No Surprises

Cite this article as: Ian Wells, No Surprises, Erkenntnis https://doi.org/10.1007/s10670-019-00110-9

Title Page

Paper title: “No Surprises”

Author: Ian Wells

Author degree information:
PhD, Massachusetts Institute of Technology, 2017

Institution at which the paper was written:
Massachusetts Institute of Technology

Current institutional affiliation:
N/A

Author address:
1148 Venice Blvd
Venice, CA 90291

Author phone number:
7812492191

Author email:
ianwells@mit.edu

Alternative email:
wells.ian.thomas@gmail.com

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

No surprises

Ian Wells

Forthcoming in *Erkenntnis*

Abstract The surprise exam paradox is an apparently sound argument to the apparently absurd conclusion that a surprise exam cannot be given within a finite exam period. A closer look at the logic of the paradox shows the argument breaking down immediately. So why do the beginning stages of the argument appear sound in the first place? This paper presents an account of the paradox on which its allure is rooted in a common probabilistic mistake: the base rate fallacy. The account predicts that the paradoxical argument should get less and less convincing as it goes along—a prediction I take to be welcome. On a bleaker note, the account suggests that the base rate fallacy may be more widespread than previously thought.

Keywords Surprise exam paradox · Base rate fallacy · Conditionals · Conditional probability · Stalnaker's thesis

Introduction

The surprise exam paradox is an apparently sound argument to the apparently absurd conclusion that a surprise exam cannot be given within a finite exam period. The paradox is falsidical in the sense that its conclusion is not only apparently absurd but actually absurd.[1] Therefore, the argument is unsound. A solution to the paradox identifies the point at which the argument breaks down and explains why the argument breaks down at that point. A complete solution explains, in addition, why ordinarily reasonable people are tempted to follow through with the paradoxical argument in spite of its defectiveness. A complete solution accounts for two aspects of the argument: its defectiveness and its allure.

[1] Quine (1966) classifies paradoxes into those that are genuine antinomies and those that are not. Among those that are not, some are falsidical and others are veridical. The conclusion of a veridical paradox is apparently absurd but actually true.

Many accounts of the paradox focus on identifying the defect in the argument. The aim of this paper is to explain the argument's allure. To this end, the paper suggests that the argument trades on a common probabilistic mistake known as the 'base rate fallacy'.

Section 1 quickly rehearses the paradoxical argument in ordinary English. Section 2 examines the logic of the argument and gives an account of where the argument breaks down. Section 3 describes an example of the base rate fallacy and regiments the fallacy in the framework of probability theory. Section 4 applies base rate fallacious reasoning to the paradox. Section 5 draws attention to two broader implications of the overall analysis.

1 The paradox

A teacher announces to her student that there will be exactly one exam on some day within an upcoming exam period. The teacher says that the exam will be given at noon on whichever day it is given, and that it will be a surprise in the sense that the student will not know, on the day before the exam is given, that it will be given the next day. For concreteness, let us assume that the exam period spans an ordinary five-day week, and that the student receives the announcement on the Sunday before the week begins.

Upon receipt of the announcement, the student reasons as follows:

> Suppose that the teacher's announcement is true. Then the exam cannot be given Friday. For if it is, it will not be a surprise: I will know after noon on Thursday that it is coming Friday, since at that point I will know that only one day remains for it to be given. But that means that the exam cannot be given Thursday either. For if it is, it will not be a surprise: I will know after noon on Wednesday that it is coming Thursday, since at that point I will know that only two days remain, the last of which I have already eliminated. So the exam cannot be given Thursday or Friday. What about Wednesday? If it is given then, it will not be a surprise either: I will know after noon on Tuesday that it is coming Wednesday, since at that point I will know that only three days remain, the last two of which I have already eliminated...

The student continues eliminating days in this way until none are left. She concludes that the teacher's announcement is inconsistent: either the exam will not be given or else it will not be a surprise.

Though clever, the student's argument arrives at a conclusion that is too good to be true. The teacher's announcement is perfectly consistent. There is a possible world at which it is true. Consider, for example, a world at which the exam is given Monday. There, the student is surprised by the exam, and not only because she does not believe on Sunday that the exam will be given Monday. Even were she to have such a belief, that belief would not

amount to knowledge, owing to insufficient evidential support, among other shortcomings.[2]

The student's conclusion is false, so her argument must go wrong somewhere. But where?

2 The logic of the paradox

The student's argument takes the form of a *reductio* on the assumption that the teacher's announcement is true. The argument proceeds through a series of lemmas, the first being that the exam will not be given Friday, the second that the exam will not be given Thursday, and so on. Together, the lemmas contradict the assumption that there will be an exam within the week.

Some notation will be useful in what follows. Let $K_i p$ be the proposition that the student knows p after noon on day i, where p is a proposition and i is an integer in the interval $[0, 5]$, with 0 representing Sunday, 1 representing Monday, 2 representing Tuesday, and so on. Let E_i be the proposition that the exam is given on day i.

The assumption for *reductio* is that the teacher's announcement is true. The announcement has two parts: that exactly one exam will be given within the week, and that, for any day of the week, if the exam is given on that day, the student will not know on the day prior that the exam is coming the next day. That is, the announcement is equivalent to the conjunction of the following claims:

(Exam) For exactly one $i \in [1, 5]$: E_i.
(Surprise) For each $i \in [1, 5]$: $E_i \supset \neg K_{i-1} E_i$.

The student's argument requires several background assumptions. The first is that the student knows propositional logic throughout the week, in the sense that she knows, on each day, every tautology of propositional logic. The second and third assumptions are that the student's knowledge is closed under conjunction and known entailment.

(Logic) For each $i \in [0, 5]$: $K_i \top$.
(Conjunction) For each $i \in [0, 5]$: $(K_i p \wedge K_i q) \equiv K_i (p \wedge q)$.
(Entailment) For each $i \in [0, 5]$: $(K_i p \wedge K_i (p \supset q)) \supset K_i q$.

When convenient, I will use 'Competence' to collectively refer to the three assumptions above.

A fourth background assumption is that the student's memory is in good working order, throughout the week, with respect to information concerning whether the exam was or was not given on a particular day. More precisely, if

[2] The student's belief is neither safe from error in nearby worlds nor sensitive to the truth value of the proposition believed in those worlds. At any rate, the point that the announcement is consistent does not depend on any particular account of knowledge. Reflection on the ease with which real teachers fulfill real announcements (not relevantly dissimilar from the one in the paradox) suffices.

the exam is given on some day i, the student will know, after noon on day i and thereafter, that the exam was given on day i; and if the exam is not given on day i, the student will know, after noon on day i and thereafter, that the exam was not given on day i.

(Memory) For each $i, j \in [1, 5]$ such that $j \geq i$: $E_i \supset K_j E_i$, and $\neg E_i \supset K_j(\neg E_i)$.

We can now reconstruct the first step of the student's argument. The first step proceeds by *reductio* on the assumption that the exam is given Friday. By Surprise, E_5 entails $\neg K_4 E_5$. The argument then tries to derive $K_4 E_5$ for a contradiction. By Exam, E_5 entails $\neg E_1 \wedge \neg E_2 \wedge \neg E_3 \wedge \neg E_4$. By Memory and Conjunction, $K_4(\neg E_1 \wedge \neg E_2 \wedge \neg E_3 \wedge \neg E_4)$ follows. This, together with the assumption that the student will know on Thursday that there will be an exam within the week, entails $K_4 E_5$ via Competence.[3]

The crucial assumption at this first step of the student's argument is that the student will know on Thursday that there will be an exam within the week. That assumption may seem innocuous. Exam is entailed by the announcement. So, by Competence, the student knows it whenever she knows the announcement. The student presumably knows the announcement on Sunday, on the basis of the teacher's testimony.[4] Why should she not continue to know it on Thursday?

The problem is that the conjunction of Exam, Surprise, and $\neg E_1 \wedge \neg E_2 \wedge \neg E_3 \wedge \neg E_4$ entails $E_5 \wedge \neg K_4 E_5$. And the latter proposition cannot be known by the student on Thursday. For suppose otherwise: $K_4(E_5 \wedge \neg K_4 E_5)$. Conjunction entails $K_4 E_5 \wedge K_4(\neg K_4 E_5)$. Since only truths are known, $K_4 E_5 \wedge \neg K_4 E_5$ follows, for a contradiction. Hence, with Competence in place, the student cannot know on Thursday both that no exam has yet been given, and also that there will be a surprise exam within the week. At least one of the following claims is false:

(i) $K_4(\neg E_1 \wedge \neg E_2 \wedge \neg E_3 \wedge \neg E_4)$,
(ii) K_4Exam, or
(iii) K_4Surprise.

Call this observation 'Quine's bind'.[5]

[3] Proof: suppose K_4Exam. By Conjunction, we derive $K_4(\text{Exam} \wedge \neg E_1 \wedge \neg E_2 \wedge \neg E_3 \wedge \neg E_4)$. Since $(\text{Exam} \wedge \neg E_1 \wedge \neg E_2 \wedge \neg E_3 \wedge \neg E_4) \supset E_5$ is a tautology, Logic entails $K_4((\text{Exam} \wedge \neg E_1 \wedge \neg E_2 \wedge \neg E_3 \wedge \neg E_4) \supset E_5)$. The desired conclusion follows by Entailment.

[4] I say 'presumably' because nothing in what follows turns on the claim that, in a Friday-exam world, the student knows the announcement on Sunday. That claim is false if the following 'Confidence' principle is true: if one knows p, one knows that one will continue to know p. For what it is worth, my opinion is that the student may know the announcement on Sunday, even in a world in which that knowledge is later defeated. Future defeat is consistent with present knowledge. However, while Confidence is false, there is a true principle in the vicinity: see Hall (1999). Anyway, the issue is orthogonal to the solution defended in this section, which relies only on the point that if the exam is given Friday, the student cannot know the announcement on Thursday, provided that she knows on Thursday that no exam was given Monday through Thursday.

[5] The reference is to Quine (1953).

Quine's bind appears to block the first step of the student's argument. After all, the first step is sound only if (i) and (ii) are true. By Quine's bind, (i) and (ii) are true only if (iii) is false. Hence, the first step is sound only if (ii) is true but (iii) is false. That is, the first step is sound only if the student knows one half of the announcement but not the other. But in the standard presentation of the paradox, the student's total evidence is exhausted by the teacher's testimony. And the teacher's testimony does not privilege one half of the announcement over the other. As such, it cannot give the student knowledge of Exam without also giving her knowledge of Surprise; either the student knows both halves of the announcement or she knows neither. Hence, on the standard presentation of the paradox, the first step of the student's argument is unsound.

However, it is possible to modify the presentation of the paradox so as to accommodate the truth of (ii) and falsity of (iii).[6]

> *Ayer's enrichment.* The student's total evidence on Thursday supports Exam to a higher degree than Surprise.

Perhaps it is a known rule of the school that an exam must be given within the week, whether or not it is a surprise. In that case, the student could reason soundly as follows: "Come Thursday, I will no longer know that the announcement is true, given that I will then know that no exam was given Monday through Thursday. But the announcement has two parts, one for which I have evidence independent of the teacher's testimony. Since I will still know the relevant part, I will know on Thursday that the exam is coming the next day. No surprise there."[7]

While Ayer's enrichment releases the first step of the student's argument from Quine's bind, the release is only temporary. Quine's bind comes back for more at the second step.

[6] This is Ayer (1973)'s response to Quine. His story is that the student witnesses the teacher shuffle an ace of spades into a pile of four other cards, where the protocol is that the teacher will keep the cards in view of the student throughout the week, draw a card without replacement from the top of the deck each day, and give the exam on the day the ace of spades is drawn. The 'rule of the school' idea is Kripke (2011)'s.

[7] Note that the first step is sound on and *only* on Ayer's enrichment. For suppose that Ayer's enrichment is false. Then there are two possible cases: in α, the student's total evidence supports Exam and Surprise to an equal degree, as in the unenriched paradox, where the student's total evidence is exhausted by the teacher's testimony. In β, the student's total evidence supports Exam to a lesser degree than it supports Surprise. Suppose that α obtains. Then (ii) and (iii) are either both true or both false, since in α the student's evidence does not distinguish between Exam and Surprise. If (ii) and (iii) are both false, the first step is unsound. If (ii) and (iii) are both true, then by Quine's bind (i) is false and the first step is again unsound. Suppose, on the other hand, that β obtains. Then (ii) is true only if (iii) is as well, since in β the student's evidence for Surprise is even stronger than her evidence for Exam. By Quine's bind, (ii) and (iii) are true only if (i) is false. Hence, in β, (ii) is true only if (i) is false. Since the first step relies on both, it is unsound in β. In either case, then, the first step is unsound. By conditional introduction, it follows that if Ayer's enrichment is false, the first step is unsound. By contraposition, it follows that if the first step is sound, Ayer's enrichment is true: i.e. the first step is sound only on Ayer's enrichment.

The second step proceeds by *reductio* on the assumption that the exam will be given Thursday. By Surprise, E_4 entails $\neg K_3 E_4$. The argument then tries to derive $K_3 E_4$ for a contradiction. By Exam, E_4 entails $\neg E_1 \wedge \neg E_2 \wedge \neg E_3$. By Memory and Conjunction, $K_3(\neg E_1 \wedge \neg E_2 \wedge \neg E_3)$ follows. Given Ayer's enrichment, we may assume K_3Exam. If we also assume $K_3(\neg E_5)$, we can derive $K_3 E_4$ via Competence.

The crucial assumption at the second step is that the student will know on Wednesday that the exam will not be given Friday. Is that true? Not obviously. In a world at which the exam is given Thursday, the student will know on Wednesday that only two days remain for the exam to be given. But nothing the teacher told her suggests that the exam will be given on one of those days rather than the other. So how could the student *know* that the exam will not be given Friday? Well, if on Wednesday the student could carry out the first step of her argument, she could deduce and thereby come to know the conclusion of that step of the argument: namely, that the exam will not be given Friday. But deduction yields knowledge only from known premises. So, the crucial assumption at the second step of the student's argument follows only on the further assumption that the student will know on Wednesday each of the premises of the first step of the argument. Can she know each of those premises on Wednesday?

Reenter Quine's bind. The conjunction of each of the first step premises, together with $\neg E_1 \wedge \neg E_2 \wedge \neg E_3$, entails $E_4 \wedge \neg K_3 E_4$. After all, as we just witnessed, the first step premises entail $\neg E_5$. Moreover, $\neg E_1 \wedge \neg E_2 \wedge \neg E_3 \wedge \neg E_5$ and Exam entail E_4. Finally, E_4 and Surprise entail $\neg K_3 E_4$. The student cannot know $E_4 \wedge \neg K_3 E_4$ on Wednesday, for the same reason that she cannot know $E_5 \wedge \neg K_4 E_5$ on Thursday. Hence, with Competence in place, the student cannot know on Wednesday both that no exam has yet been given, and also that each of the first step premises is true. At least one of the following claims is false:

(i) $K_3(\neg E_1 \wedge \neg E_2 \wedge \neg E_3)$, or
(ii) $K_3(\text{Exam} \wedge \text{Surprise} \wedge \text{Competence} \wedge \text{Memory} \wedge K_4\text{Exam})$.

Since the second step relies on both, it is unsound.

Ayer's enrichment cannot release the second step from Quine's bind. It released the first step by surrendering knowledge of Surprise. The loss was tolerable because the first step did not require *knowledge* of Surprise, only Surprise itself. The same move does not help at the second step, since that step requires that the student know each premise of the first step, and one such premise is Surprise.

So we have an account of where the student's argument breaks down.[8] The argument breaks down either at the first or second step, depending on how

[8] The account sketched in this section draws from Kripke (2011), Jackson (1987), and Wright and Sudbury (1977). A virtue of the account is that it can explain several different versions of the paradox. Some commentators define "surprise" in terms of justified belief (J), rather than knowledge. Even though J is non-factive, the account still applies, provided that J abides by an enkratic constraint according to which, for any i, $\neg J_i(p \wedge \neg J_i p)$. Other commentators have shown that the diachronic aspect of the paradox is inessential:

the story is fleshed out. Either way, the defect in the argument is the same. The argument illicitly assumes, about a case in which the exam is given late in the week, that the student will know the announcement on the day before the exam is given.

Although I find much to like about this account, I cannot help but feel that something is missing.[9] Even after thinking through and appreciating the logical flaw in the student's argument, I still find the beginning steps of the argument compelling. Why is that?

The answer requires a brief detour into the world of conditionals and conditional probability.

3 The base rate fallacy

Suppose that one in every thousand drivers drive drunk.[10] A driver is selected at random and tested by a 95% reliable breathalyzer. Nothing else is known. Now consider the following claim:

> (DRUNK) If the breathalyzer indicates that the driver is drunk, the driver is drunk.

There is a simple and intuitive line of reasoning according to which this claim is highly probable. It goes like this: the breathalyzer is highly reliable; therefore, if it says that the driver is drunk, the driver probably is drunk.

Surprisingly, this line of reasoning is fallacious. Those who reason in this way are said to commit the base rate fallacy. The correct way to reason is rather as follows: drunk drivers are extremely rare—indeed, much rarer than mistaken indications of the breathalyzer; therefore, even if the breathalyzer indicates that the driver is drunk, it is still highly unlikely that the driver is drunk; it is more likely that the breathalyzer is simply mistaken on this occasion.

The explanation of why we misjudge the probability of DRUNK has partly to do with a thesis in the literature on indicative conditionals known as 'Stalnaker's Thesis'. According to this thesis, the degree of belief one should assign

see Sorensen (1988)'s version of the paradox involving multiple students facing forward in a line. The account still applies in the synchronic case. As in the original paradox, the student relies on an illicit assumption concerning an epistemic state distinct from her own at the time of her reasoning. The only difference is that, in the synchronic case, the assumption concerns the epistemic state of another student at the same time, rather than the student's own epistemic state at a later time. See Williamson (2000), p. 138, on this point.

[9] Let it be understood that, in what follows, I am taking on board the canonical solution to the paradox, as presented in §2. Now, this solution follows from a purely logical, non-probabilistic interpretation of the paradox. In subsequent sections, I give a particular probabilistic interpretation of the paradox, and I use it to explain the allure of the student's reasoning. But the solution to the paradox given in §2 does not depend on my particular probabilistic interpretation of the paradox or any probabilistic interpretation whatsoever. Further, let it be understood that my specific probabilistic interpretation of the paradox is not the only one available. For a different probabilistic account, see Hall (1999).

[10] The case is from the Wikipedia entry on the base rate fallacy.

to the proposition expressed by an indicative conditional sentence, $A \rightarrow C$, should match one's conditional degree of belief in C given A.[11]

> *Stalnaker's Thesis.* The degree of belief one should assign to a conditional $A \rightarrow C$ should match one's conditional degree of belief in C given A. In symbols:
>
> $$P(A \rightarrow C) = P(C \mid A),$$
>
> where P is a probability function representing one's degrees of rational belief and $P(C \mid A) = P(C \wedge A)/P(A)$ when $P(A) > 0$.

Stalnaker's Thesis accords with intuition in the vast majority of cases.[12] For example, suppose that a card is drawn at random from a regular 52-card deck. Nothing else is known. What is the probability that the card is a six, if it is a club? 1/13, we say, since the card's being a six is one of 13 possible outcomes in which the card is a club. And Stalnaker's Thesis agrees:

$$\begin{aligned} P(CLUB \rightarrow SIX) &= P(SIX \mid CLUB) \\ &= P(SIX \wedge CLUB)/P(CLUB) \\ &= (1/52)/(13/52) = 1/13. \end{aligned}$$

That Stalnaker's Thesis accords with intuition in the vast majority of cases suggests that, typically, we match our opinion of a conditional with our estimate of the corresponding conditional probability.[13] Thus, if our estimate of the conditional probability is off-base, so too will be our opinion of the conditional. That is where the base rate fallacy comes in. The base rate fallacy is a propensity to miscalculate conditional probabilities. When we misjudge the probability of DRUNK, we are doing one thing right but another thing wrong. We are rightly using our estimate of the relevant conditional probability to shape our opinion of the conditional. But we are miscalculating that

[11] In stating his thesis, Stalnaker (1970) applies degrees of belief to conditional sentences, rather than conditional propositions. Alternatively, some understand the thesis in terms of degrees of assertability of conditional sentences. For more on this latter approach, see DeRose (2010). The differences between these versions of the thesis will not matter for present purposes, although it is worth noting that the propositional version of the thesis has been called into question by the triviality results of, for example, Lewis (1976). However, those results may be avoidable on a contextualist theory of conditionals, such as the one developed in Bacon (2015).

[12] There may be some exceptions. See the purported counterexamples of McGee (2000) and Kaufmann (2004).

[13] Note that a simple alternative to Stalnaker's Thesis, according to which the probability of a conditional is equal to the probability of the associated material conditional, falters badly in the card case:

$$\begin{aligned} P(CLUB \rightarrow SIX) &= P(CLUB \supset SIX) = P(\neg CLUB \vee SIX) \\ &= P(\neg CLUB) + P(SIX) - P(\neg CLUB \wedge SIX) \\ &= 39/52 + 4/52 - 3/52 = 10/13. \end{aligned}$$

Of course, one should not be 77% confident that the card is a six if it is a club.

conditional probability. In the remainder of this section I will examine the breathalyzer case in more detail, first identifying the correct way to calculate the relevant conditional probability and then identifying the incorrect, base rate fallacious calculation.

Let D be the proposition that the driver is drunk. Let I be the proposition that the breathalyzer indicates that the driver is drunk. Let R be the proposition that the breathalyzer is "right" (or, correct) on this particular occasion:

$$R = (D \wedge I) \vee (\neg D \wedge \neg I).$$

Stalnaker's Thesis entails that the rational degree of belief to have in DRUNK is $P(D \mid I)$. Using the probability calculus and the definition of conditional probability, we can expand $P(D \mid I)$ as follows:[14]

$$P(D \mid I) = P(D \mid (I \wedge R)) \times P(R \mid I) + P(D \mid (I \wedge \neg R)) \times P(\neg R \mid I). \quad (1)$$

Note that $P(D \mid (I \wedge R)) = 1$ and $P(D \mid (I \wedge \neg R)) = 0$. After all, the breathalyzer's correctly indicating that the driver is drunk entails that the driver is drunk, and the breathalyzer's incorrectly indicating that the driver is drunk entails that the driver is not drunk. Substituting values, the right-hand side of equation 1 reduces to $P(R \mid I)$, the conditional probability that the breathalyzer is right on this occasion, conditional on its indicating that the driver is drunk.

By the definition of conditional probability, $P(R \mid I) = P(R \wedge I)/P(I)$. Since $R \wedge I$ is logically equivalent to $D \wedge I$, it follows that

$$P(R \mid I) = P(D \wedge I)/P(I).$$

We can calculate $P(D \wedge I)$ using the definition of conditional probability and the two stipulations of the case: namely, that the breathalyzer is 95% reliable, and that the base rate of drunk drivers is 1/1000.

$$\begin{aligned} P(D \wedge I) &= P(I \mid D) \times P(D) \\ &= .95 \times .001 = .00095. \end{aligned}$$

[14] A more general proof, following Edgington (2013): given any propositions X, Y and Z:

$$\begin{aligned} P(X \mid Y) &= \frac{P(X \wedge Y)}{P(Y)} = \frac{P(X \wedge Y \wedge Z) + P(X \wedge Y \wedge \neg Z)}{P(Y)} \\ &= \frac{P(X \mid (Y \wedge Z)) \times P(Y \wedge Z) + P(X \mid (Y \wedge \neg Z)) \times P(Y \wedge \neg Z)}{P(Y)} \\ &= P(X \mid (Y \wedge Z)) \times P(Z \mid Y) + P(X \mid (Y \wedge \neg Z)) \times P(\neg Z \mid Y). \end{aligned}$$

We can calculate $P(I)$ using the law of total probability and the two stipulations:

$$\begin{aligned} P(I) &= P(I \mid D) \times P(D) + P(I \mid \neg D) \times P(\neg D) \\ &= .95 \times .001 + .05 \times .999 \approx .05. \end{aligned}$$

Substituting values, $P(R \mid I) = .00095/.05 \approx .02$. Hence, $P(D \mid I) \approx .02$. In other words, even though the breathalyzer is generally reliable, the probability that the driver is drunk, given that the breathalyzer indicated as much, is quite low: around 2%.

That is how we should calculate $P(D \mid I)$. But what are we doing when we miscalculate that conditional probability? Informally speaking, we are failing to properly take into account the 'base rate' of drunk drivers among all drivers. In particular, we are neglecting the fact that the breathalyzer's indication of drunkenness is evidence that its indication is incorrect, given the overall rarity of drunk drivers.

Formally, the base rate fallacious judgment is the result of calculating an equation that looks very similar to equation 1, only it is not a consequence of the probability calculus. To keep straight the distinction between true conditional probability and the base rate fallacious method of calculating conditional probability, I will make use of a new operator ($\nmid$) to symbolize what we might call 'base rate fallacious conditional probability'. In the breathalyzer case, the base rate fallacious conditional probability of the driver being drunk, conditional on the breathalyzer indicating as much, is given by the following equation:

$$P(D \nmid I) = P(D \mid (I \wedge R)) \times P(R) + P(D \mid (I \wedge \neg R)) \times P(\neg R). \quad (2)$$

Because $P(D \mid (I \wedge R)) = 1$ and $P(D \mid (I \wedge \neg R)) = 0$, the right-hand side of equation 2 reduces to $P(R)$, the unconditional probability that the breathalyzer is right on this occasion, i.e. 95%. This value supports the base rate fallacious judgment according to which it is highly likely that the driver is drunk, given that the breathalyzer indicated as much.

Notice that the right-hand sides of equations 1 and 2 are almost identical.

$$P(D \mid I) = P(D \mid (I \wedge R)) \times P(R \mid I) + P(D \mid (I \wedge \neg R)) \times P(\neg R \mid I). \quad (1)$$

$$P(D \nmid I) = P(D \mid (I \wedge R)) \times P(R) + P(D \mid (I \wedge \neg R)) \times P(\neg R). \quad (2)$$

Both are weighted sums of the same conditional probabilities. The difference is that the weights on the addends in equation 2 are *un*conditional probabilities. This formal feature of the equation corresponds to the mistake of neglecting the evidential impact of the breathalyzer's indication on its correctness.

Before returning to the surprise exam paradox, it will be helpful to have general versions of equations 1 and 2 on the table. Given two propositions X

and Y and a partition of propositions $\mathcal{Z} = \{Z_1, ..., Z_n\}$, equations 1 and 2 generalize as follows:[15]

$$P(X \mid Y) = \sum_{i=1}^{n} P(X \mid (Y \wedge Z_i)) \times P(Z_i \mid Y). \tag{3}$$

$$P(X \nmid Y) = \sum_{i=1}^{n} P(X \mid (Y \wedge Z_i)) \times P(Z_i). \tag{4}$$

Again, equation 3 is a consequence of the probability calculus and represents the correct way to calculate conditional probability. Equation 4 represents the mistaken, base rate fallacious way to calculate conditional probability.

4 Paradox revisited

Recall the way in which the first step of the student's argument is presented in ordinary English: if the exam is given Friday, it will not be a surprise; but it will be a surprise; so it will not be given Friday.[16] The major premise of this step of the argument is:

(NOSUR) If the exam is given Friday, it will not be a surprise.

The account of the paradox given in section 2 suggests that NOSUR is false and that, on careful reflection, we ought to assign it a low probability.[17] This suggestion fits with Stalnaker's Thesis, since, as we will see momentarily, the conditional probability of the exam not being a surprise, conditional on its being given on Friday, is low—indeed, zero.

In the notation of section 2, we can rewrite NOSUR as $E_5 \rightarrow K_4E_5$. We are interested in the value of $P(K_4E_5 \mid E_5)$, where P represents the degrees of belief that we ought to have at the time at which we are presented with

[15] Kaufmann (2004) introduced versions of equations 3 and 4 while discussing purported counterexamples to Stalnaker's Thesis. Equation 3 is partition-invariant in the sense that the choice of $\mathcal{Z}$ is irrelevant to the value of $P(X \mid Y)$. But equation 4 is partition-dependent. The choice of $\mathcal{Z}$ matters very much to the value of $P(X \nmid Y)$. For example, letting $\mathcal{Z} = \{X, \neg X\}$, equation 4 entails that $P(X \nmid Y) = P(X)$. Douven (2008) presents another example that dramatizes the partition-dependence of equation 4. See Khoo (2016) for a discussion of partition selection.

[16] My account of the first step of the student's argument relies on some claims about conditionals. Because of this, one might worry that the account is parochial. For the first step of the student's argument could just as easily be presented without any conditionals, as a disjunctive syllogism, in the following way: either the exam will not be given Friday or else it will not be a surprise; it will be a surprise; so it will not be given Friday. However, it is well-known that English speakers are quick to infer from $X \vee Y$ to $\neg X \rightarrow Y$ (see, e.g., Stalnaker (1975)'s "direct argument", Jackson (1987)'s "passage principle", and Bennett (2003)'s "or-to-if inference"). This inference is so automatic that, I suspect, any temptation to accept the disjunction is, in the first instance, a temptation to accept the inferred conditional. Thanks to [removed] for helpful discussion on this point.

[17] More carefully, the account suggests that NOSUR is false in the canonical version of the paradox, sans Ayer's enrichment.

the first step of the student's argument. Recall that equations 3 and 4 are stated in terms of a partition of propositions, $\mathcal{Z}$, which we can think of as a background question (in the breathalyzer case, the question is whether the breathalyzer is correct on a particular occasion). In this case, we will use as a partition the question of whether the student will know on Thursday that an exam will be given within the week: $\{K_4\text{Exam}, \neg K_4\text{Exam}\}$.[18] We can now instantiate equation 3 as follows:

$$\begin{aligned} P(K_4E_5 \mid E_5) = {} & P(K_4E_5 \mid (E_5 \wedge K_4\text{Exam})) \times P(K_4\text{Exam} \mid E_5) \\ & + P(K_4E_5 \mid (E_5 \wedge \neg K_4\text{Exam})) \times P(\neg K_4\text{Exam} \mid E_5). \end{aligned} \tag{5}$$

To reduce the right-hand side of this equation, note, first, that

$$P(K_4E_5 \mid (E_5 \wedge K_4\text{Exam})) = 1.$$

After all, by Memory, E_5 entails $K_4(\neg E_1 \wedge \neg E_2 \wedge \neg E_3 \wedge \neg E_4)$, which, together with K_4Exam, entails K_4E_5 via Competence. Moreover,

$$P(K_4E_5 \mid (E_5 \wedge \neg K_4\text{Exam})) = 0.$$

After all, by Entailment, $\neg K_4\text{Exam}$ entails that $\neg K_4E_5 \vee \neg K_4(E_5 \supset \text{Exam})$. But $E_5 \supset \text{Exam}$ is a tautology, so, by Logic, the student knows it on Thursday. Hence, $\neg K_4E_5$.

Substituting values into equation 5, we have:

$$P(K_4E_5 \mid E_5) = P(K_4\text{Exam} \mid E_5).$$

In other words, the probability of NOSUR equals the conditional probability that the student will know on Thursday that an exam will be given within the week, conditional on the exam being given Friday.

Quine's bind, together with the setup of the unenriched paradox, ensures that the relevant conditional probability equals zero. By Memory, E_5 entails

$$K_4(\neg E_1 \wedge \neg E_2 \wedge \neg E_3 \wedge \neg E_4).$$

By Quine's bind, the latter proposition is inconsistent with the proposition that the student knows the announcement on Thursday. But if the student does not know the announcement on Thursday, then, absent Ayer's enrichment, she does not know on Thursday that there will be an exam. Hence,

$$P(K_4\text{Exam} \mid E_5) = 0.$$

[18] Why use this partition? Khoo (2016) argues that the context in which a conditional is uttered supplies a salient question which serves as the appropriate partition. The question of whether the student will know a crucial part of the announcement on Thursday strikes me as a particularly salient question in the context of the first step of the student's argument.

Hence, $P(K_4E_5 \mid E_5) = 0$.

Here is a different way to see why $P(K_4\text{Exam} \mid E_5) = 0$. By the definition of conditional probability, $P(K_4\text{Exam} \mid E_5)$ is equal to a ratio the numerator of which is $P(E_5 \wedge K_4\text{Exam})$. The argument of the previous paragraph showed that E_5 and K_4Exam are incompatible. If the exam is given Friday, the student's knowledge on Thursday of past exam-less days precludes her from also knowing on Thursday that an exam will be given within the week. So $P(E_5 \wedge K_4\text{Exam}) = 0$. Since the numerator is zero, the ratio is too.

Recall that equation 3 represented the probabilistically correct way to calculate conditional probability. So the probability that we should assign to NOSUR is zero. But as we saw in the previous section, we do not always assign probabilities to conditionals as we should. Sometimes we evaluate conditionals in the base rate fallacious way captured by equation 4. What happens if we commit the base rate fallacy in the surprise exam paradox?

Instantiating equation 4, we have:

$$\begin{aligned} P(K_4E_5 \mathbin{\text{!`}} E_5) = {} & P(K_4E_5 \mid (E_5 \wedge K_4\text{Exam})) \times P(K_4\text{Exam}) \\ & + P(K_4E_5 \mid (E_5 \wedge \neg K_4\text{Exam})) \times P(\neg K_4\text{Exam}). \end{aligned} \tag{6}$$

Substituting values from above, the right-hand side of equation 6 reduces to $P(K_4\text{Exam})$. In other words, the base rate fallacious probability of NOSUR equals the probability that the student will know on Thursday that an exam will be given within the week.

The question is: how confident should we be that the student will know on Thursday that an exam will be given within the week? Note that the question is not how confident we should be in K_4Exam *conditional on* the exam being given Friday. Rather, the question concerns our *un*conditional confidence in K_4Exam.

Here is an argument that $P(K_4\text{Exam})$ is high. It is highly likely that the exam will be given on one of the first four days of the week. Consider one such day n. If the exam is given on n, then, by Memory, the student will know after noon on Thursday that the exam was given on n. And if the student knows on Thursday that the exam was given on n, then, by Competence, she knows Exam. Generalizing, it is highly likely that the student will know Exam on Thursday.

The argument goes by way of the law of total probability. There are six possibilities: either the exam is given on one of the five days of the week, or it is not given at all. Since these possibilities are mutually exclusive and jointly exhaustive, the law of total probability instructs us to consider the conditional probability of K_4Exam on each of them, weight each conditional probability by the corresponding unconditional probability of the possibility conditioned on, and then sum the resulting values to arrive at the unconditional probability of K_4Exam. We showed above that, for any n such that $1 \leq n \leq 4$, the conditional probability of K_4Exam on E_n is one. We also showed that the conditional probability of K_4Exam on E_5 is zero. Additionally, it is clear that the conditional probability of K_4Exam on $\neg$Exam is zero, since only truths

are known. Supposing, for simplicity, that our degrees of belief are distributed evenly over the six possibilities, it follows that $P(K_4\text{Exam}) = 2/3$.[19]

$$\begin{aligned} P(K_4\text{Exam}) &= P(K_4\text{Exam} \mid E_1) \times P(E_1) \\ &+ P(K_4\text{Exam} \mid E_2) \times P(E_2) \\ &+ P(K_4\text{Exam} \mid E_3) \times P(E_3) \\ &+ P(K_4\text{Exam} \mid E_4) \times P(E_4) \\ &+ P(K_4\text{Exam} \mid E_5) \times P(E_5) \\ &+ P(K_4\text{Exam} \mid \neg\text{Exam}) \times P(\neg\text{Exam}) \\ &= 1/6 + 1/6 + 1/6 + 1/6 + 0 + 0 = 2/3. \end{aligned}$$

Since $P(K_4\text{Exam})$ is high, $P(K_4E_5 \nmid E_5)$ is too.

Result: if we commit the base rate fallacy at the first step of the student's argument, we will mistakenly assign high probability to the major premise of that step of the argument, when really we should assign it low probability.

That ordinarily rational people are tempted by base rate fallacious reasoning is well-documented.[20] This temptation, together with the foregoing result, explains why we are willing to follow through with the first step of the student's argument in spite of its defectiveness.

5 Conclusion

I have offered an account of the surprise exam paradox on which its allure is rooted in a common probabilistic fallacy. In closing, I want to explore some of the consequences of this account.

In my experience, both thinking about the paradox on my own and presenting the paradox to others, there is a feeling that the student's argument gets less and less convincing as it goes along.[21] While the first step seems unassailable, some doubt creeps in at the second, and even more at the third. By the end of the argument, many people are unwilling to acquiesce with the student's reasoning. This aspect of the paradox cries out for explanation.

On the present account, there is a simple diagnosis. The major premise of the second step of the student's argument is:

(NOSUR2) If the exam is given Thursday, it will not be a surprise.

Recall that P represents the degrees of belief that we ought to have when presented with the first step of the student's argument. By the time we are

[19] In fact, $P(K_4\text{Exam})$ should be greater than 2/3, since the possibility in which no exam is given deserves lower credence than the other possibilities. After all, the teacher promised an exam and, at least at the outset of the student's argument, we have no reason to doubt that she will follow through with that promise.

[20] For empirical evidence of the base rate fallacy, see, among others, Kahneman and Tversky (1973), Lyon and Slovic (1976), Casscells et al (1978) and Bar-Hillel (1980).

[21] This sentiment is echoed in Kripke (2011), p. 29.

presented with the second step, we have a new probability function, updated on the conclusion of the first step. Let P' be that function.

The allure of the second step of the argument hinges on the base rate fallacious conditional probability associated with NOSUR2. This probability reduces to $P'(K_3\text{Exam})$, just as the base rate fallacious conditional probability associated with NOSUR reduced to $P(K_4\text{Exam})$.[22] As before, we can calculate $P'(K_3\text{Exam})$ using the law of total probability. The difference is that, by the second step, we have ruled out the possibility of a Friday exam, so instead of six possibilities there are now only five: that the exam is given on one of the first four days of the week, or that it is not given at all.

$$\begin{aligned} P'(K_3\text{Exam}) &= P'(K_3\text{Exam} \mid E_1) \times P'(E_1) \\ &\quad + P'(K_3\text{Exam} \mid E_2) \times P'(E_2) \\ &\quad + P'(K_3\text{Exam} \mid E_3) \times P'(E_3) \\ &\quad + P'(K_3\text{Exam} \mid E_4) \times P'(E_4) \\ &\quad + P'(K_3\text{Exam} \mid \neg\text{Exam}) \times P'(\neg\text{Exam}) \\ &= 1/5 + 1/5 + 1/5 + 0 + 0 = 3/5. \end{aligned}$$

Hence, the present account of the paradox predicts a drop in confidence between the first and second step of the student's argument: from 67% confidence in the major premise of the first step to 60% confidence in the major premise of the second step.

Analogous reasoning shows that an even larger drop should occur at the third step:

$$\begin{aligned} P''(K_2\text{Exam}) &= P''(K_2\text{Exam} \mid E_1) \times P''(E_1) \\ &\quad + P''(K_2\text{Exam} \mid E_2) \times P''(E_2) \\ &\quad + P''(K_2\text{Exam} \mid E_3) \times P''(E_3) \\ &\quad + P''(K_2\text{Exam} \mid \neg\text{Exam}) \times P''(\neg\text{Exam}) \\ &= 1/4 + 1/4 + 0 + 0 = 1/2. \end{aligned}$$

And an even larger drop at the fourth step:

[22] Here I assume that we are confident that the student will be able to run the first step of the argument on Wednesday, so as to come to know that the exam will not be given Friday. This assumption is also implicit in the claim that $P(K_3\text{Exam} \mid E_4) = 0$. Indeed, throughout this section I make similar assumptions, according to which we are confident that the student will be able to run various steps of the argument on various days of the week. For reasons discussed at the end of section 2, we may not be justified in being so confident. Still, the assumptions are innocuous here, as the purpose of this section is not to identify what we are justified in believing about the paradox but rather to identify what we might be doing when we mistakenly accommodate the student's reasoning. (The base rate fallacious probabilities are fallacious, after all.)

$$\begin{aligned} P'''(K_1\text{Exam}) &= P'''(K_1\text{Exam} \mid E_1) \times P'''(E_1) \\ &+ P'''(K_1\text{Exam} \mid E_2) \times P'''(E_2) \\ &+ P'''(K_1\text{Exam} \mid \neg\text{Exam}) \times P'''(\neg\text{Exam}) \\ &= 1/3 + 0 + 0 = 1/3. \end{aligned}$$

The fact that $P'''(K_1\text{Exam}) = 1/3$ suggests that even those under the grip of the base rate fallacy should become skeptical of the student's conclusion at the fourth step. However, if they accept that conclusion nevertheless, pushing onward to the fifth and final step of the argument, something interesting happens.

The major premise of the fifth step is:

(NOSUR5) If the exam is given Monday, it will not be a surprise.

The base rate fallacious conditional probability associated with NOSUR5 reduces to $P''''(K_0\text{Exam})$, while the true conditional probability reduces to $P''''(K_0\text{Exam} \mid E_1)$. But $P''''(K_0\text{Exam}) = 0$. After all, having eliminated all but a Monday exam, the student cannot know the teacher's announcement, for the familiar reason that such knowledge would put the student squarely in the grip of Quine's bind. Of course, if $P''''(K_0\text{Exam}) = 0$ then $P''''(K_0\text{Exam} \mid E_1) = 0$ as well. Hence, both the base rate fallacious and the true condition probability of NOSUR5 equal zero.

The convergence of $P(. \mid .)$ and $P(. \upharpoonright .)$ at the final step of the argument is exactly what we should expect. For by the final step, even those initially convinced by the early stages of the argument become unwilling to acquiesce with the student's reasoning. The present account of the paradox accommodates this datum while simultaneously explaining the allure of the early stages.

The application of base rate fallacious reasoning to the surprise exam paradox sheds light not only on the paradox but on the fallacy as well. Standard treatments of the fallacy depict a purely probabilistic phenomenon, arising in cases where people are presented with statistical information and are unable to properly digest new evidence in light of that information. The canonical victims of the fallacy are clinicians, psychologists, legal experts and other professionals dealing with diagnostic tests. Our findings suggest that the scope of the fallacy is considerably wider than that. The fallacy afflicts conditional probabilities, but conditional probabilities guide our judgments of indicative conditionals. We should therefore expect to find instances of the fallacy not only in statistical settings but also in the context of deductive arguments involving indicative conditionals. The ubiquity of such arguments in ordinary discourse makes this expectation rather worrisome.

6 Objections

In giving my account of the allure of the paradox, I have appealed to a probabilistic framework. Some may worry about the very idea of translating the

paradox into this framework. After all, no probabilities are suggested in the original formulation of the paradox.[23] So why think that probabilities are relevant to the paradox in the first place?

It is true that the standard natural language presentation of the paradox given in §1 contains no probabilistic language. That is, the propositions expressed in the student's argument do not contain explicitly probabilistic concepts. Is it, therefore, inappropriate to appeal to probabilities in an attempt to elucidate the paradox? I think not. For one, the probabilistic apparatus I employ is used to model our degrees of belief in the propositions expressed by the student's reasoning. Now, there is nothing odd about having degrees of belief in propositions that do not contain probabilistic concepts. It is not as if we only have degrees belief in propositions expressed by sentences such as "It is probably going to rain." So the lack of probabilistic language in the statement of the paradox is no obstacle to modeling the degrees of belief of someone who is presented with the statement of the paradox.

Second, the natural language statement of the paradox contains conditionals, such as "If the exam is given Friday, it will not be a surprise." The appearance of conditionals in the paradox opens the door for a probabilistic treatment of the paradox, since it has been theorized that our opinions of conditionals are intimately connected to our conditional beliefs, as modeled by conditional probabilities. For this reason, I believe that it is both appropriate and important to model the paradox probabilistically, despite the superficial lack of any probabilistic language in the statement of the paradox.

A related objection stems from the fact that several of the probabilities that I have appealed to, in modeling the paradox, are equal to 1 or 0, due to relations of deductive entailment or inconsistency among the propositions. One might see these maximal probabilities and get the feeling that the appeal to probabilities is forced and unnatural.

I can see where this feeling is coming from, if we are thinking about an agent working through the equations of §4 and §5 in their head as I do in the paper. After all, it would be odd for someone presented with the paradox to work through these maximal probabilities rather than simply reasoning deductively. But that is really not how we should be thinking about the equations presented above. The conditional probability in equation (5), for example, represents a mental state of someone who is presented with the students paradoxical argument: that is, it represents this persons conditional degree of belief. Now, the right-hand side of the equation in (5) explains one way for us, as theorists, to calculate that conditional probability. But the equation is simply an artifact of our model. We should not take it so seriously as to think that the agent being modeled must herself reason through that equation. After all, many different equations are mathematically equivalent. Who is to say which one the agent reasons through to arrive at the level of confidence that she has? What matters is that the agent has a certain level of confidence, and that we

[23] Thanks to an anonymous referee for raising this challenge and the related challenge below.

offer an equation that models that level of confidence accurately. The equation itself is not offered as a psychological model of how the agent arrived at their particular level of confidence. So I think that the intuition—that probabilities are not relevant because some equal to 1 or 0—arises from a misunderstanding of what I am doing in §4 and §5.

I think that reflecting on the breathalyzer example further supports this response to the objection. After all, I take it as evident that equations (1) and (2) accurately capture the conflicting opinions over (DRUNK). Of course, some of the probabilities contained in those equations are equal to 1 or 0. But we do not conclude from this that the breathalyzer example has nothing to do with probability, or that probabilities are inappropriate to use when modeling the example. The maximal probabilities are simply part of the equation that we are using to model the agents degrees of belief. The breathalyzer example very clearly has to do with probability, yet in modeling the example we occasionally appeal to maximal probabilities. So too, in modeling the surprise exam paradox, we may occasionally appeal to maximal probabilities, but that does not mean that a probabilistic approach to the paradox is inappropriate.

References

Ayer AJ (1973) On a supposed antinomy. Mind 82(325):125–126
Bacon A (2015) Stalnaker's thesis in context. Review of Symbolic Logic 8(1):131–163
Bar-Hillel M (1980) The base-rate fallacy in probability judgments. Acta Psychologica 44:211–233
Bennett J (2003) A Philosophical Guide to Conditionals. Clarendon Press
Casscells W, Schoenberger A, Graboys T (1978) Interpretations by physicians of clinical laboratory results. The New England Journal of Medicine 299(18):999–1001
DeRose K (2010) Conditionals of deliberation. Mind 119(473):1–42
Douven I (2008) Kaufmann on the probabilities of conditionals. Journal of Philosophical Logic 37:259–266
Edgington D (2013) Estimating conditional chances and evaluating counterfactuals. Studia Logica
Hall N (1999) How to set a surprise exam. Mind 108(432):647–703
Jackson F (1987) Conditionals. Oxford: Blackwell
Kahneman D, Tversky A (1973) On psychology of prediction. Psychological Review 80:237–251
Kaufmann S (2004) Conditioning against the grain. Journal of Philosophical Logic 33:583–606
Khoo J (2016) Probabilities of conditionals in context. Linguistics and Philosophy 39(1):1–43
Kripke S (2011) Two paradoxes of knowledge. In: Kripke S (ed) Philosophical Troubles: Collected Papers Vol I, Oxford University Press

Lewis D (1976) Probabilities of conditionals and conditional probabilities. The Philosophical Review 85(3):297–315
Lyon D, Slovic P (1976) Dominance of accuracy information and neglect of base rates in probability estimation. Acta Psychologica 40:287–298
McGee V (2000) To tell the truth about conditionals. Analysis 60:107–111
Quine WV (1953) On a so-called paradox. Mind 62(245):65–67
Quine WV (1966) The ways of paradox. In: Quine WV (ed) The Ways of Paradox and Other Essays, New York: Random House, originally published in Scientific American 206, 196.
Sorensen R (1988) Blindspots. New York: Oxford University Press
Stalnaker R (1970) Probability and conditionals. Philosophy of Science 37:64–80
Stalnaker R (1975) Indicative conditionals. Philosophia 5(3):269–286
Williamson T (2000) Knowledge and its Limits. Oxford University Press
Wright C, Sudbury A (1977) The paradox of the unexpected examination. Australasian Journal of Philosophy 55:41–58