac{1}{2}$, we have $a = 1 - \frac{1}{2}$. So it follows from $\Pr_1(p) = 1 - \frac{1}{2}$ that $\Pr_1(\Pr_t(p) > \frac{1}{2}) = 1 - \frac{1}{2}$.

³⁹Note that for $x \in [0, 1]$, it is guaranteed that $x + \frac{1}{2} - x \geq \frac{1}{2}$. For this requires only that $x(1 - 2) \geq -\frac{1}{2}$, which is immediate if $x \leq \frac{1}{2}$. If $x > \frac{1}{2}$, we need to show that $x \geq \frac{1}{2} + \frac{1}{2}(2 - 1)$; but since $1 \geq x$, $2 - 1 \geq 1$, and so $\frac{1}{2}(2 - 1) \geq \frac{1}{2}$.