**Noname manuscript No.**
(will be inserted by the editor)

# How to Expect a Surprising Exam

**Brian Kim · Anubav Vasudevan**

Received: date / Accepted: date

**Abstract** In this paper, we provide a Bayesian analysis of the well-known surprise exam paradox. Central to our analysis is a probabilistic account of what it means for the student to accept the teacher's announcement that he will receive a surprise exam. According to this account, the student can be said to have accepted the teacher's announcement provided he adopts a subjective probability distribution relative to which he expects to receive the exam on a day on which he expects not to receive it. We show that as long as expectation is not equated with subjective certainty there will be contexts in which it is possible for the student to accept the teacher's announcement, in this sense. In addition, we show how a Bayesian modeling of the scenario can yield plausible explanations of the following three intuitive claims: (1) the teacher's announcement becomes easier to accept the more days there are in class; (2) a strict interpretation of the teacher's announcement does not provide the student with any categorical information as to the date of the exam; and (3) the teacher's announcement contains less information about the date of the exam the more days there are in class. To conclude, we show how the surprise exam paradox can be seen as one among the larger class of paradoxes of doxastic fallibilism, foremost among which is the paradox of the preface.

**Keywords** Surprise Exam Paradox, probability, fallibilism, Preface Paradox, Bayesianism

## 1 Introduction

A teacher announces to one of his students prior to the first meeting of class that at some point during the term the student will be given a surprise exam. The teacher's announcement, as innocuous as it may seem, gives rise to a well-known paradox in light of the following argument purporting to establish that what the teacher has said cannot be true:

> The surprise exam cannot be given on the last day of class. For if this were so then just prior to the last class, having attended all the previous classes without receiving the exam, the student would expect to receive the exam that day, in which case the exam would not

Brian Kim
Oklahoma State University
E-mail: brian.kim@okstate.edu

Anubav Vasudevan
University of Chicago
E-mail: anubav@uchicago.edu

> be a surprise.[1] Neither, however, can the exam be given on the next-to-last day of class. For if this were so then just prior to the next-to-last class, having attended all but the last two days of class without receiving the exam and having already concluded that the exam cannot be given on the last day of class, the student would expect to receive the exam that day, in which case the exam would not be a surprise. Continuing in this way, it follows that no matter when the exam is given it cannot be a surprise.

The conclusion of this argument is, of course, absurd since it is clearly possible for the student to receive a surprise exam. Where then does the above line of reasoning go wrong?[2]

As a first attempt at answering this question it may be noted that the above argument appears to conflate (1) the claim that the student will receive a surprise exam with (2) the claim that the student *accepts* that this will be so. While it is the first of these two claims that is the intended target of the reductio, it is the second that is appealed to in order to show that the student cannot receive a surprise exam on the last day of class.

Thus, one way of diagnosing the fallacy in the argument is to note its reliance on the unstated assumption that the student *accepts* the teacher's announcement.[3] Unfortunately, this observation alone does not constitute a satisfactory resolution of the paradox. For while it does allow one to avoid the absurd conclusion that no exam administered to the student can ever be a surprise, it does nothing to address the equally absurd claim that the student can only receive a surprise exam provided he does not accept what the teacher has told him. It is, after all, clearly possible for the student to accept the teacher's announcement and yet be surprised when the teacher gives him the exam on, say, the third or the fourth day of class.

A satisfactory resolution to the paradox must therefore explain not only why the paradoxical argument does not rule out the possibility of the student receiving a surprise exam, but also why it does not rule out the possibility of a surprise exam being given to a student to whom it has been announced and *by whom it has been accepted* that this will be so. Assuming that the student is clever enough to detect any inconsistency that might exist between his acceptance of the teacher's announcement and this announcement's being true, the challenge raised by the surprise exam paradox can thus be put more succinctly as that of explaining how the student can *coherently accept* that he will receive a surprise exam.

In order to respond to this challenge a more precise account must be offered of the constraints imposed on the student's doxastic state by the requirement that he "accept" the teacher's announcement. In section 2, we propose a Bayesian account of what it means for the student to accept that he will receive a surprise exam, and show how this account successfully manages to avoid the paradox. In sections 3 and 4, we defend the plausibility of this Bayesian model by explaining how it can make sense of the following three intuitions:

[1] Strictly speaking, the teacher's announcement does not imply that the student will receive exactly one surprise exam during the term. In what follows, when we speak of "the surprise exam" (or "the exam") this phrase may be understood as referring to the *first* surprise exam that the student will receive.

[2] The paradox in it various forms has a long and well-documented history. While a similar scenario to the surprise exam paradox was described in [O'Connor, 1948], curiously, it seems to have only been first acknowledged that a surprise event could occur despite its having been announced in [Scriven, 1951]. A detailed history of the paradox's origins can be found in [Sorensen, 1988], ch. 7. For a survey of the various approaches to the problem appearing in the literature prior to 1983, see [Margalit and Bar-Hilel, 1983]. A more recent survey can be found in [Chow, 1998]. The most recent online version of Chow's paper (which can be found at `arXiv:math/9903160v4 [math.LO]`) contains a comprehensive bibliography listing all the papers that are relevant to the paradox published prior to June 2011.

[3] This is essentially the solution proposed by [Quine, 1953] except that Quine describes the illicit assumption as the assumption that the student *knows* (rather than accepts) that he will receive a surprise exam prior to the last day of class. Quine rejects this assumption on the grounds that if the student is prepared to accept the teacher's announcement as contradictory after following through with the fallacious line of reasoning, then he could not have known that he would receive a surprise exam on the basis of the teacher's announcement. [Kripke, 2011] criticizes Quine's response to the problem on the grounds that under normal conditions the student does, in fact, know that he will receive a surprise exam upon hearing the teacher's announcement.

1. The teacher's announcement becomes easier to accept the more days there are in class.
2. A strict interpretation of the teacher's announcement does not provide the student with any categorical information as to the date of the exam.
3. The teacher's announcement contains less information about the date of the exam the more days there are in class.

In section 5, we discuss the connection between the surprise exam paradox and other paradoxes of doxastic fallibilism, foremost among which is the paradox of the preface.

## 2 How to Expect a Surprising Exam

In order to describe the constraints imposed on the student's doxastic state as a result of accepting the teacher's announcement, we must first come to a clearer understanding of what is meant by a "surprise exam". A surprise exam, in the usual sense of the phrase, refers to an exam that is not explicitly announced in advance of its being given. This, however, cannot be the phrase's intended meaning in the context of the paradox, for there is no inconsistency in the student receiving an unannounced exam on a day on which he expects to receive it and hence no inconsistency in such an exam being given on the last day of class.

At a minimum, in order for the paradox to arise it must be the case that an exam's being a "surprise" carries with it the implication that on the day on which the exam is given the student does not expect to receive it. Even granting this much, however, there remain a number of distinct ways of understanding what is meant by a "surprise exam". While each of these interpretations gives rise to a slightly different form of the paradox, in what follows we will opt for what we take to be the most literal construal of the phrase, namely, that according to which a surprise exam is an exam that the student is surprised to receive or an exam that the student finds *surprising*.

As natural a proposal as this may seem, it differs in certain crucial respects from the definitions appearing most often in the literature. Take for example the oft-proposed definition of a surprise exam as an exam of which the student has no prior knowledge. The first point to note about this definition is that it lacks precision. Since knowledge is a compound notion involving both doxastic and justificatory components, denials of knowledge have a disjunctive form. Thus, a student has no prior knowledge of an exam if at the time of the exam either (1) the student does not believe (or expect) that he will receive an exam, or (2) his belief (or expectation) that he will receive an exam is unwarranted. Clearly, however, for the purpose of defining what it means for the student to be surprised to receive the exam, it is only the first of these two scenarios that is relevant. Thus, a more precise definition along these lines ought to make it explicit that in denying an agent's prior knowledge of an event the intent is to deny the doxastic and not the justificatory implications of the rejected knowledge claim.

Once the definition has been made more precise in this way, it is still to be distinguished from the literal interpretation of a surprise exam as one which the student is surprised to receive. After all, to say of someone that he had no prior belief or expectation that a certain event would occur is not the same as to say that he found that event surprising. One does not, for example, expect that a fair coin will land heads, but one would not be surprised to discover that it had done so. Or, again, in a three-horse race in which two of the three horses are equally favored to win, the third being a significant underdog, while it may be true that one does not expect of any one of the three horses that it will be the winner (the choice between the two favorites being too close to call), the claim that one would be surprised to discover which horse won is not uninformative. On the contrary, it clearly indicates that it was the underdog that has won the race.

Such examples suggest that to say of an event that one found it surprising is not merely to say that one did not expect that event to occur, but is rather to make the stronger claim that one expected that that event would *not* occur.[4] It is this fact that accounts for the infelicitousness of the phrase: “I would find it surprising were I to receive an exam today, but at the same time I would find it surprising were I not to.” If surprise merely denoted a lack of prior expectation on the part of an agent such a phrase would express a perfectly ordinary state of mind since it often happens (as in the case of a fair coin’s landing heads) that one has no prior expectation one way or the other as to whether or not a given event will occur. The infelicitousness of the phrase instead turns on the fact that if expectation is given its usual meaning, one cannot both expect an event not to occur (and so be surprised when it does) and at the same time expect its non-occurrence not to occur (and so be surprised when it does not).

In light of the above remarks, we will not adopt the usual (negative) definition of a surprise exam as one which the student does not expect to receive, but instead will define a surprise exam as one which the student (positively) expects not to receive at the time at which he receives it.[5] The most natural framework in which to model such an expectation is the framework of subjective probabilities. In the context of this framework, we say that an agent expects an event to occur provided the subjective probability he assigns to that event exceeds a certain context-sensitive threshold. Correspondingly, an event is expected not to occur—and is thus surprising—provided the probability of its non-occurrence exceeds this same threshold.[6]

In order to determine whether or not the student would be surprised to receive the exam on the $n$th day of class, we must therefore determine the subjective probability that he would assign to the event that the exam will be given on that day, supposing he had attended the first $n-1$ days of class without receiving the exam. To compute this probability, we will make the following two simplifying assumptions: (1) at any given time, apart from the teacher’s announcement, the only relevant information which the student possesses is whether or not he has as yet received the exam; and (2) the student processes this information by updating his initial subjective probabilistic assessment of when the exam will be given via the rule of Bayesian conditionalization.[7]

[4] Admittedly, there is plausible reading of surprise according to which a surprising event is one which is unexpected. But here “unexpected” has a meaning akin to that of the term “undesirable”, where to describe a certain state of affairs in this way is not merely to say that one lacks any desire that it should be so, but is rather to claim that one desires that it should *not* be so.

[5] There are at least three limitations of this account that prevent it from serving as a satisfactory general definition of a surprising event. First, whereas a surprising event is here defined relative to his expectations at the time of that event’s occurrence, strictly speaking, a surprising event ought to be defined relative to the expectations of the agent at the time at which he *learns* of that event’s occurrence. Second, the proposed account ignores the possibility of an agent’s being surprised by an event the occurrence of which he has simply failed to consider. Thus, for example, John may be surprised to discover that there is a spaceship hovering in the sky outside his window despite the fact that, prior to looking out his window, he may not have explicitly formed any expectation of observing a spaceshipless sky. Third, to describe an event as surprising not only implies something about one’s epistemic state at the time at which one learned of its occurrence, but also says something about the phenomenology of the discovery itself. A surprising event is not just unexpected but is also, in some sense, startling or shocking. We will assume that for the purposes of analyzing the paradox such subtleties may be ignored.

[6] Probabilistic approaches to the paradox are not new. In particular, [Clark, 1994a], [Clark, 1994b], [Hall, 1999], [D. Borwein and Marechal, 2000] and [Schumacher and Westmoreland, 2008] all adopt probabilistic approaches to the problem. The latter two, in particular, are worth mentioning. [D. Borwein and Marechal, 2000] proposes a fascinating analysis of the paradox in which no categorical criterion is proposed as to when an exam is to count as a surprise exam, but is instead framed entirely in terms of the degree to which an exam is surprising. The specific question that the paper asks is what probability distribution maximizes the student’s expected degree of surprisal, as measured by the negative log of the probability. [Schumacher and Westmoreland, 2008] proposes a very similar framework to that developed in this paper. However, their account of what it means to coherently accept the teacher’s announcement differs substantially from our own.

[7] In what follows, we are not always careful to distinguish between conditionalization as a norm governing the static process of suppositional reasoning, and conditionalization as a norm governing the dynamic process of belief revision. While we often give weight to considerations of fluency or style over those of accuracy, in fact, it would be more true to our intention to think of conditions (1) and (2) as norms governing the student’s suppositional reasoning immediately following the teacher’s announcement, so that all that we are concerned with is the internal coherence of the student’s doxastic state at that particular time.

Suppose that the class meets $N$ times. Let $E_n$ $(n = 1, 2, \dots, N)$ be the event that the exam is administered on the $n$th day of class, and let $E_{N+1}$ be the event that no exam is given at all. We will model the doxastic state of the student just after the teacher has made his announcement by an assignment of probabilities to the events $E_1, E_2, \dots, E_{N+1}$. Given the above two assumptions, each such assignment of probabilities uniquely determines how confident the student will be just prior to the $n$th day of class that he will not receive the exam that day, on the condition that he has not yet received it. In particular, for the probability $P$, this value is given by:[8]

$$\sigma_P(E_n) = P(\neg E_n | \neg E_1 \wedge \cdots \wedge \neg E_{n-1})$$

If $\alpha$ is the threshold above which an assignment of probability to an event implies that its non-occurrence will be found surprising, then the event which corresponds to a student with probability $P$ receiving a surprise exam is the event that the exam is given on a day $n$ for which:

$$\sigma_P(E_n) \geq \alpha.$$

In what follows, we will refer to this event as the "surprise event" and denote it by $S(P, \alpha)$.

The dependence of the surprise event on the probability distribution $P$ marks the first noteworthy feature of the teacher's announcement, namely, that its truth conditions depend on the doxastic state adopted by the student in response to this very announcement. This is a feature shared by all future-tensed predictions of surprise, provided that the person of whom the surprise is predicted is one to whom the prediction is made known.[9]

The second important point to note about the surprise event is its dependence on the parameter $\alpha$. In what follows, we refer to $\alpha$ as the *surprise index.* The value of this parameter determines how confident the student must be that he will not receive the exam, at the time at which it is given, in order for the exam to count as a surprise. If $\alpha = 1$, then the exam only counts as a surprise if it is administered on a day on which the student is absolutely certain that he will not receive it. If, on the other hand, $\alpha = 2/3$, it is enough for the exam to count as a surprise that it be administered on a day on which the student judges it to be at least twice as likely as not that he will not receive it. The intuition that one cannot find surprising both the occurrence and non-occurrence of an event, requires that $\alpha$ be assigned a value in the half-open interval $(1/2, 1]$, but, apart from this fact, we leave it open what additional factors may be relevant for determining the specific value of $\alpha$. In our view, $\alpha$ should not be thought of as a semantic constant, the value of which is fixed by the meaning of the term 'surprise', but rather as a context-sensitive parameter, whose evaluation is one essential part of the interpretive task of making sense of the teacher's announcement. In this sense, the value of $\alpha$ reflects a judgment, on the part of the student, as to how "strictly" he takes the teacher to be speaking in characterizing the exam as a surprise. In forming such a judgment, the student must appeal not only to the content of the teacher's assertion but also to various incidental features of the context in conjunction with certain rules governing conversational implicature.

[8] We adopt the usual convention that the conjunction of an empty set of events is the necessary event, $\top$, and that the disjunction of an empty set of events is the impossible event, $\bot$. In accordance with this convention, it follows from the above definition that $\sigma_P(E_1) = P(\neg E_1|\top) = P(\neg E_1)$. The quantity $\sigma_P(E_n)$ is only well defined if $P(\neg E_1 \wedge \cdots \wedge \neg E_{n-1}) > 0$. If this is not the case, we will stipulate that $\sigma_P(E_n) = 0$.

[9] If, for example, John is told by Lev that he will be surprised by the ending of the book, then what must hold of the conclusion of the book in order for Lev's statement to be true depends on how John responds to Lev's remark. If, in response to Lev's remark, John is led to expect that it was the Butler who committed the murder, then Lev has spoken truly just in case the Butler did not do it, for it is only in this case that John's expectations will be confounded. If, on the other hand, John comes to believe on the basis of Lev's remark that it was the Colonel who did it, then if it turns out that the true culprit was the Butler, Lev will have spoken truly. It may be noted, in contrast, that if John is told by Lev that he *would have been* be surprised by the ending of the book, then what must hold of the conclusion of the book in order for Lev's statement to be true, depends only on John's doxastic state prior to Lev's remark.

As an illustration of such context-sensitivity consider the three-horse race example described above. In that case, the conversational maxim that instructs the listener to interpret the speaker as supplying some non-trivial information requires that the surprise index be set no lower than the probability with which either of the two favorites will lose the race. Otherwise, the claim that one would be surprised to discover which horse had won would be compatible with any one of the three horses being the winner and thus would be wholly uninformative.[10] At the same time, the requirement that the speaker's assertion be given a consistent interpretation requires that the surprise index be set at a value no higher than the probability with which the underdog will lose the race. Otherwise, the speaker's assertion would contradict the claim that one of the three horses must be the winner.

At present, we need not enter into a discussion of the specific contextual factors that ought to inform the student's choice of $\alpha$ in the specific case of the surprise exam paradox. For, as it turns out, the paradoxical argument does not depend on the value of this parameter. In probabilistic terms, what this argument shows is that the student cannot coherently accept *with certainty* that he will receive a surprise exam. This fact is summarized by the following proposition:

**Theorem 1** *There is no probability $P$ and $\alpha \in (1/2, 1]$, such that $P(S(P,\alpha)) = 1$.*

*Proof* Suppose, for contradiction, that $P(S(P,\alpha)) = 1$. We will show that $S(P,\alpha)$ implies $\neg E_n$, for all $n = 1, \ldots, N+1$. The proof will proceed by backwards induction on $n$.

Since, for any probability $P$, $\sigma_P(E_{N+1}) = 0$, it follows from the fact that $\alpha > 0$ that $S(P,\alpha)$ implies $\neg E_{N+1}$.

Now, choose $k$ such that $0 \leq k < N$ and suppose that $S(P,\alpha)$ implies $\neg E_n$, for $n = N-k+1, \ldots, N+1$. Then, $S(P,\alpha)$ implies $\neg E_{N-k+1} \wedge \cdots \wedge \neg E_{N+1}$ . If $P(\neg E_1 \wedge \cdots \wedge \neg E_{N-k-1}) = 0$, then $\sigma_P(E_{N-k}) = 0$ (see fn. 13), and so $S(P,\alpha)$ implies $\neg E_{N-k}$. If $P(\neg E_1 \wedge \cdots \wedge \neg E_{N-k-1}) > 0$, then since $P(S(P,\alpha)) = 1$, it follows that $P(\neg E_{N-k+1} \wedge \cdots \wedge \neg E_N) = 1$, and so

$$\begin{aligned}\sigma_P(E_{N-k}) &= P(\neg E_{N-k} | \neg E_1 \wedge \cdots \wedge \neg E_{N-k-1}) \\ &= P(\neg E_{N-k} | \neg E_1 \wedge \cdots \wedge \neg E_{N-k-1} \wedge \neg E_{N-k+1} \wedge \cdots \wedge \neg E_{N+1}) \\ &= P(\neg E_{N-k} | E_{N-k}) = 0.\end{aligned}$$

Hence, $S(P,\alpha)$ implies $\neg E_{N-k}$. This completes the induction. Since $S(P,\alpha)$ implies $\neg E_n$, for all $n = 1, \ldots, N+1$, $S(P,\alpha)$ is an impossible event and so must be assigned a probability of 0. But this contradicts the assumption that $P(S(P,\alpha)) = 1$.

We include the proof of this claim, not only to indicate how the student's argument can be formulated in the Bayesian framework, but also to indicate more clearly what is responsible for the difficulty. The crucial step in the argument relies on the fact that if an event is assigned a probability of 1, Bayesian conditionalization cannot alter this fact, so that such an event must still be assigned a probability of 1 no matter what else is supposed to have occurred. Thus, if the student is absolutely certain that he will receive a surprise exam, he must still be certain of this fact even on the supposition that the exam will not be given prior to the last day of class. From this, however, it follows that the student is certain that an exam given on the last day of class cannot be a surprise (for any value of $\alpha$). Thus, if the student is absolutely certain that he will receive a surprise exam, he is committed to being absolutely certain that the exam will be given prior to the last day of class, and this initiates the paradoxical induction.

[10] A similar pragmatic explanation can be given as to why it is wrong to assert of an arbitrarily chosen ticket in a fair lottery that one was surprised to discover that that ticket had won (despite the probability of its winning being extremely low). Since the lottery is fair, this assertion implies that the surprise index $\alpha$ has been set at a value such that the claim "the outcome of the lottery was surprising" supplies no information about the actual outcome of the lottery.

On a Bayesian analysis of the paradox, the fallacy thus consists in the assumption that after hearing the teacher's announcement the student is justified in assigning a subjective probability of 1 to the event that he will receive a surprise exam. Now, on its face, it might seem rather strange to describe the fallacy in this way. After all, numerically precise probabilities play no essential role in the argument as it was originally described. Thus, one might reasonably object that insofar as the Bayesian formulation of the paradox assumes a more specific modeling of the scenario than what is required in order to clearly express the paradoxical argument, any Bayesian diagnosis of the paradox will be insufficiently general.

Admittedly, were the sole aim of our analysis to identify the fallacy in the paradoxical argument such an objection would be well-put. With respect to this specific aim, a more perspicuous approach would be to depict the fallacy as resulting from the mistaken assumption that the teacher's announcement licenses the student in accepting "dogmatically" that he will receive a surprise exam (i.e., accepting not only that he will receive a surprise exam but that he will *always* accept that he will receive a surprise exam). The argument's reliance on this assumption can be made clear in a fully general, abstract setting in which only minimal assumptions are made about what it means to accept that an event will occur.[11] In the context of such a minimal framework, the assignment of probability 1 to an event can be seen as just one particular example of a dogmatic standard for acceptance.[12]

There exist in the literature several discussions of the paradox which diagnose the fallacy in precisely this way. As diagnoses of the fallacy, we are in complete agreement with these approaches to the paradox, and we concede that formulating the analysis in probabilistic terms does nothing to further clarify the issue. In our view, however, a complete resolution of the paradox should do more than correctly diagnose the source of the error—it must also propose a possible remedy. It is therefore not enough to point out that the argument assumes an absurdly dogmatic standard for acceptance relative to which it is (perhaps not surprisingly) impossible for the student to accept

[11] In this more general framework, the student's epistemic state is not modeled by a probability function, but instead by a sequence of modal operators $A_1, \ldots, A_N$, where $A_n(E)$ means that the student accepts $E$ just prior to the $n$th day of class. To say that acceptance is dogmatic is to say that $A_n(E)$ implies $A_n(A_m(E))$, for all $1 \leq n \leq m \leq N$. If acceptance is dogmatic, then given only very modest assumptions about acceptance, it can be shown that the student cannot coherently accept that he will receive a surprise exam. In other words, it can be shown that the following two claims imply that the student's acceptations at the time of the teacher's announcement are contradictory.

1. $A_1(E_1 \vee \cdots \vee E_N)$
2. $A_1(E_n \to A_n(\neg E_n))$, for $n = 1, \ldots, N$.

For an analysis of the paradox along these lines, see [Kripke, 2011].

[12] This can be made precise as follows. We define a probabilistic model of the modal language described in fn. 11 as a pair $(E_n, P)$, where $P$ is a probability function defined on the non-modal sentences of the language. We say that $(E_n, P)$ satisfies the proposition $A_m(\varphi)$ if a Bayesian agent with prior probabilities $P$ would assign a probability of 1 to $\varphi$ conditional on what he would know just prior to the $m$th day of class were the exam to be administered on day $n$, i.e.:

$$(E_n, P) \models A_m(\varphi) \text{ iff } P\left(\bigvee \{E_i : (E_i, P) \models \varphi\} | \psi_{n,m}\right) = 1$$

where:

$$\psi_{n,m} = \begin{cases} \neg E_1 \wedge \cdots \wedge \neg E_{m-1} & \text{if } m \leq n \\ E_n & \text{if } n < m \end{cases}$$

It is easy to confirm that, relative to this semantics, acceptance is dogmatic, i.e., for any $n \leq m$, and any sentence $\varphi$:

$$A_n(\varphi) \models A_n(A_m(\varphi))$$

what the teacher has said. One must also supply a more plausible standard for acceptance relative to which it can be shown that the student *can* coherently accept the teacher's announcement.

It is with regard to this latter objective that the Bayesian framework is useful, for it allows us to formulate a natural, non-dogmatic standard for acceptance in terms of probabilistic expectation. Specifically, in place of the requirement that the student assign a probability of 1 to the surprise event in order to accept the teacher's announcement, we may instead require only that he expect this event to occur in the sense of assigning to it a probability in excess of some context-sensitive threshold.

In what follows, we denote this threshold by $\beta$ and refer to it as the *acceptance index.* The value of the acceptance index determines how confident the student must be that he will receive a surprise exam in order to count as having "accepted" the teacher's announcement. If $\beta = 1$, then the student must be absolutely certain that he will receive a surprise exam (an attitude, which, as we have just seen, is incoherent) to qualify as having accepted the teacher's announcement. On the other hand, if $\beta = {}^2\!/_3$, then the student will count as having accepted the teacher's announcement provided he judges it to be at least twice as likely as not that he will receive a surprise exam.

So as to respect the intuition that there is no meaningful sense of acceptance according to which the student can be said to have accepted the teacher's announcement despite not expecting to receive a surprise exam, we will assume that $\beta$ must be assigned a value in the half-open interval $({}^1\!/_2, 1]$. As in the case of the surprise index $\alpha$, however, we will regard any further specification of $\beta$ as an interpretive act on the part of the student. In other words, the value of $\beta$ reflects a judgment on the part of the student as to how "strictly" he takes the teacher to be speaking in asserting that that the student *will*, in fact, receive a surprise exam. In order to form such a judgment, the student must appeal not only to the content of the teacher's assertion, but also to various incidental features of the context in conjunction with certain rules governing conversational implicature.

To illustrate the context-sensitivity of the acceptance index, consider a three-horse race in which one of the three horses is strongly favored to win, and suppose that we are conversing with a horse-racing expert about whether to place a bet that the favored horse will win. If it is understood by all that any information that the expert provides to us will be used to inform our betting behavior, then if the expert asserts that the favored horse will win, this can be taken to provide us with prima facie justification in placing a bet on the favored horse. Now, if the betting odds that the favored horse will win are 2:1, then in order to justify this bet, the expert's testimony would have to license an assignment of probability of at least ${}^2\!/_3$ to the event that the favored horse will win. So, in this context, the acceptance index would have to be greater than or equal to ${}^2\!/_3$. Suppose, however, that the odds on the favored horse winning were instead 3:1. In this context, relative to the same assertion by the expert, the acceptance index would have to be set at least ${}^3\!/_4$ in order to justify placing the bet. In this way, we can see how features of the context wholly incidental to the meaning of the expert's claim (e.g., the betting odds on the favored horse) can impose constraints on what counts as a reasonable standard for acceptance.[13]

We are now in a position to state precisely what it means for the student to accept the teacher's announcement. The student accepts the teacher's announcement (relative to $\alpha$ and $\beta$) provided he expects to receive a surprise exam, i.e., provided the student's probabilities satisfy the condition:

$$P(S(P, \alpha)) \geq \beta. \tag{1}$$

[13] Strictly speaking, these examples do not imply that $\beta$ depends on these contextual features, since, in both of these examples, the constraints only fix the lower-bound for the acceptance index. Scenarios in which precise (non-trivial) constraints are imposed on both the upper- and lower-bounds of $\beta$ are not so easy to fabricate. A general (albeit vague) constraint that applies in almost all contexts of assertion is that the value of $\beta$ be less than 1 since it is only in highly contrived settings that an assertion is understood as licensing a bet on the asserted claim at any odds whatsoever.

Since the only non-pragmatic constraint that we have imposed on the choice of $\alpha$ and $\beta$ is that they both be greater than $1/2$, we will adopt as a minimal standard for the coherent acceptance of the teacher's announcement that condition (1) be satisfied for some $\alpha, \beta \in (1/2, 1]$.

It is not difficult to see that, relative to this standard, there are many contexts in which it is possible for the student to coherently accept the teacher's announcement. Suppose, for example, that based on the teacher's announcement, the student adopts the uniform probability distribution given by:

$$P(E_n) = \begin{cases} 1/N & \text{if } n = 1, \ldots, N \\ 0 & \text{if } n = N+1 \end{cases} . \tag{2}$$

Then, by a simple calculation, it can be shown that $S(P, 2/3)$ is the event $E_1 \vee \cdots \vee E_{N-2}$, from which it follows that

$$P(S(P, 2/3)) = \frac{N-2}{N}.$$

Consequently, for $N \geq 6$:

$$P(S(P, 2/3)) \geq 2/3.$$

Thus, if the class meets more than six times, a student who believes that the exam is no more likely to be given on one day than on any other will judge it to be at least twice as likely as not that he will receive the exam on a day on which he will judge it to be at least twice as likely as not that he will not. Since, in this case, the student may coherently be described as expecting to receive a surprise exam, we regard this assignment of probabilities as a coherent way of accepting the teacher's announcement.

The argument purporting to show that the student cannot be given a surprise exam is paradoxical in that it appears to imply that the student cannot coherently accept the teacher's announcement—a fact which is, on its face, absurd. As we have seen, however, if we interpret the student's doxastic state probabilistically, then all that the argument actually implies is that the student cannot be absolutely certain that he will receive a surprise exam. The contradiction in this case derives from the fact that to be absolutely certain that an event will occur is, for a self-reflective Bayesian, to be certain that one will *always* be certain that that event will occur. The teacher's announcement that there will be a surprise exam provides one example of a claim towards which one cannot coherently adopt such a dogmatic attitude.

In order to avoid the paradox, we must therefore adopt a non-dogmatic standard for acceptance relative to which the student may count as having accepted the teacher's announcement despite not being absolutely certain that he will receive a surprise exam. The minimal standard for acceptance that we have proposed in this section is that the student judge it to be more likely than not that he will receive the exam on a day on which he will judge it to be more likely than not that he will not. As we have just seen, with respect to this non-dogmatic standard of acceptance, there are many contexts in which it is possible for the student to coherently accept the teacher's announcement.

Of course, in order to constitute a satisfactory resolution of the paradox it must further be argued that the proposed standard for acceptance is a plausible one. In other words, it must be shown that this standard for acceptance can make sense of our various intuitions concerning the conditions under which the teacher's announcement can be accepted and the consequences of doing so. One such intuition that has been discussed by many commentators is that the teacher's announcement seems to become "easier" to accept the more days there are in class. In the following section, we will examine various proposals for how to make sense of this intuition in the Bayesian framework introduced above.

## 3 Accepting the Teacher's Announcement as a function of $N$

The intuition that if the class meets only once, the student cannot coherently accept the teacher's announcement can be be given a straightforward justification by appeal to the standard for coherent acceptance proposed in the previous section. If $N = 1$:

$$P(S(P,\alpha)) = \begin{cases} P(E_1) & \text{if } P(E_1) \leq 1-\alpha \\ 0 & \text{otherwise} \end{cases}.$$

So, for any $\alpha \in (1/2, 1]$:

$$P(S(P,\alpha)) \leq 1-\alpha < 1/2.$$

Hence, if $N = 1$, there is no probability function satisfying the condition $P(S(P,\alpha)) \geq \beta$, for both $\alpha, \beta \in (1/2, 1]$. Thus, in this case, the student cannot coherently expect to receive a surprise exam. [14] But what if the class meets more than once? In this case, can the student coherently accept the teacher's announcement? The answer to this question is yes, i.e., if $N \geq 2$, then there do exist probabilities satisfying the condition $P(S(P,\alpha)) \geq \beta$, for some $\alpha, \beta \in (1/2, 1]$.[15]

Despite this fact, there remains something intuitively odd about the teacher's announcement in the case where the class meets only two or three times. Various proposals have been put forward in the literature for how to make sense of this intuition. In this section, we will examine three such proposals. Sections 3.1 and 3.2 discuss proposals inspired by the analyses offered in [Kripke, 2011] and [Hall, 1999], respectively. Readers who are interested in our preferred solution may skip these discussions and proceed directly to section 3.3.

### 3.1 An appeal to what can be accepted with reasonable certainty

So as to make matters more concrete, let us focus specifically on the case in which the class meets only twice. In the two-day case, in order to rule out the possibility that the student will receive the exam on the last (second) day of class, it must be assumed that if the student were to attend the first day of class without receiving the exam, he would expect to receive the exam the following day. This, however, assumes that the student is reasonably certain that he will receive an exam at some point during the term, for otherwise, having not received the exam on the first day of class, he may instead be led to conclude that he will not receive an exam at all.

Now, as it turns out, the minimum probability that the student must assign to the event that he will receive an exam in order for him to expect the exam on the second day of class having not received it on the first, exceeds that which must be assigned to this event if the student is to expect to receive a surprise exam. While this observation helps to explain how it is possible for the student to expect to receive a surprise exam in the two-day case, it also brings to light the following curious fact: if the class meets only twice, then if the student is even reasonably certain that he will be given an exam on either one of the two days on which the class meets, he cannot coherently accept the teacher's announcement. More specifically:

[14] If $P(E_1) = 1/2$, then $P(S(P, 1/2)) = 1/2$. This means that in the one-day case, the student can judge it to be as likely as not that he will receive the exam on a day on which he will judge it to be as likely as not that he will not receive the exam. This, however, is not a possible explication of the claim that the student expects to receive a surprise exam, since, in this case, the condition $P(S(P,\alpha)) \geq \beta$ is only satisfied for $\alpha = \beta = 1/2$. Given, however, that the student could be said to expect to receive a surprise exam if the surprise and acceptance indices could both be raised by any (infinitesimal) amount, it may perhaps be more accurate to describe an acceptance of the teacher's announcement in the one-day case as "borderline" incoherent.

[15] Suppose, for example, that $N = 2$. Then, if the student's probabilities are given by $P(E_1) = 0.45$; $P(E_2) = 0.1$; and $P(E_3) = 0.45$, then $P(S(P, 0.55)) \geq 0.55$, and so the student can coherently be described as expecting to receive a surprise exam. The general claim that for all $N \geq 2$ there exists a probability function satisfying $P(S(P,\alpha)) \geq \beta$, for some $\alpha, \beta \in (1/2, 1]$ is a corollary of Theorem 5 below, the proof of which is given in Appendix A.

**Theorem 2** *If $N = 2$ and if $P(S(P, \alpha)) \geq \beta$, for some $\alpha, \beta \in (1/2, 1]$, then $P(E_1 \vee E_2) < 3/4$.*[16]

Thus, in the two-day case, if the student is to accept the teacher's announcement, he cannot claim to know with any reasonable degree of certainty (i.e., with probability $\geq 3/4$) that the teacher will give him an exam at all.

Here then is one possible explanation of what seems odd about the teacher's announcement in the two-day case: if the class meets only twice, the teacher's announcement can only be accepted provided the student refuses to accept as certain what appears to be a perfectly straightforward and intelligible consequence of what the teacher has said, namely, that he will receive an exam.

Is this explanation satisfactory? At first blush, it may seem obvious that based on the teacher's announcement, the student ought to feel certain that he will, in fact, receive an exam. After all, in most real-world settings when a teacher announces to one of his students that he will receive a surprise exam this typically implies a commitment on the teacher's part to give the student an exam at some point during the term. If the additional fact that the exam is to be a surprise complicates matters, in such ordinary contexts, it does so only with respect to the question of when, and not whether, the exam will be given.

As intuitively plausible as this may seem, it is important to note that this intuition relies on certain tacit assumptions which are not strictly implied by the original formulation of the paradox. In particular, the student must have reason to believe that the causal factors determining whether he is to receive an exam not only ensure that this will be so, but they do so *independently* of whether or not the exam turns out to be a surprise. It is, however, not difficult to imagine a scenario in which such an assumption is unwarranted. Suppose, for example, that prior to class each day, the teacher employs a specially devised brain-scanner that measures how confident the student is that he will receive an exam that day, and that the teacher has decided in advance to only administer the exam on a day on which the results of the scan reveal that the exam would, in fact, take the student by surprise. Nothing in the way in which the paradoxical scenario was originally described precludes this possibility, and yet, were the student to be informed of the teacher's methods, he could be no more certain that he would receive an exam than he could be that he would receive a surprise exam.

While such a scenario is far-fetched, it serves to emphasize an important point. The proposed explanation of what is odd about the teacher's announcement in the two-day case relies upon certain assumptions concerning the causal mechanism by means of which it is determined whether and when the exam will be given.[17] Contrast this with the above explanation of the incoherence of the teacher's announcement in the one-day case. In that case, the justification of the intuition was purely linguistic in that it relied only on facts concerning what it means for the student to coherently accept what the teacher has said.

But does the causal nature of the proposed explanation constitute grounds for rejecting it?[18] It is, of course, difficult to investigate the extent to which our pre-theoretical intuitions in fact rely

[16] *Proof.* Suppose for contradiction that $P(E_1 \vee E_2) \geq 3/4$. It follows from the fact that $P(S(P, \alpha)) > 1/2$ that $P(E_1) < 1/2$. But then $P(E_2) > P(E_3)$, which means that $\sigma_P(E_2) < 1/2$, and so, since $\alpha > 1/2$, $S(P, \alpha) = E_1$. But this means that $P(S(P, \alpha)) = P(E_1) < 1/2$. Contradiction.

[17] Following the approach adopted in [Kripke, 2011], we may formulate such auxiliary assumptions by appeal to the notion of a "school rule", where the function of such a rule is to provide the student with grounds for dogmatically accepting a given proposition independently of any additional information that he may receive. Thus, if we suppose that there is a school rule in place requiring that, in every class, an exam must be given (so that the effect of the teacher's announcement is simply to inform the student of the fact that the exam that he is certain that he will receive, will also be a surprise) then, for the reasons outlined above, the student cannot coherently accept the teacher's announcement if the class meets twice. Hence, an alternative way of formulating the above explanation of what is odd about the teacher's announcement in the two-day case is that the announcement cannot be accepted in any context in which there is (in effect) a school rule in place requiring that the student be given an exam.

[18] Not everyone thinks so. For example, [Thalos, 1997] provides a resolution to the puzzle that is explicit in its appeal to causal intuitions.

on such causal assumptions. One might, for example, ask whether the intuition that the student cannot accept the teacher's announcement in the two-day case loses its force in such pathological contexts as the brain-scanning scenario described above. It would, however, be unreasonable to place too much weight on such subtle introspections. A more modest approach would be to inquire into whether an alternative linguistic explanation of the intuition in the two-day case can be found without making any further claim to the effect that the discovery of such an explanation would thereby invalidate the account proposed above. From a strictly theoretical point of view, such a non-causal explanation seems desirable if only because of its wider scope of applicability.

In assessing the plausibility of such alternative explanations one consideration to keep in mind is the extent to which they can be generalized so as to apply to the case in which the class meets more than twice. In other words, can the proposed explanation account for the fact that there still seems to be something odd about the student accepting the teacher's announcement if the class meets three times, or four, or five? On this score, at least, it would seem that the above explanation fails, for if the class meets more than two times, then it is easy to see that the student can even be absolutely certain that he will receive an exam while still expecting the exam to be a surprise.[19] Thus, if this approach is to be used to explain why the teacher's announcement becomes easier to accept the more days there are in class, it must be assumed that the student can be reasonably certain of more than just the claim that he will receive an exam. But it is not clear what else can be accepted with reasonable certainty based only on the teacher's announcement.[20]

### 3.2 An appeal to the teacher's honesty

Let us now consider a second way of explaining what seems odd about the teacher's announcement in the two-day case. Regardless of how many times the class meets, in order for the student to accept the teacher's announcement, he cannot expect to receive the exam on the first day of class since this would be inconsistent with him expecting to receive the exam on a day on which he expects not to receive it. But, if the class meets only twice, then, after the first day of class, there is only one day of class left, and so the student can no longer coherently expect to receive a surprise exam. Thus, in the two-day case, while the student can coherently expect to receive a surprise exam, he can only do so provided he does not expect to receive the exam prior to the time at which he can no longer coherently expect to receive it.

Now, provided the student is aware of this fact, he may be led to conclude that the teacher is not acting honestly. For it may seem as if the teacher's plan is to withhold the exam until after the point at which the student can no longer coherently accept the teacher's announcement. In this case, even if it can rightly be said of the student that he has accepted the teacher's announcement

[19] This is obvious, since to assign a probability of 1 to the event $E_1 \vee E_2 \vee E_3$ has the effect of reducing the problem to the two-day case, and, as we have already seen, the student can expect to receive a surprise exam if the class meets only twice.

[20] Kripke suggests that this approach can be generalized so as to apply to the case in which $N > 2$. He writes:

> What if there are many days?... Once again, we could invoke the "rule of the school" device. We can suppose that it is long-settled school policy that the exam must be given on a day when the students do not know that it will be given, even on the day before. And, of course, we also suppose that school policy demands an exam. Given these things, the supposition that the exam will be given on day $N - 1$ will lead to a contradiction of the appropriate premises. This type of idea could be iterated to exclude successive days from the list. The rule of the school will get successively more complicated and involve iterations of knowledge about knowledge, lack of knowledge, and the preservation of the situation. ([Kripke, 2011], p. 38)

This last sentence is vague, and it is unclear how exactly Kripke intends to complicate the school rules in a natural way so as to exclude successive days of class, one at a time. The rule that he appeals to in the three-day case is simply the rule requiring that the student be given, not only an exam, but a surprise exam. But this rule, in fact, makes the teacher's announcement incoherent no matter how many days there are in class.

that he will receive a surprise exam, associated with this acceptance is an attendant belief that the teacher may, in some sense, be out to mislead him.[21]

Here then is a second potential way of explaining what is odd about the teacher's announcement in the two-day case: if the class meets only twice, the student cannot accept what the teacher has said and at the same time accept that the teacher has spoken honestly, i.e., without any intent to mislead.[22]

Before considering this proposal any further, let us first state more clearly the mathematical fact upon which it is based. Let $S_0(P)$ be the event that the exam will be given on a day $n$ for which:

$$\sigma_P(E_n) > {}^1\!/\!{}_2.$$

The condition that a probability function $P$ satisfies the inequality

$$P(S_0(P)) > {}^1\!/\!{}_2, \tag{3}$$

is equivalent to the condition that $P(S(P, \alpha)) \geq \beta$, for some $\alpha, \beta \in ({}^1\!/\!{}_2, 1]$. Hence, condition (3) expresses the minimal standard for coherent acceptance of the teacher's announcement proposed in the previous section.

For any probability $P$ and any $n$, let $P_n$ describe what the doxastic state of a student with initial probabilities $P$ would be, were he to have attended the first $n-1$ days of class without receiving the exam, i.e.:

$$P_n(E_m) = P(E_m | \neg E_1 \wedge \cdots \wedge \neg E_{n-1}).$$

Let $S_1(P)$ be the event that the exam will be given on a day $n$ for which:[23]

$$P_n(S_0(P_n)) > {}^1\!/\!{}_2$$

$S_1(P)$ corresponds to the event that the exam will be given on a day on which the student can still coherently be described as expecting to receive a surprise exam (i.e., on a day on which he will judge it to be more likely than not that he will receive an exam on a day on which he will judge it to be more likely than not that he will not). We can coherently describe the student as expecting *this* event to occur, provided he assigns to it a probability in excess of ${}^1\!/\!{}_2$, i.e.:

$$P(S_1(P)) > {}^1\!/\!{}_2. \tag{4}$$

[21] [Sober, 1998] offers an analysis of the surprise exam paradox that appeals to this type of adversarial relationship between the student and teacher.

[22] In [Hall, 1999], a similar rationale is offered in support of the claim that, in response to the teacher's announcement, the student ought to conclude that the exam is unlikely to be given on the last day of class (in Hall's example, the class meets five times, beginning on Monday and ending on Friday):

> 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 [the probability of a Friday exam] should . . . be small . . .. ([Hall, 1999], p. 689)

Hall's point, in this passage, is that the student ought to assign a low probability to receiving the exam on the last day of class, since were he to receive the exam that day, he would know that the teacher's plan was to trick him, and presumably he has no reason to believe that the teacher is anything but honest.

[23] If $P(\neg E_1 \wedge \cdots \wedge \neg E_{n-1}) = 0$, then $P_n$ is not well defined. In this case, as before, we will simply stipulate that $E_n$ is not included in $S_1(P)$. This stipulation has no effect on $P(S_1(P))$.

The explanation proposed above of what is odd about the teacher's announcement in the two-day case is based on the mathematical fact that if $N = 2$, there is no probability that satisfies both (3) and (4). On the other hand, if $N = 3$, then there do exist such probabilities.[24]

As it turns out, this result can be straightforwardly generalized to the multi-day case (see Appendix A). If the class meets only once, there is no way for the student to expect to receive a surprise exam, and hence no way for the student to coherently accept the teacher's announcement. If the class meets twice, the student can expect to receive a surprise exam, but he cannot expect to receive it prior to the time at which he can no longer coherently expect to receive it. If the class meets three times, not only can the student expect to receive a surprise exam, but he can also expect to receive the exam on a day on which he can still coherently expect to receive it. However, he cannot expect to receive the exam prior to the time at which he can no longer coherently expect *this* to be the case, that is, prior to the time at which he can no longer coherently expect to receive the exam on a day on which he can coherently expect to receive it. And so forth.

How does this help to explain the intuition that the more days the class meets, the "easier" it is to accept the teacher's announcement? In the two-day case, as we have observed, the student cannot coherently accept that the teacher is being honest in any sense of the term that would preclude him from administering the exam on the last day of class. Moreover, in the three-day case, he cannot coherently accept that he will receive the exam before the time at which he will be forced to concede that the teacher may not have spoken honestly in this sense. But perhaps there is still something dishonest about the teacher withholding the exam until after the point at which the student can no longer coherently accept that the teacher has spoken honestly (even if he still accepts that the teacher has spoken truly). If so, then even in the three-day case there is a sense in which the student is still forced to acknowledge that the teacher may have it in mind to mislead him.

Note, however, that as $N$ increases, the precise sense in which the student is forced to regard the teacher's announcement as potentially misleading grows increasingly complex and subtle, and consequently, it grows increasingly implausible that the teacher should have formed the explicit intention to mislead the student in this way. Indeed, even in the three-day case, it is already a far stretch to imagine that the teacher's explicit plan is to withhold the exam until after the time at which the student must attribute to him the plan to withold the exam until after the time at which he can no longer coherently expect to receive it! Any student who attributed such a plan to the teacher could be rightly accused of being, at best, paranoid and, at worst, a kook!

We may thus try to account for the intuition that the more days the class meets, the "easier" it is to accept the teacher's announcement along roughly the following lines: The more days the class meets, the more conspiratorial the student must be in order to interpret the expectations forced upon him by an acceptance of the teacher's announcement as resulting from an explicit plan on the teacher's part to mislead him.

Is this explanation satisfactory? On the one hand, it has the virtue of generalizing to the multi-day case. Nevertheless, it is still susceptible to the objection that it relies on assumptions that are not strictly speaking essential to the paradox. Consider again the two-day case. In this context, it was assumed that if the teacher were to administer the exam on the last day of class, he would somehow be acting dishonestly by withholding the exam until after a time at which the student could no longer coherently accept the teacher's announcement. Such an assumption, however, takes it for granted that it lies within the teacher's *power* to administer the exam on a day of his own choosing, for otherwise he could not be accused of "withholding" the exam, and so, could not be accused of engaging in any intentional act of deception.

Of course, in real-world classroom settings, the teacher is usually the individual who determines when an exam is to be given. Nevertheless, one can easily imagine scenarios in which this is not

[24] This is a direct corollary of Theorem 5, the proof of which is given in Appendix A.

the case.[25] Perhaps it is the principal of the school who informs the teacher prior to class each day whether or not he is to administer an exam, or perhaps the date of the exam is determined by means of some random process, like the toss of a weighted coin. In the latter case, especially, were the student to receive the exam on the last day of class, there would be no agent to whom he could in virtue of this fact address a charge of deceptiveness, and hence no prior presumption in favor of any particular agent's intention to honestly administer the exam that could justify the student's prior acceptance that the exam will not be given on the last day of class.

The explanation offered in this section thus has a limited scope of applicability. In particular, it applies only in those contexts in which the student is justified in assuming that the agent who determines the day on which the exam will be given (be it the teacher or someone else) can be expected to administer the exam in an honest and straightforward manner. Again, this can be contrasted with the one-day case, where the student's inability to coherently accept the teacher's announcement is simply a consequence of what it means to expect to receive a surprise exam.

### 3.3 An appeal to how strictly the teacher's announcement can be interpreted

We are thus left searching for a purely linguistic explanation of why it seems odd for the student to accept the teacher's announcement in the case where the class meets only a few times, and why this apparent oddness seems to diminish the more days there are in class. The sort of explanation we seek is completely general in that it would apply even in a context in which no additional assumptions are made as to the causal mechanism by which it is determined whether and when the exam will be given. In particular, this explanation should apply even if the teacher is nothing more than a reliable informant.

As we have seen, in the one-day case, there is no way for the student to accept that he will receive a surprise exam in even the loosest, most inclusive sense of the terms "acceptance" and "surprise". That is, even if we adopt a standard for acceptance relative to which an agent is said to have accepted any claim which he judges to be more likely than not to be true, and even if we count as a "surprise" any exam which the student judges to be less likely than not to be given on the day on which he receives it, then the student *still* cannot accept that he will receive a surprise exam if the class meets only once.

If the class meets multiple times, this is no longer so. In this case, the student can accept that he will receive a surprise exam in at least some sense of "acceptance" and "surprise". This corresponds to the fact, noted above, that for $N \geq 2$, there exist probabilities satisfying the condition that $P(S(P, \alpha)) \geq \beta$, for some $\alpha, \beta \in (1/2, 1]$. Nevertheless, if the class meets only a few times, then there are still quite significant constraints imposed on how strictly the terms "acceptance" and "surprise" can be interpreted if the student is to accept that he will receive a surprise exam. If we refer to the pair of values $(\alpha, \beta)$ as *coherent* just in case there exists a probability $P$ satisfying $P(S(P, \alpha)) \geq \beta$, then we have the following result:

**Theorem 3** *The pair of values* $(\alpha, \beta)$ *is coherent if and only if:*

$$\beta \leq 1 - \alpha^N.$$

(For a proof of this claim, see Appendix B).

For example, in the 2-day case there is no probability satisfying the condition $P(S(P, 2/3)) \geq 2/3$. Thus, if the class meets only twice, the student can coherently accept that he will receive a surprise exam, only provided he does not adopt particularly restrictive standards for "acceptance" and "surprise". Why is this odd? It is often the case that a speaker does not intend for his words to

[25] For example, [Williamson, 2000] presents a version of the paradox where the school's custodian has seen, from a distance, a date for a surprise exam marked on the calendar and has informed the students of what he has seen.

be given the strictest possible meaning. One may, for instance, describe a glass of water as full, despite the fact that it is close to full or mostly full. However, it would be odd to describe the glass as full if it were, in fact, close to empty. This is because, while there may be interpretive contexts in which a close to empty glass counts as full, they are uncommon.[26] On the other hand, every such context is also one in which a close to full glass counts as full, and besides these contexts there are many more. This is reflected in the fact that one glass is fuller than another, in any context, if the former is closer to a completely full glass (and farther from a completely empty one) than the latter.[27]

A similar observation applies to the notions of surprise and acceptance. While there may be contexts in which an agent counts as accepting a claim provided he judges it to be only slightly more likely than not to be true, there are many more contexts in which acceptance demands a higher degree of confidence. Similarly, while there may be contexts in which an event counts as surprising provided one just barely expects it not to occur, typically a higher degree of expectation is required. What is odd about the teacher's announcement in the 2-day case is that its coherent acceptance is not possible in such typical contexts of interpretation.

What is crucial to note, however, is that the interpretive constraints imposed on the student become less restrictive the more days there are in class. This is because, as $N$ increases, the space of coherent choices for $\alpha$ and $\beta$ expands to allow for increasingly strict interpretations of the teacher's announcement (see Figure 1).

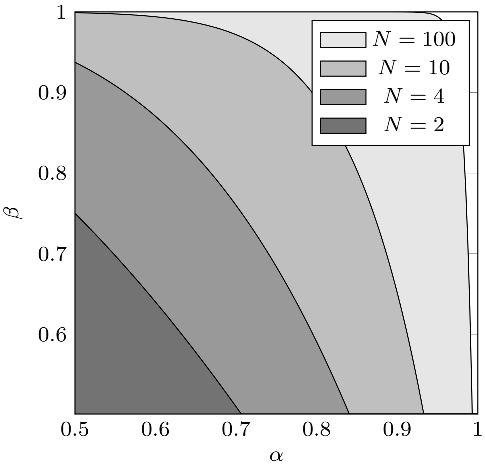

**Fig. 1** The regions indicated in the above graph correspond to the coherent values of $(\alpha, \beta)$ as a function of $N$. As $N$ increases, the space of coherent values expands to allow for increasingly strict interpretations of the teacher's announcement.

Thus, when the class meets only a few times, the student cannot help but interpret the teacher as speaking loosely. At best, he can interpret the teacher as asserting that it is somewhat likely that he will receive the exam on a day on which he will judge it to be somewhat likely that he will not. But as the number of days in class increases, he can interpret the teacher's announcement in increasingly strict terms. If the class meets many times, the student can be all but certain that

[26] Consider, for example, a laboratory setting in which test tubes are used to store very small amounts of liquid. In such a setting, when one says that all the test tubes are full, this can be understood to mean that all the test tubes contain a small amount of liquid.

[27] For a semantics of gradable adjectives that accounts for this fact about the use of such adjectives in comparative constructions, see [Kennedy, 2007]

he will receive the exam on a day on which he will judge it to be all but certain that he will not. It is this fact which, in our view, best accounts for the intuition that the teacher's announcement becomes easier to accept the more days there are in class.

## 4 Further Intuitions Concerning the Acceptance of the Teacher's Announcement

In this section, we offer explanations, in our proposed framework, for the following two intuitions concerning the acceptance of the teacher's announcement:

1. A strict interpretation of the teacher's announcement does not provide the student with any categorical information as to the date of the exam.
2. The teacher's announcement contains less information about the date of the exam the more days there are in class.

To account for the first intuition, we must first explain what we mean by a "strict" interpretation of the teacher's announcement. The relevant notion of strictness will be formalized in terms of the student's choice of the surprise and acceptance indices, $\alpha$ and $\beta$. What does it means for a particular choice of $\alpha$ and $\beta$ to correspond to a strict interpretation of the teacher's announcement?

As suggested in the previous section, the strictness of an interpretation of the teacher's announcement should increase with both $\alpha$ and $\beta$. Thus, we may think to define a strict interpretation of the teacher's announcement as one which maximizes both of these quantities. It is, however, a consequence of Theorem 3 that, under the constraint that the teacher's announcement be given a coherent interpretation, $\alpha$ and $\beta$ cannot both be maximized. This is because, when the choice of $\alpha$ and $\beta$ is restricted to the set of coherent values, the maximum value of $\alpha$ is a decreasing function in $\beta$ and vice-versa.

This complementarity of $\alpha$ and $\beta$ makes the question of what it means to give a strict interpretation of the teacher's announcement, a non-trivial one. If, on the one hand, the student attempts to increase the probability of the event that he will receive a surprise exam, he will be led to lower the standard for how unexpected an exam must be in order to count as a surprise. If, on the other hand, he opts to interpret a surprise exam as one which is highly unexpected, he will be led to lower the standard for how likely he must judge the teacher's announcement to be in order to count as having accepted it. In giving a strict interpretation to the teacher's words, the student is thus forced to compromise between strictly interpreting the teacher's claim that he will, in fact, receive a surprise exam, and strictly interpreting the claim that the exam he receives will come as a surprise.

For this reason, there is no specific choice of $\alpha$ and $\beta$ that corresponds to a strict interpretation of the teacher's announcement. Nevertheless, we can at least require that a strict interpretation satisfies the constraint that the student cannot increase both $\alpha$ and $\beta$ without thereby rendering the teacher's announcement as incoherent. In other words, a strict interpretation is one in which $\alpha$ and $\beta$ are chosen in such a way that there is no coherent pair of values $(\alpha', \beta')$ such that both $\alpha < \alpha'$ and $\beta < \beta'$. If this latter condition is satisfied, then we will say that the pair of values $(\alpha, \beta)$ is "strict". With reference to Figure 1, the only coherent pairs of values $(\alpha, \beta)$ that are strict are those which fall somewhere on the curved boundary of the shaded region.

As the following proposition indicates, the requirement that the teacher's announcement be given a strict interpretation imposes quite severe constraints on the student's probabilities:

**Theorem 4** *If the pair of values $(\alpha, \beta)$ is strict, then $P(S(P, \alpha)) \geq \beta$ if and only if:*

$$P(E_n) = \begin{cases} \alpha^{n-1}(1-\alpha) & n = 1, \ldots, N \\ \alpha^N & n = N+1 \end{cases} .$$

(For a proof of this proposition, see Appendix B).

Thus, if the student is to give a strict interpretation to the teacher's announcement, he is left with only one remaining degree of freedom with which to assign probabilities. If we take this degree of freedom to apply to the student's choice of $\alpha$, then the problem of strictly interpreting the teacher's announcement reduces to that of choosing the value of $\alpha$ from among those in the open interval $(1/2, 1/2^{1/N})$. Once this choice has been made, the student's probabilities are uniquely given by the distribution described in Theorem 4.[28] We will refer to this probability as $P_\alpha$.

The first intuition stated above can now be explained by appeal to the following fact. For every probability distribution of the form $P_\alpha$:

$$S(P_\alpha, \alpha) = E_1 \vee \cdots \vee E_N.$$

In other words, on any strict interpretation of the teacher's announcement, the surprise event is simply the event that the student will be given an exam. Why is this intuitive? Suppose that this were not so and that the student instead took the surprise event to correspond to the fact that the exam will be administered during some proper subset of the days on which the class meets. Then, based on the teacher's announcement, the student would have reason to expect that he will be given an exam on one of these specific days. But, on a strict interpretation of the teacher's words, the teacher has told him no such thing. If, in response to the teacher's announcement, the student were to ask the teacher whether, strictly speaking, he had meant to imply that the student would receive an exam during any particular subset of the days on which the class meets, the student could reasonably expect the teacher to reply in the negative. As we have seen, on any strict interpretation of the teacher's announcement, this response by the teacher is just what the student would expect. Interpreted strictly, the teacher's words do not provide the student with sufficient grounds to rule out any of the days on which the class meets as possible dates for when the exam might be given.[29]

Let us now turn to the second intuition described above. After hearing the teacher's announcement, how much more information does the student now possess about the date of the exam than what he would have possessed had the teacher merely told him that he would receive an exam at some point during the term? In other words, how much additional information does the student acquire about the date of the exam by being informed of the fact that it will be a surprise? Intuitively, the answer to this question seems to depend on how many times the class meets. In the case where the class meets only a few times, the fact that the exam will come as a surprise

[28] It may be noted that this probability distribution is a simple Bernoulli distribution, i.e., it is that which the student would adopt were he to believe that prior to each class the teacher decides whether or not to give the student the exam that day by tossing a coin with bias $\alpha$ in favor of heads, and administering the exam just in case the coin lands showing tails. This may strike the reader as odd since it may seem as if the student is somehow obtaining substantial information as to the nature of the chance mechanism by which the date of the exam is to be determined simply by interpreting the teacher's announcement strictly. However, in the Bayesian framework, it is not assumed that an agent's subjective probabilities reflect any judgment as to the objective chances with which various events will occur.

[29] In the 5-day case, Hall considers the following probability distribution:

$$P(E_n) = \begin{cases} 0.31 & n = 1, 2, 3 \\ 0.062 & n = 4 \\ 0.007 & n = 5 \\ 0.001 & n = 6 \end{cases}$$

He notes that, "... it is not difficult to argue that [this distribution] 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 [the teacher] untruthful, deceptive, or devious." ([Hall, 1999], p. 687) However, putting $\beta = 0.99$ (as Hall does), we can see that, in this case, the student not only expects (or to use Hall's language, "knows") that he will receive a surprise exam, but that he also expects ("knows") that he will receive the exam prior to the last day of class.

conveys to the student quite a lot of additional information. However, as the number of days in class increases, the information contained in this fact grows smaller and smaller. After all, if the class meets a very large number of times, even absent any announcement by the teacher, the student can safely assume that he is likely to be surprised by *any* exam that he is given.[30]

Let us assume that if the student were merely told that he will receive an exam and nothing more, he ought to adopt the uniform distribution given by:

$$P(E_n) = \begin{cases} 1/N & \text{if } n = 1, \ldots, N \\ 0 & \text{if } n = N+1 \end{cases} .$$

The question then is to what extent the student's probabilities must deviate from this uniform state of ignorance when he is further informed of the fact that the exam will be a surprise? To answer this question in numerically precise terms, we must appeal to some quantitative measure of the amount of non-uniformity that is exhibited by the student's probabilities. The standard way of measuring the non-uniformity of a probability distribution is by its information entropy.[31] Adopting this measure, it can be shown that, for any fixed value of $(\alpha, \beta)$, the most uniform (and hence least informative) probability distribution that the student can coherently adopt grows increasingly uniform as a function of $N$. Figure 2 illustrates this point for the specific case in which both $\alpha$ and $\beta$ are equal to $3/4$.

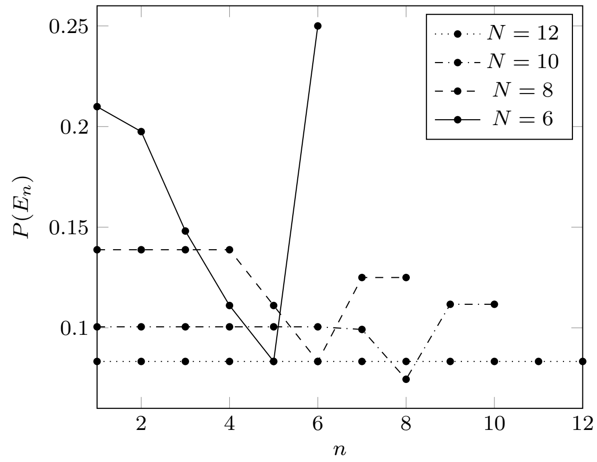

**Fig. 2** The lines in this graph indicate the probability distribution that minimizes information under the constraint $P(S(P, 3/4)) \geq 3/4$ for different values of $N$. Note the increasing uniformity of the distribution as a function of $N$.

[30] [Hall, 1999] criticizes the analysis of the paradox given in [Wright and Sudbury, 1978], claiming that: "They are at pains to insist that an adequate account of the surprise exam paradox must show how the [teacher's] announcement can be *informative*. But their account does nothing whatsoever to explain what is distinctively informative about [his] claim that the exam will *come as a surprise*." (p. 684) While Hall is quite right to insist that an account of the paradox be able to explain what is informative about the claim that the exam will come as a surprise, a satisfactory account should further explain why the informativeness of this claim decreases with increasing $N$. Hall's account cannot explain this fact.

[31] The information entropy of a probability distribution $P$ is given by:

$$H(P) = \sum_{n=1}^{N+1} P(E_n) \log P(E_n)$$

The higher the entropy of $P$, the less information it contains about when the exam will be given, with $H$ taking its maximum value of $\log(N+1)$ when $P$ is the uniform distribution.

Suppose that the student interprets the teacher's announcement at levels $\alpha = 3/4$ and $\beta = 3/4$. Then, as Figure 2 indicates, if the class meets less than twelve times, the student can only accept the teacher's announcement provided he adopts a non-uniform probability distribution. In other words, he can only accept this announcement provided he takes it to provide him with some non-trivial information about the date of the exam. If the class meets twelve or more times, however, the uniform probability distribution counts as a coherent way of accepting the teacher's announcement. Thus, in these cases, the student can accept the teacher's announcement without taking it to provide him with any information about the date of the exam.

This fact supports the intuition that if the class meets sufficiently many times, the teacher's announcement need not be taken to provide the student with any information whatsoever about the date of the exam. The exact number of times that the class must meet before such uniform ignorance counts as a way of accepting the teacher's announcement depends on the values of $\alpha$ and $\beta$. As a function of $N$, the values of $\alpha$ and $\beta$ for which the uniform distribution counts as a coherent way of accepting the teacher's announcement are illustrated in Figure 3.

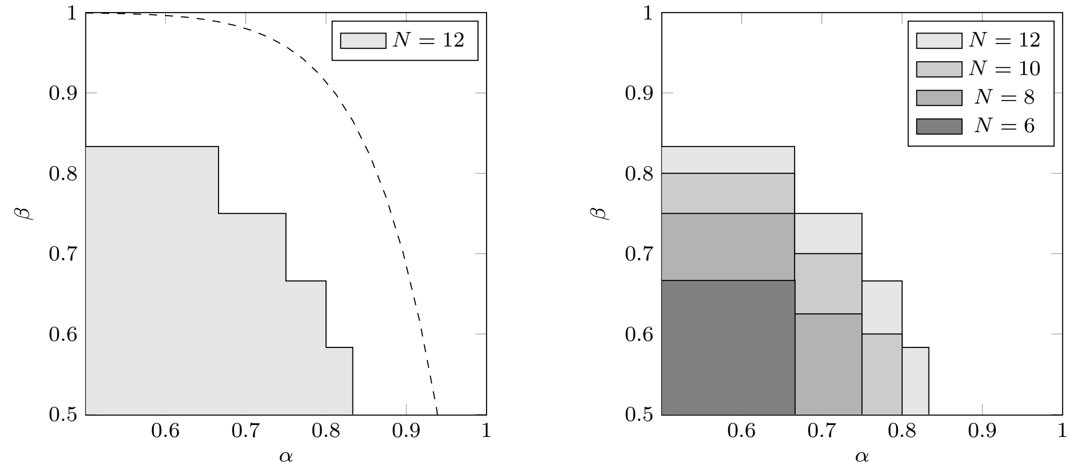

**Fig. 3** The region under the dashed line in the left graph contains all coherent pairs $(\alpha, \beta)$ for $N = 12$. The shaded region in both graphs contains all pairs $(\alpha, \beta)$ for which the uniform distribution satisfies $P(S(P, \alpha)) \geq \beta$ as a function of $N$.

## 5 The Surprise Exam, the Preface, and Fallibilism

The analysis of the surprise exam paradox offered in this paper is based on a definition of surprise that differs from many that have appeared in previous discussions of the paradox. Whereas most authors define a surprise exam to be an exam which the student does not expect to receive at the time at which it is given, we instead define a surprise exam to be one which the student expects *not* to receive at this time, i.e., one which the student finds surprising. Since many of the authors who have written on the subject readily acknowledge the ambiguity associated with the notion of a surprise exam, it is a legitimate question to ask why we have chosen to resolve the ambiguity this way. The answer is in part based on the fact that the specific version of the paradox discussed in this paper exhibits a close structural similarity to other well-known epistemological puzzles relating to the issue of how to accept the fallibility of one's own beliefs.

To see the connection between the surprise exam paradox and the general issue of accepting one's own fallibility, let us first observe that, given our particular definition of surprise, there is a natural relationship that obtains between the notions of surprise and belief. In particular, a person can be said to believe a claim to be true just to the extent that that person would be surprised to discover that what he believes is, in fact, false.[32] If belief and surprise stand in this simple relationship to one another, then the teacher's announcement can be viewed as instructing the student to believe, for each day of class, that he will not receive an exam that day, while at the same time accepting that one of these beliefs will ultimately be confounded. In other words, the teacher's announcement instructs the student to adopt a fallibilist attitude towards a certain set of his own beliefs.

The problem of how to coherently adopt such a fallibilist attitude is, of course, a general problem which appears in many guises in the epistemological literature. The most well-known of these is the preface paradox, which arises in following sort of scenario:

> Professor Lee has written a book of considerable length. Being the culmination of her life's work, she has researched each and every claim to the best of her ability, and being a careful, skilled scholar, she remarks in the preface of her book that she has only made claims that she has very good evidence to believe are true. However, in recognition of her own fallibility, she further observes that, while she has benefited greatly from the help of her colleagues, errors will inevitably be found, for which she alone is responsible.

The puzzle arises from the need to assign a coherent interpretation to Professor Lee's prefatory remark. How, on the one hand, can she claim to believe each of the claims that she makes in the book, while at the same time acknowledge that it is almost certainly the case that at least one these claims is wrong?[33]

In order to solve the preface paradox, one must explain how it is possible for the author to coherently accept her own announcement of fallibility without thereby undermining her commitment to any one of the book's claims. Analogously, the surprise exam paradox raises the question of how it is possible for the student to coherently believe that he will receive an exam at some point during the term, without thereby undermining his commitment to at least one of the claims

[32] We acknowledge that "belief" is an ambiguous term that may refer to a variety of different cognitive states, not all of which stand in this particular relation to surprise. For example, the sense in which one may be said to have a "belief" in what is, at present, our best scientific account of the world, is one which seems compatible with an expectation that these accounts will ultimately turn out to be false. Consequently, the question of how to model scientific fallibilism (when it is grounded on a pessimistic meta-induction, for example) is not the same sort of question as how to model the student's acceptance of the teacher's announcement. [Arló-Costa and Pedersen, 2012] offers a brief survey of the types of qualitative beliefs that have been discussed in the literature. They consider *full beliefs* (c.f. [Levi, 1980]), *plain beliefs* (c.f. [Spohn, 2012]), and *expectations* (c.f. [Gärdenfors, 1994]) as distinct notions of qualitative belief. Finally, it is worth pointing out one consequence of considering beliefs that have the proposed inverse relation to surprise. Since one would not and should not be surprised to learn that an arbitrary ticket in a fair lottery had won (see fn. 10), then one neither would nor should believe (in the relevant sense) that each ticket did not win. Therefore, the lottery paradox is unmotivated for this notion of belief since the lottery paradox relies upon the intuition that one is rational to believe that each ticket did not win.

[33] The standard presentations construe the paradox as resulting from two intuitions concerning rational belief. The first is the conjunction principle, which states that if a subject is rational to believe that $P$ and a rational to believe that $Q$, then that subject is rational to believe that $P$ and $Q$. The second is the no-contradiction principle, which states that one is never rational to believe an explicit contradiction, e.g., that $P$ and not-$P$. The problem then is to explain how, in light of these two intuitions, the author can both possess a rational belief in each of the claims that she makes in her book, and also a rational belief to the effect that at least one these claims is false, since the conjunction of all of these beliefs yields an explicit contradiction. The common response, defended in [Makinson, 1965], [Klein, 1985], [Kyburg, 1997], and [Foley, 1993], is to resolve the paradox by rejecting the conjunction principle. Our intent in this section is not to discuss the preface paradox as it is traditionally presented, but rather to highlight the ways in which our resolution of the surprise exam paradox can be brought to bear on this particular problem.

asserting that, for each day on which the class meets, he should expect not to receive an exam that day.

The puzzle of accepting one's own fallibility is thus akin to the puzzle of accepting that a surprising event will occur. To make this point explicit, we may note that the author in the preface paradox could just as well express her fallibility in terms of surprise.[34] She could, for instance, announce in the preface of her book, that any error she might later discover would come as a surprise to her, and yet, given her own fallibility, future inquiry will, no doubt, yield such surprising results. In other words, she could announce that she expects to receive a surprising correction. This last way of formulating the author's preface remark shows how similar her announcement truly is to that of the teacher in the surprise exam paradox. Whereas the student expects to be surprised in the future by an exam, the author expects to be surprised by the news that one of the claims in her book is false.[35]

In light of the structural isomorphism between the preface paradox and the surprise exam paradox, our resolution of the latter can be directly applied to the former.[36] For example, by adopting an analogous Bayesian modeling of the preface paradox, we could show that, as long as the book consists of more than one claim, its author can coherently accept her own fallibility. We can also explain why prefatory qualifications are easier to make the more claims there are in the book. If the book contains only a few claims, then the author's remark can only be assigned a coherent interpretation provided the reader assumes that the author is speaking somewhat loosely in either claiming to believe that each of the claims made in the book are true, or else in asserting that at least one of these claims will ultimately be proven false. If, on the other hand, the book contains many claims, then the author's remark can be assigned a coherent interpretation even on a restrictive interpretation of both the author's claim to sincerely believe in each claim in the book and her remark that, owing to her own fallibility, the book will almost certainly contain some errors.

The two intuitions concerning the acceptability of the teacher's announcement discussed in the previous section also have natural applications in the context of the preface paradox. On the one hand, any strict interpretation of the author's remark should not take this remark to imply the existence of an error in any proper part of the book, i.e., the author's fallibilism should be interpreted globally. As we have seen, this intuition can be justified by appeal to the fact that, on any strict interpretation of the author's remark, the event whose occurrence is affirmed by this remark is simply the event that the book in its entirety contains at least one error.

The second intuition concerned the amount of information about the date of the exam that is contained in the announcement that it will be a surprise. As we observed, this information decreases the more days there are in class. In the context of the preface paradox, an analogous claim can be made to the effect that the larger and more complicated the book, the less information about the location of potential errors in the book is contained in the fact that the existence of such errors is announced by one (viz., the author) who would find such errors surprising. In other words,

[34] An added benefit of viewing the author as expressing her fallibility in terms of surprise is that we avoid some hard to interpret expressions of fallibility that are used in most presentations of the preface paradox. See [Kim, 2015] for a detailed discussion.

[35] The line of reasoning employed by the student to produce a paradoxical conclusion can also be employed by the author to the same effect. We simply have to situate the story in a thought experiment in which an omniscient oracle calls our historian and tells her that she will receive a call once a day and be told, starting with the first claim and ending with the last, whether or not a particular claim is true. The oracle then announces that the author will receive news of a surprising error. It is then straightforward to see that the student's reasoning can be used, mutatis mutandis, by the author.

[36] It is worth reminding the reader that in our treatment of the surprise exam paradox, we do not assume that there is only one surprise exam during the term (see fn. 1). Instead, we simply focus our discussion on the first surprise exam. Therefore, we can apply our resolution straightforwardly to the preface paradox so long as we make it clear that we are talking about the first surprising error. In doing so, we allow for the possibility that there is more than one error in the book.

in the case of a complicated book, there is not much difference between an author admitting her own fallibility and someone else acknowledging her fallibility on her behalf. In this sense fallibilism becomes a less subjective and more intersubjective phenomenon the more complex is the system of belief towards which such fallibilism is addressed. This perhaps can help to account for the apparent banality of preface claims at least when they are addressed to long and complicated works. In such contexts, these claims provide us with virtually no information as to the author's own attitudes concerning the reliability of the various claims made in her book.

There are many apparently puzzling consequences associated with the acknowledgment of one's own fallibility. How is it that one can believe each of a number of individual claims, and at the same time acknowledge that at least one among these claims must surely turn out to be false? In order to respond to this question, a detailed account must be offered of the constraints on an agent's doxastic state that are implied by such an acknowledgment. In this paper, we have offered one such account that is grounded in the observation that an acceptance of one's own fallibility is, in many ways, similar to adopting an attitude relative to which one would be surprised not to be surprised by the occurrence of one among a number of events. Regardless of whether one accepts the details of our analysis, we hope that it is progress in itself to recognize that the fundamental question that the paradox of the surprise exam raises is really no different than that which is raised by numerous other problems appearing in the epistemological literature.[37]

## A

For a given probability $P$, let $\sigma_P^0 = \sigma_P$ and let:

$$S_0(P) = \bigvee\{E_n : \sigma_P^0(E_n) > 1/2\}.$$

Given, the event $S_{m-1}(P)$ $(m \geq 1)$, put

$$\sigma_P^m(E_n) = \begin{cases} 1 \text{ if } P(\neg E_1 \wedge \cdots \wedge \neg E_{n-1}) > 0 \text{ and } P_n(S_{m-1}(P_n)) > 1/2 \\ 0 \text{ otherwise} \end{cases}$$

and

$$S_m(P) = \bigvee\{E_n : \sigma_P^m(E_n) > 1/2\}.$$

In this appendix, we will prove the following result:

**Theorem 5** *For any $m \geq 0$, there exists a probability function $P$ satisfying:*

$$P(S_i(P)) > 1/2$$

*for all $i = 0, \ldots, m$, if and only if $N > m + 1$.*

We first establish the following lemma:

**Theorem 6** *For any probability $P$ and any $m \geq 0$, $S_{m+1}(P) \models S_m(P)$.*

[37] There are some advantages to considering the surprise version of the preface paradox and adopting our proposed solution. First, the surprise version makes no assumptions about what is expressed by the author's expression of fallibility. In this version, determining what is expressed becomes a crucial part of explaining how the author coherently accepts her own announcement of fallibility. In contrast, one problem facing many discussions of the preface paradox is that they are grounded in contentious assumptions about what is expressed by an author's expression of fallibility. Furthermore, by leaving what can be expressed by the author's prefatory remarks as an interpretative task, we are viewing the puzzle of accepting one's own fallibility not as a purely epistemic puzzle. Rather, on our view of the puzzle, part of the difficulty in accepting an announcement of one's fallibility is determining how strictly one can be speaking when expressing one's fallibility. Our solution relies upon the idea that we can express varying degrees of fallibility and as a result, there are varying degrees to which one may accept one's own fallibility.

*Proof* The proof will proceed by induction on $m$. Suppose, for contradiction, that $E_n \models S_1(P)$ but $E_n \not\models S_0(P)$. From the fact that $E_n \models S_1(P)$, it follows that $P_n(S_0(P_n)) > 1/2$. But since $E_n \not\models S_0(P)$:

$$\sigma_P(E_n) = P_n(\neg E_n) \leq 1/2.$$

Thus, $P_n(E_n) \geq 1/2$, and so, since $P_n(S_0(P_n)) > 1/2$, $E_n \models S_0(P_n)$. But since $\sigma_P(E_n) = \sigma_{P_n}(E_n)$, this contradicts the assumption that $E_n \not\models S_0(P)$. Hence, $S_1(P) \models S_0(P)$.

Now, assume that, for any probability $P$, $S_m(P) \models S_{m-1}(P)$ $(m \geq 1)$ and let $E_n \models S_{m+1}(P)$. Then, $P(\neg E_1 \wedge \cdots \wedge \neg E_{n-1}) > 0$ and $P_n(S_m(P_n)) > 1/2$. But since, by assumption, $S_m(P) \models S_{m-1}(P)$ it follows that $P_n(S_{m-1}(P_n)) > 1/2$, and so $E_n \models S_m(P)$. Hence, $S_{m+1}(P) \models S_m(P)$, which completes the induction.

We now proceed to the proof of Theorem 5. To prove the only-if direction of the claim, it suffices to show that for any probability $P$:

$$P(S_{N-1}(P)) \leq 1/2$$

The proof proceeds by induction on $N$. If $N = 1$, then

$$P(S_0(P)) \leq 1/2,$$

for any probability $P$. Now let $N > 1$, and suppose that, for all $i = 1, \ldots, N-1$, there is no probability $P$ with $P(E_1 \vee \cdots \vee E_{N-i+1}) = 1$ satisfying

$$P(S_{N-1-i}(P)) > 1/2. \tag{5}$$

It follows that

$$P_{i+1}(S_{N-1-i}(P_{i+1})) \leq 1/2, \tag{6}$$

for $i = 1, \ldots, N-1$, for otherwise, the function defined by

$$P(E_n) = P_{i+1}(E_{n+i}) \qquad (n = 1, \ldots, N-i+1)$$

would satisfy (5). But (6) implies that

$$E_{i+1} \not\models S_{N-1-i}(P),$$

for $i = 1, \ldots, N-1$, and so, from Theorem 6, we have

$$E_{i+1} \not\models S_{N-1}(P),$$

for $i = 1, \ldots, N-1$. Moreover, since $E_{N+1} \not\models S_0(P)$, by Theorem 6, $E_{N+1} \not\models S_{N-1}(P)$, and so we have $S_{N-1}(P) \models E_1$. Now, suppose that $P(S_{N-1}(P)) > 1/2$. Then $P(E_1) > 1/2$, which implies that $E_1 \not\models S_0(P)$. But then, by Theorem 6, $E_1 \models S_{N-1}(P)$, and so $S_{N-1}(P) \equiv \bot$. Contradiction. Hence, $P(S_{N-1}(P)) \leq 1/2$.

We now prove the if direction of Theorem 5. Suppose that $m + 1 < N$. We will construct a probability $P$ satisfying:

$$P(S_i(P)) > 1/2$$

for $i = 0, 1, \ldots, m$.

Let $c$ be any value in the open interval $(1/2, 1/\sqrt{2})$, and for each $n = 1, \ldots, N+1$, let

$$P(E_n) = \begin{cases} c^{n-1}(1-c) & 1 \leq n \leq N \\ c^N & n = N+1 \end{cases}$$

Then $P$ is a probability function. Moreover, for any $k$ such that $1 \leq k \leq N+1$:

$$\sigma_{P_k}(E_n) = \begin{cases} c \text{ if } k \leq n \leq N \\ 0 \text{ otherwise} \end{cases}.$$

Since $c > 1/2$, it follows that $S_0(P_k) = E_k \vee \cdots \vee E_N$, and so:

$$P_k(S_0(P_k)) = 1 - c^{N-k+1}.$$

But since $c$ is in the interval $(1/2, 1/\sqrt{2})$, it follows that

$$\sigma^1_{P_k}(E_n) = \begin{cases} 1 \text{ if } k \leq n \leq N-1 \\ 0 \text{ otherwise} \end{cases}.$$

Now suppose that

$$\sigma^i_{P_k}(E_n) = \begin{cases} 1 \text{ if } k \leq n \leq N-i \\ 0 \text{ otherwise} \end{cases} . \tag{7}$$

for some $1 < i < m$. Then $S_i(P_k) = E_k \vee \cdots \vee E_{N-i}$, and so

$$P_k(S_i(P_k)) = \begin{cases} 1 - c^{j-k+1} & \text{if } k \leq N-i \\ 0 & \text{otherwise} \end{cases}$$

But since $c$ is in the interval $(1/2, 1/\sqrt{2})$, it follows that

$$\sigma^{i+1}_{P_k}(E_n) = \begin{cases} 1 \text{ if } k \leq n \leq N-(i+1) \\ 0 \text{ otherwise} \end{cases} .$$

Hence, by induction, (7) holds for all $i = 1, \ldots, m$. This (together with the fact that $S(P_0) = E_1 \vee \cdots \vee E_N$) implies that $S_i(P) = E_1 \vee \cdots \vee E_{N-i}$, for $i = 0, 1, \ldots, m$. But then, since $m < N-1$:

$$P(S_i(P)) \geq P(E_1 \vee E_2) > 1/2,$$

for $i = 0, 1 \ldots, m$.

## B

In this appendix, we will prove Theorems 3 and 4. Both propositions follow straightforwardly from the following lemma:

**Theorem 7** *For each $\alpha \in (1/2, 1]$, let*

$$P_\alpha(E_n) = \begin{cases} \alpha^{n-1}(1-\alpha) & i = 1, \ldots, N \\ \alpha^N & i = N+1 \end{cases}$$

*Then, $P_\alpha(S(P_\alpha, \alpha)) \geq P(S(P, \alpha))$, for any probability $P$. Moreover, the inequality is strict unless $P = P_\alpha$.*

*Proof* It is easy to verify that $\sigma_{P_\alpha}(E_n) = \alpha$, for $n = 1, \ldots, N$. Hence, $S(P_\alpha, \alpha) = E_1 \vee \cdots \vee E_N$ and $P_\alpha(S(P_\alpha, \alpha)) = 1 - \alpha^N$.

Fix $\alpha \in (1/2, 1]$ and let $P$ be any probability. For $n = 1, \ldots, N+1$, put

$$q_n = \sum_{j=n}^{N+1} P(E_j).$$

Then, for $n = 1, \ldots, N$:

$$\sigma_P(E_n) = \begin{cases} q_{n+1}/q_n & q_{n+1} \neq q_n \\ 0 & \text{otherwise} \end{cases} .$$

Thus, $P(S(P, \alpha))$ is equal to the probability of the set:

$$\{E_n : n \leq N \text{ and } q_{n+1} \geq \alpha q_n\}.$$

For each $n = 1, \ldots, N$, let

$$\gamma_n = \max\left\{\alpha, \frac{q_{n+1}}{q_n}\right\}$$

(here we assume that $0/0 = 1$), and define $q^*_n$ inductively as follows:

$$q^*_1 = 1$$

$$q^*_{n+1} = \gamma_n q^*_n$$

A simple induction confirms that, for $n = 1, \dots, N+1$, both (i) $q_n^* \geq q_n$; and (ii) $q_n^* \geq \alpha^{n-1}$. Hence:

$$\begin{aligned}
P(S(P,\alpha)) &= \sum_{\substack{n \leq N \\ q_{n+1} \geq \alpha q_n}} q_n - q_{n+1} \\
&= \sum_{\substack{n \leq N \\ q_{n+1} \geq \alpha q_n}} \left(1 - \frac{q_{n+1}}{q_n}\right) q_n \\
&= \sum_{\substack{n \leq N \\ q_{n+1} \geq \alpha q_n}} (1 - \gamma_n) q_n \\
&\leq \sum_{\substack{n \leq N \\ q_{n+1} \geq \alpha q_n}} (1 - \gamma_n) q_n^* \\
&= \sum_{\substack{n \leq N \\ q_{n+1} \geq \alpha q_n}} q_n^* - q_{n+1}^* \\
&\leq \sum_{n=1}^{N} q_n^* - q_{n+1}^* = q_1^* - q_{N+1}^* \leq 1 - \alpha^N = P_\alpha(S(P_\alpha, \alpha))
\end{aligned}$$

Note that the inequality is strict unless $q_{n+1} = \alpha q_n$, for $n = 1, \dots, N$. But this is equivalent to the claim that $P = P_\alpha$.

To prove Theorem 3, we first take note of the obvious fact that (for a given $N$) if $(\alpha, \beta)$ is coherent, then $(\alpha, \beta')$ is coherent, for all $\beta'$ such that $1/2 < \beta' \leq \beta$. In addition, we note that for any $\alpha \in (1/2, 1]$, it follows from Theorem 7 that $(\alpha, 1-\alpha^N)$ is coherent, but that if $\beta > 1 - \alpha^N$, $(\alpha, \beta)$ is incoherent. These two facts together entail Theorem 3.

To prove Theorem 4, simply note that Theorem 3 implies that if $(\alpha, \beta)$ satisfies the condition stated in the proposition, then $\beta = 1 - \alpha^N$. The result then follows from Theorem 7.

## References

[Arló-Costa and Pedersen, 2012] Arló-Costa, H. and Pedersen, A. P. (2012). Belief and probability: A general theory of probability cores. *International Journal of Approximate Reasoning*, 53(3):293–315.

[Chow, 1998] Chow, T. Y. (1998). The surprise examination or unexpected hanging paradox. *American Mathematical Monthly*, 105:41–51.

[Clark, 1994a] Clark, D. (1994a). How expected is the unexpected hanging? *Mathematics Magazine*, pages 55–58.

[Clark, 1994b] Clark, R. (1994b). Pragmatic paradox and rationality. *Canadian journal of philosophy*, 24(2):229–242.

[D. Borwein and Marechal, 2000] D. Borwein, J. B. and Marechal, P. (2000). Surprise maximization. *The American Mathematical Monthly*, 107(6):517–527.

[Foley, 1993] Foley, R. (1993). *Working without a Net.* Oxford University Press.

[Gärdenfors, 1994] Gärdenfors, P. (1994). *The role of expectations in reasoning.* Springer.

[Hall, 1999] Hall, N. (1999). How to set a surprise exam. *Mind*, 108(432):647–703.

[Kennedy, 2007] Kennedy, C. (2007). Vagueness and grammar: the semantics of relative and absolute gradable adjectives. *Linguistics and Philosophy*, 30(1):1–45.

[Kim, 2015] Kim, B. (2015). This paper surely contains some errors. *Philosophical Studies*, 172(4):1013–1029.

[Klein, 1985] Klein, P. (1985). The virtues of inconsistency. *The Monist*, 68(1):105–135.

[Kripke, 2011] Kripke, S. (2011). On two paradoxes of knowledge. In *Philosophical Troubles: Collected Papers, Volume 1.* Oxford University Press.

[Kyburg, 1997] Kyburg, J. H. E. (1997). The rule of adjunction and reasonable inference. *Journal of Philosophy*, 94(3):109–125.

[Levi, 1980] Levi, I. (1980). *The Enterprise of Knowledge.* MIT Press, Cambridge, MA.

[Makinson, 1965] Makinson, D. C. (1965). The paradox of the preface. *Analysis*, 25:205–207.

[Margalit and Bar-Hilel, 1983] Margalit, A. and Bar-Hilel, M. (1983). Expecting the Unexpected. *Philosophia*, 13(3–4):263–288.

[O'Connor, 1948] O'Connor, D. J. (1948). Pragmatic paradoxes. *Mind*, 57(227):358–359.

[Quine, 1953] Quine, W. V. (1953). On a so-called paradox. *Mind*, 62(245):65–67.

[Schumacher and Westmoreland, 2008] Schumacher, B. and Westmoreland, M. (2008). Reverand Bayes takes the Unexpected Examination. *Math Horizons*, 16(1):26–27.

[Scriven, 1951] Scriven, M. (1951). Paradoxical announcements. *Mind*, 60(239):403–407.

[Sober, 1998] Sober, E. (1998). To give a surprise exam, use game theory. *Synthese*, 115(3):355–373.

[Sorensen, 1988] Sorensen, R. A. (1988). *Blindspots.* Oxford University Press.

[Spohn, 2012] Spohn, W. (2012). *The Laws of Belief: Ranking Theory and Its Philosophical Applications.* OUP Oxford.

[Thalos, 1997] Thalos, M. (1997). Conflict and co-ordination in the aftermath of oracular statements. *The Philosophical Quarterly*, 47(187):212–226.

[Williamson, 2000] Williamson, T. (2000). *Knowledge and its Limits.* Oxford University Press.

[Wright and Sudbury, 1978] Wright, C. and Sudbury, A. (1978). The paradox of the unexpected examination. In *The Philosopher's Annual, Vol. 1.* Rowman and Littlefield.