The Surprise Examination or Unexpected Hanging Paradox

Timothy Y. Chow

The American Mathematical Monthly, Vol. 105, No. 1. (Jan., 1998), pp. 41-51.

Stable URL:
http://links.jstor.org/sici?sici=0002-9890%28199801%29105%3A1%3C41%3ATSEOUH%3E2.0.CO%3B2-N

The American Mathematical Monthly is currently published by Mathematical Association of America.

The Surprise Examination or Unexpected Hanging Paradox

Timothy Y. Chow

Many mathematicians have a dismissive attitude towards paradoxes. This is unfortunate, because many paradoxes are rich in content, having connections with serious mathematical ideas as well as having pedagogical value in teaching elementary logical reasoning. An excellent example is the so-called "surprise examination paradox" (described below), which is an argument that seems at first to be too silly to deserve much attention. However, it has inspired an amazing variety of philosophical and mathematical investigations, that have in turn uncovered links to Gödel's incompleteness theorems, game theory, and several other logical paradoxes (e.g., the liar paradox and the sorites paradox). Unfortunately, most mathematicians are unaware of this because most of the literature has been published in philosophy journals.

In this article, I describe some of this work, emphasizing the ideas that are particularly interesting mathematically. I also try to dispel some of the confusion that surrounds the paradox and plagues even the published literature. However, I do not try to correct every error or explain every idea that has ever appeared in print. Readers who want more comprehensive surveys should see [**30**, chapters 7 and 8], [**20**], and [**16**].

At times I assume some knowledge of mathematical logic (such as may be found in Enderton [**10**]), but the reader who lacks this background may safely skim these sections.

1. THE PARADOX AND THE META-PARADOX. Let us begin by recalling the paradox. It has many variants, the earliest probably being Lennart Ekbom's surprise drill, and the best known to mathematicians (thanks to Quine and Gardner) being an unexpected hanging. We shall give the surprise examination version.

> *A teacher announces in class that an examination will be held on some day during the following week, and moreover that the examination will be a surprise. The students argue that a surprise exam cannot occur. For suppose the exam were on the last day of the week. Then on the previous night, the students would be able to predict that the exam would occur on the following day, and the exam would not be a surprise. So it is impossible for a surprise exam to occur on the last day. But then a surprise exam cannot occur on the penultimate day, either, for in that case the students, knowing that the last day is an impossible day for a surprise exam, would be able to predict on the night before the exam that the exam would occur on the following day. Similarly, the students argue that a surprise exam cannot occur on any other day of the week either. Confident in this conclusion, they are of course totally surprised when the exam occurs (on Wednesday, say). The announcement is vindicated after all. Where did the students' reasoning go wrong?*

The natural reaction to a paradox like this is to try to resolve it. Indeed, if you have not seen this paradox before, I encourage you to try to resolve it now before reading on. However, I do not want to discuss the resolution of the paradox right away. Instead, for reasons that should become apparent, I discuss what I call the "meta-paradox" first.

The meta-paradox consists of two seemingly incompatible facts. The first is that the surprise exam paradox seems easy to resolve. Those seeing it for the first time typically have the instinctive reaction that the flaw in the students' reasoning is obvious. Furthermore, most readers who have tried to think it through have had little difficulty resolving it to their own satisfaction.

The second (astonishing) fact is that to date nearly a hundred papers on the paradox have been published, and still no consensus on its correct resolution has been reached. The paradox has even been called a "significant problem" for philosophy [**30**, chapter 7, section VII]. How can this be? Can such a ridiculous argument really be a major unsolved mystery? If not, why does paper after paper begin by brusquely dismissing all previous work and claiming that it alone presents the long-awaited simple solution that lays the paradox to rest once and for all?

Some other paradoxes suffer from a similar meta-paradox, but the problem is especially acute in the case of the surprise examination paradox. For most other trivial-sounding paradoxes there is broad consensus on the proper resolution, whereas for the surprise exam paradox there is not even agreement on its proper formulation. Since one's view of the meta-paradox influences the way one views the paradox itself, I must try to clear up the former before discussing the latter.

In my view, most of the confusion has been caused by authors who have plunged into the process of "resolving" the paradox without first having a clear idea of what it *means* to "resolve" a paradox. The goal is poorly understood, so controversy over whether the goal has been attained is inevitable. Let me now suggest a way of thinking about the process of "resolving a paradox" that I believe dispels the meta-paradox.

In general, there are two steps involved in resolving a paradox. First, one establishes precisely *what the paradoxical argument is*. Any unclear terms are defined carefully and all assumptions and logical steps are stated clearly and explicitly, possibly in a formal language of some kind. Second, one *finds the fault in the argument*. Sometimes, simply performing step one reveals the flaw, e.g., when the paradox hinges on confusing two different meanings of the same word, so that pointing out the ambiguity suffices to dispel the confusion. In other cases, however, something more needs to be done; one must locate the bad assumptions, the bad reasoning, or (in desperate circumstances) the flaw in the structure of logic itself.

These two steps seem straightforward, but there are a few subtleties. For example, if, in the second step, the flaw is caused by bad assumptions, it may be hard to isolate a unique culprit. Sometimes what we discover is a set of mutually incompatible assumptions such that rejecting any one of them suffices to eliminate the contradiction. When this occurs, however, notice that while it may be an interesting question to decide which assumption to reject, such a decision is *not* usually needed to resolve the paradox. It is usually enough to exhibit the incompatible assumptions and state that their joint inconsistency is the source of the paradox.

The first step of resolving a paradox can also be subtle. As many investigators of the surprise exam paradox have noted, formal versions of a paradox sometimes miss the essence of the original informal version. Such a mistranslation evades the paradox instead of resolving it. Certainly, this is a real danger, and numerous authors have fallen into this trap. However, there is a simple but important point here that is often overlooked: the question of whether or not a particular formalization of a paradox "captures its essence" is to some extent a matter of opinion. Given two formalizations of the paradox, one person may think that the first captures the essence better but another may prefer the second. One cannot say who is objectively right, since there is always some vagueness in the original informal account. To be sure, one can sometimes argue that a particular formalization is inadequate by proposing a variation of the paradox that seems to retain its essence but for which the particular formalization fails. Even here, though, there is some room for differences of opinion, because one can sometimes argue that the variant paradox does not in fact retain the essence but is actually a different paradox that requires a different solution.

Thus, sometimes there exist multiple formalizations of a paradox that all capture its essence reasonably well. In such cases I believe it is misguided to speak of *the* resolution of the paradox. This point has also been made by Kirkham **[16]**.

With these ideas in mind we can easily explain the meta-paradox. A careful look at the literature confirms our suspicion that the paradox is not hard to resolve, because most authors have succeeded in finding resolutions. Most of the controversies have been false controversies. For example, there has been much debate between what I call the "epistemological school," (which formalizes the paradox using concepts such as knowledge, belief and memory) and the "logical school" (which avoids such concepts) over who has the "right" formalization. But both approaches are reasonable and neither is guilty of evasion.

Also, within the epistemological school there has been much debate over which axiom of a certain set of mutually inconsistent axioms about knowledge should be rejected. The question is an interesting one from the point of view of philosophically analyzing the concept of knowledge, but if we agree that identifying the "right" axiom to reject is not essential to resolving the paradox then this debate need not trouble us.

Having dealt with the meta-paradox, we now turn to the paradox itself and explore several different approaches.

2. THE LOGICAL SCHOOL.

We mathematicians have a firm belief that logic and mathematics are consistent. When we are confronted with a paradox, therefore, our tendency is to assume, even before analyzing the paradox, that either the paradox cannot be translated into a purely logical or mathematical argument, or that if it can be so translated, the faulty step or assumption will become immediately apparent. So a natural reaction to the surprise examination paradox (at least for a mathematician) is to take the students' argument and try to convert it into a rigorous proof in order to find the flaw. Let us now do this and see what happens.

Every proof begins with axioms. The students' argument seems to deduce a contradiction from the teacher's announcement, so it seems that the axioms in this case ought to be some formalization of the announcement. Now, part of the announcement—the claim that an examination will take place some time during the following week—is not difficult to formalize, but the part that says that the examination will be a surprise is not as clear. What is meant by "surprise"?

Whatever "surprise" means, it must at least mean that the students will not be able to deduce logically the date of the examination ahead of time, for if the students could *prove* that the date of the examination were such-and-such before the date arrived, they would surely not be at all surprised by the exam. So a first step towards formalizing the teacher's announcement might be, "There will be an examination next week and its date will not be deducible in advance."

This is not sufficient, however, because every proof begins with axioms. To say that the date of the examination will not be deducible in advance is a vague statement until the axioms from which the date cannot be deduced are specified precisely. Now, it is not completely clear which axioms are in question here; the informal word "surprise" is too vague to give us many clues. However, if our formalization is to be at all true to the original paradox, it should at least allow us to formalize the students' argument to some degree. Formalizing the teacher's announcement as

(a) There will be an examination next week and its date will not be deducible in advance from an empty set of assumptions

is certainly not satisfactory because it does not allow the students' argument even to begin. This would evade the paradox and not resolve it.

A better attempt at formalization might be something like

(b) There will be an examination next week and its date will not be deducible in advance from the assumption that the examination will occur some time during the week.

This formalization allows at least the first step of the students' argument to be carried out: given this announcement, the students can deduce that the examination will not occur on the last day of the week. However, if we try to reproduce the next step of the students' argument—the step that eliminates the penultimate day of the week—we find ourselves stuck. In order to eliminate the penultimate day, the students need to argue that their ability to deduce, from statement (b), that the examination will not occur on the last day implies that a last-day examination *will not be surprising*. But since we have restricted "surprising" to mean "not deducible from the assumption that the examination will occur sometime during the week" instead of "not deducible from *statement (b)*," the students' argument is blocked. To continue the argument we need to be able to use the nondeducibility from the announcement as an assumption, i.e., we must embed the nondeducibility from the announcement into the announcement itself.

It now becomes clear that to carry out the students' argument, one needs a formalization that is something like

(c) There will be an examination next week and its date will not be deducible in advance using *this announcement* as an axiom.

In other words, the announcement must be formulated as a *self-referential* statement!

There is a temptation to end the analysis here with a comment that the self-referential nature of statement (c) is the source of the paradox. After all, if a self-referential definition like this were to be presented in a mathematical paper, we would surely reject it instantly as illegal. Indeed, Shaw concludes his paper [25] with just such a comment.

However, we need to be careful. For it *is* possible in mathematics to formalize certain kinds of self-referential statements. Indeed, this was one of the crucial ideas in Gödel's proof of his incompleteness theorems, and it is now a standard technique in mathematical logic. It is natural to ask if this technique can be used to obtain a completely formal version of statement (c). The answer is yes; we give the construction (due to Fitch **[11]**; **[4]** and **[32]** have similar constructions) in some detail since it is rather interesting.

Let us reduce the number of days to two for simplicity (we consider one-day weeks shortly), and let Q_1 and Q_2 be statements representing the occurrence of the exam on days one and two, respectively. Then what we are seeking is a statement S such that

$$S \equiv \big(Q_1 \,\&\, ([S \Rightarrow Q_1] \text{ is unprovable})\big) \text{ or else}$$
$$\big(Q_2 \,\&\, ([S \,\&\, \sim Q_1 \Rightarrow Q_2] \text{ is unprovable})\big).$$

Given a first-order language that contains enough elementary arithmetic to handle primitive recursive functions, together with some Gödel numbering of the formulas, it is straightforward to formalize most aspects of this statement. There are primitive recursive functions "Neg," "Conj," and "Imp" encoding negation, conjunction, and implication (i.e., if q is the Gödel number of Q then Neg q is the Gödel number of the negation of Q, and so on), and a primitive recursive relation R that relates i to j if and only if i is the Gödel number of a proof of the sentence whose Gödel number is j. The only tricky part is the self-reference, and this is achieved using the usual (primitive recursive) "diagonalization" operator D: $D(m, n)$ is the Gödel number of the sentence obtained by replacing the free variable in the formula having Gödel number m by the name of the number n.

Now let q_1 and q_2 be the Gödel numbers of Q_1 and Q_2 respectively, let $\not\equiv$ denote exclusive or, and let $P[x]$ abbreviate $\exists y$: yRx. P stands for provable. Then we can formulate the following formula with the free variable x:

$$\big(Q_1 \,\&\, \sim P[D(x,x)\text{Imp}\, q_1]\big) \not\equiv \big(Q_2 \,\&\, \sim P[(D(x,x)\text{Conj}\,\text{Neg}\, q_1)\text{Imp}\, q_2]\big).$$

Let h be the Gödel number of this formula, and let S be the sentence obtained by substituting h for x, i.e.,

$$\big(Q_1 \,\&\, \sim P[D(h,h)\text{Imp}\, q_1]\big) \not\equiv \big(Q_2 \,\&\, \sim P[(D(h,h)\text{Conj}\,\text{Neg}\, q_1)\text{Imp}\, q_2]\big). \quad (\dagger)$$

Then, by definition of D, $D(h, h)$ is the Gödel number of S. The clincher is that $D(h, h)$ also appears on the right-hand side of (†) exactly where we want it to appear. Using "#" to denote "the name of the Gödel number of"we can rewrite S as

$$\big(Q_1 \,\&\, \sim P[\#(S \Rightarrow Q_1)]\big) \not\equiv \big(Q_2 \,\&\, \sim P[\#((S \,\&\, \sim Q_1) \Rightarrow Q_2)]\big).$$

This completes the formalization of statement (c). We can now imitate the students' argument to show that S is logically false, i.e., that $\sim S$ is a tautology. Using the definition of S, we can prove

$$(S \,\&\, \sim Q_1) \Rightarrow Q_2. \qquad (1)$$

Let a be the Gödel number of (1). By the nature of the relation R, the provability of (1) implies $P(a)$. But observe that $\sim P(a)$ appears in the second disjunct in the definition of S. It follows that

$$S \Rightarrow Q_1. \qquad (2)$$

The rest of the argument is now clear: if b is the Gödel number of (2), then the provability of (2) implies $P(b)$, but $\sim P(b)$ appears in the first disjunct of S. Therefore $\sim S$.

Thus, although self-reference is not illegitimate in all circumstances, it is illegitimate here because this particular self-referential statement is self-contradictory. Fitch's proof has a satisfying air of definitiveness, and seems to vindicate Shaw.

However, various authors have raised objections to this analysis. The most important is that the proof does not give any explanation for why the teacher's announcement appears to be vindicated after the fact. It appears to pin the blame on the teacher's announcement instead of on the students, and surely this cannot be correct.

A related objection rests on the observation that if the teacher had not announced the exam to the class but had simply decided in secret to give a surprise exam, then no paradox would have occurred. Therefore the trouble cannot be attributed solely to the propositional content of the teacher's announcement; the act of announcing it to the students must play a crucial role. The purely logical analysis seems to ignore this.

These objections have convinced many to reject entirely the "purely logical" approach, and to propose a different, "epistemological" approach.

Before moving on to a discussion of the epistemological school, however, I want to point out that the objections *can* be met. For example, the first objection indicates a misunderstanding of the purely logical approach. The conclusion of the logical analysis is *not* that the teacher's announcement is self-contradictory and is the source of the paradox. Rather, the conclusion is that *in order for the students to carry out their argument* that the teacher's announcement cannot be fulfilled, they must *interpret* the teacher's announcement as saying something like (c). If the teacher intended (c) when making the announcement, then it would be contradictory, and would remain so after the examination. However, a more reasonable assumption is that the teacher's announcement, whatever it means, does not mean (c), and that therefore the students misinterpret the announcement when they make their argument. The announcement appears to be vindicated afterwards, but the statement that is actually vindicated is something like "the students will be *psychologically* surprised by the exam," and such a statement does not permit the students' argument to be carried out. Similar observations are made in **[4]** and **[9]**.

As for the objection about the role of the act of making the announcement, observe that the same sequence of words can have different meanings depending on context, and that in the case of the teacher's announcement, the public utterance of the sentence changes its propositional content from "there will be a surprise exam" to something like "there will be a surprise exam in spite of the fact that I am now telling you that there will be a surprise exam." The logical analysis therefore *does* take into account the act of making the announcement, albeit implicitly, in its definition of the word "surprise." Ignoring the act of making the announcement would leave us stuck at (a).

3. THE EPISTEMOLOGICAL SCHOOL. The purely logical approach is attractive to a mathematician both because it shows exactly what problems arise from trying to convert the paradoxical argument into a mathematical proof and because it has connections to nontrivial theorems of logic. However, it has one serious disadvantage: certain aspects of the paradox—the act of announcing the exam, the belief or disbelief that the students have in the announcement, their assumption that they will remember the announcement during the course of the week, and so on—are taken into account only implicitly and not explicitly. It is therefore natural to ask if we can formalize the paradox in a way that lays bare these "epistemic" aspects.

Various epistemological formalizations have been proposed in the literature; we give just one here (taken from **[29]**) to illustrate the idea. As before, reduce the number of days to two for simplicity; let "1" denote "the exam occurs on the first day" and let "2" denote "the exam occurs on the second day." Let "Ka" denote "on the eve of the first day the students will know" and let "Kb" denote "on the eve of the second day the students will know." The announcement can then be written

$$[1 \Rightarrow \sim \mathrm{Ka}\ 1] \,\&\, [2 \Rightarrow (\sim \mathrm{Kb}\ 2 \,\&\, \mathrm{Kb} \sim 1)] \,\&\, [1 \vee 2]. \qquad (\ddagger)$$

We now introduce certain assumptions about knowledge and add them to our list of rules of inference in our logic.

KD: If one knows $A \,\&\, B$, then one knows A and one knows B. Similarly, if one knows that A implies B and one knows A, then one knows B.
KI: All logical truths are known.
KE: It is not possible to know something that is false.

We begin the argument with a lemma: $\mathrm{Kb}(\ddagger) \Rightarrow \sim 2$; remember that "$\Rightarrow$" here encompasses our new rules of logic KD, KI, and KE. Assume that $\mathrm{Kb}(\ddagger)$ is true. By KD, it follows that $\mathrm{Kb}[1 \vee 2]$. Now assume towards a contradiction that 2 is true, i.e., the exam is held on the last day. From $\mathrm{Kb}(\ddagger)$ and KE, $(\ddagger)$ follows, and 2 together with second conjunct of $(\ddagger)$ implies $\sim \mathrm{Kb}\ 2 \,\&\, \mathrm{Kb} \sim 1$; in particular, $\sim \mathrm{Kb}\ 2$. On the other hand, using KD, we deduce from $\mathrm{Kb}[1 \vee 2]$ and $\mathrm{Kb} \sim 1$ that Kb 2, a contradiction. Thus, $\mathrm{Kb}(\ddagger)$ implies ~ 2, and by KI we can infer $\mathrm{Ka}[\mathrm{Kb}(\ddagger) \Rightarrow \sim 2]$.

Now we can proceed with the crux of the argument: deducing a contradiction from the assumption $\mathrm{KaKb}(\ddagger)$. Assume $\mathrm{KaKb}(\ddagger)$. From KD and $\mathrm{Ka}[\mathrm{Kb}(\ddagger) \Rightarrow \sim 2]$ it follows that $\mathrm{Ka} \sim 2$. It is one of our logical truths (KE) that $\mathrm{Kb}(\ddagger) \Rightarrow (\ddagger)$, so from KI we conclude that $\mathrm{Ka}[\mathrm{Kb}(\ddagger) \Rightarrow (\ddagger)]$. By KD and our assumption that $\mathrm{KaKb}(\ddagger)$, this implies $\mathrm{Ka}(\ddagger)$ and in particular (by KD again) that $\mathrm{Ka}(1 \vee 2)$. Since we know that $\mathrm{Ka} \sim 2$, it follows from KD that Ka 1. But since $\mathrm{Ka}(\ddagger)$ is true, $(\ddagger)$ is true (by KE), and in particular its first disjunct $1 \Rightarrow \sim \mathrm{Ka}\ 1$ is true. Then from Ka 1 we deduce 1 (from KE) and hence $\sim \mathrm{Ka}\ 1$, a contradiction.

This shows that certain plausible assumptions about knowledge—KI, KD, and KE, together with the assumption that the students know that they will know the content of the announcement throughout the week—are inconsistent. Pointing out to the students that they are making these internally inconsistent assumptions about knowledge is enough to dissolve the paradox; we do not necessarily have to decide which assumption is the "wrong" one.

It is still interesting, however, to see if one of the assumptions appears to be a particularly promising candidate for rejection. Perhaps the most popular candidate has been the assumption that after hearing the announcement, the students "know" the content of the announcement. Those who maintain that we can never "know" things by authority or that we can never "know" things about the future (at least not with the same certainty that we can know many other things) naturally find this approach attractive. However, even those who are less skeptical have reason to reject the assumption, because the statement that the students are supposed to "know" is a statement that says something about the students' *inability* to "know" certain things. For comparison, consider the statement, "It is raining but John Doe does not know that it is raining." Clearly, John Doe cannot know the content of this statement even if the statement is true and it is uttered in his hearing by an extraordinarily reliable source. This curious phenomenon is known as a "Moore paradox" or a "blindspot," and the surprise exam paradox may be viewed as simply a more intricate version of this situation. The easiest way to see the connection is to reduce the length of the week to one day, so that the announcement becomes, "There will be an exam tomorrow but you do not know that." This approach is essentially the one offered in [3], [6], [18], [21], [22], [24], and [29].

Others have argued that the assumption KaKb(‡) is plausible only if one invokes the "temporal retention principle" (the students know that they will not forget the announcement during the week) or "Hintikka's KK principle" (if one knows something then one knows that one knows it), and that one or both of these assumptions should be discarded. I do not discuss this in detail here since I feel it is of limited mathematical interest, but I mention a brilliant variation of the paradox concocted by Sorensen [28], which suggests that rejecting these assumptions may be missing the point.

Exactly one of five students, Art, Bob, Carl, Don, and Eric, is to be given an exam. The teacher lines them up alphabetically so that each student can see the backs of the students ahead of him in alphabetical order but not the students after him. The students are shown four silver stars and one gold star. Then one star is secretly put on the back of each student. The teacher announces that the gold star is on the back of the student who must take the exam, and that that student will be surprised in the sense that he will not know he has been designated until they break formation. The students argue that this is impossible; Eric cannot be designated because if he were he would see four silver stars and would know that he was designated. The rest of the argument proceeds in the familiar way. The significance of this variation is that in our preceding formalization we can let Ka mean "Art knows" and Kb means "Bob knows" and then KaKb(‡) appears to be immediately plausible without reference to time or the KK principle. Thus, the problem remains even if those principles are rejected; see [28] and [16] for more discussion.

A very interesting variant of the epistemological approach, that of Kaplan and Montague [15], is a kind of hybrid of the logical and epistemological schools. They prove a theorem called "the Paradox of the Knower" that is reminiscent of Tarski's theorem on the indefinability of the truth predicate. Suppose we have a first-order language and we wish to introduce a knowledge predicate K. There are certain reasonable-sounding conditions that we might want to place on K:

(A) $K(\#Q) \Rightarrow Q$;
(B) (A) is known, i.e., $K(\#A)$;
(C) if Q can be proved from P and $K(\#P)$, then $K(\#Q)$.

Unfortunately, these assumptions cannot be satisfied. Using a diagonalization argument, we can construct a sentence S such that $S \equiv K(\#(\sim S))$, and then derive a contradiction by substituting this S for Q and A for P in (A), (B), and (C). Thus, no such knowledge predicate is possible.

One might think at first that (C) is the dubious assumption since certainly nobody knows all the logical consequences of what he knows, but (C) can be weakened to the assumption that the logical conclusion of a *particular* explicitly given proof is known, so the theorem is quite a strong one. The Paradox of the Knower has inspired some sophisticated work in logic; see [1], [2], or [13].

4. GAME THEORY.

Some authors have made the fascinating suggestion that the surprise exam paradox may be related to the iterated prisoner's dilemma. The prisoner's dilemma is a two-player game in which each player has the choice of either defecting or cooperating and must make the choice without communicating with the other player and without prior knowledge of the other player's choice. If one player defects and the other player cooperates, then the defector enjoys a large payoff and the cooperator suffers a large loss. If both players defect then both payoffs are zero, and if both players cooperate then they both earn a moderate payoff. It is easy to show that each player has a dominant strategy (i.e., one that is better than any other strategy regardless of the opponent's strategy): to defect.

Intuitively, defection is the best choice because the prisoner's dilemma is a "one-shot" game; there is no incentive for players to build up a cooperative relationship since they are guaranteed never to meet again. This suggests considering the iterated prisoner's dilemma, in which there are n rounds instead of just one (and the fact that there are exactly n rounds is public knowledge). The payoffs in each round are as in the usual prisoner's dilemma, and the two players are still not allowed to communicate with each other, but at each round they do know and remember the results of all previous rounds. One might think that in this case, occasional cooperation would be superior to invariable defection—the idea being that a cooperative move in an early stage, even if the opponent defects, encourages future cooperation that counterbalances earlier losses.

Consider the following "surprise examination" argument that even in the n-round prisoner's dilemma, the optimal strategy is invariable defection. The last round of an iterated prisoner's dilemma is identical to the "one-shot" prisoner's dilemma, since there is no hope of future cooperation. Hence the optimal last-round strategy is to defect. But since defection in the last round is certain, there is no incentive in the penultimate round to cooperate, for doing so cannot possibly encourage future cooperation. Thus, the optimal strategy in the penultimate round is also defection. Proceeding by induction, we conclude that perfect players always defect.

The analogy between this argument and the standard surprise examination argument is quite striking at first. Indeed, Sorensen [**30**] has argued that the two are really the same, and has substantially revised his analysis of the surprise exam as a result. There is, however, an important disanalogy. In the iterated prisoner's dilemma, the conclusion about invariable defection is *counterintuitive*, but it does not lead to an explicit *contradiction*. It is not difficult to adapt our argument to give a fully rigorous mathematical proof that in the iterated prisoner's dilemma, a Nash equilibrium is possible only if both players defect in every round; see [**12**, p. 166]. (A Nash equilibrium is a situation in which if the strategies of all but one player are held fixed, that player cannot do better by changing strategies. One reason that the "surprise exam" argument we presented is not rigorous as it stands is that the word "optimal" is imprecise. A Nash equilibrium is a precise concept that captures some—though not all—of the connotations of the word "optimal.") Therefore, I believe that the iterated prisoner's dilemma is essentially distinct from the surprise examination paradox and is not just a variant; see [**23**].

Nevertheless, one might be able to exploit the parallel between the surprise examination and the iterated prisoner's dilemma to obtain some new ideas for game theory. After all, cooperation is observed in the real world, and this suggests that the usual mathematical model of the iterated prisoner's dilemma might ignore some crucial point. For some interesting ideas in this direction, see [**17**].

Finally, I want to mention an unpublished idea of Karl Narveson that illustrates how the surprise exam paradox can inspire new mathematics. A teacher gives a quiz every week, with probability p_1 on Monday, p_2 on Tuesday, and so on. The teacher's goal is to find a probability distribution that maximizes the absolute value of the expected surprise when the quiz is announced. Here "surprise" is based on Shannon entropy, so the surprise on Monday is log p_1, the surprise on Tuesday is the log of the probability that the exam occurs on Tuesday given that it has not occurred on Monday, and so on until Friday, when the quiz becomes a certainty and its announcement no longer comes as a surprise.

Let $q_{n,m}$ be the probability that the exam occurs on the nth-to-the-last day of an m-day week given that it has not occurred on any previous days, where n ranges from zero to $m - 1$. As one may easily show, the optimal value of $q_{n,m}$ is independent of m, so we drop the second subscript.

Now set $s_0 = 0$. Narveson has shown that q_n is given by the mutual recursions

$$q_n = \exp(s_{n-1} - 1)$$
$$s_n = s_{n-1} - q_n$$

The p's may then be recovered from the q's. For a five-day week, the probabilities for each of the five days are about 0.1620, 0.1654, 0.1713, 0.1844, and 0.3169.

5. FURTHER READING.

The literature contains a wide variety of other approaches to the surprise examination paradox. Cargile [5] is the first paper in the literature to mention game theory. Clark [7] remarks that strictly mathematical analyses of the surprise exam are rare in the literature and he tries to fill this gap. Smullyan [27] weaves Gödel's theorem, brainteasers, the surprise exam, and other "epistemic" and "doxastic" paradoxes into a delightful tapestry. Some have seen connections between the surprise exam and the sorites paradox ("removing one grain of sand from a heap of sand leaves it a heap so zero grains of sand is still a heap"); see [8], [26], and [31]. A connection with the paradox of Schrödinger's cat is discussed in [14] and [19].

ACKNOWLEDGMENTS. The USENET newsgroup `rec.puzzles` has played an important role in my understanding of the surprise examination paradox. I am indebted to an unknown contributor to the `rec.puzzles` archive who compiled a list of about thirty references; reading these papers some five years ago constituted my first serious attempt to understand the paradox. I am grateful for the encouragement of Chris Cole, the maintainer of the `rec.puzzles` archive, and of Bill Taylor, whose enlightening and witty correspondence has had a lot of influence on this paper. Roy Sorensen, Dale Jacquette, and Martin Gardner have helped with my literature search. Thanks finally to Jim Propp for suggesting that I submit this paper to the MONTHLY, and to all my friends who have patiently endured my ramblings on the paradox over the years.

REFERENCES

1. C. A. Anderson, The paradox of the knower, *J. Phil.* **80** (1983), 338–355.
2. N. Asher and H. Kamp, The knower's paradox and representational theories of attitudes, in *Theoretical Aspects of Reasoning About Knowledge: Proceedings of the 1986 Conference, March 19–22, 1986, Monterey, California*, ed. J. Y. Halpern, Morgan Kaufmann, Los Altos, California, 1986, pp. 131–147.
3. R. Binkley, The surprise examination in modal logic, *J. Phil.* **65** (1968), 127–136.
4. J. Bosch, The examination paradox and formal prediction, *Logique et Analyse* **15** (1972), 505–525.
5. J. Cargile, The surprise test paradox, *J. Phil.* **64** (1967), 550–563.
6. J. M. Chapman and R. J. Butler, On Quine's 'so-called paradox,' *Mind* **74** (1965), 424–425.
7. D. Clark, How expected is the unexpected hanging? *Math. Mag.* **67** (1994), 55–58.
8. P. Dietl, The surprise examination, *Educational Theory* **23** (1973), 153–158.
9. M. Edman, The prediction paradox, *Theoria* **40** (1974), 166–175.
10. H. B. Enderton, *A Mathematical Introduction to Logic*, Academic Press, New York, 1972.
11. F. Fitch, A Goedelized formulation of the prediction paradox, *Amer. Phil. Quart.* **1** (1964), 161–164.
12. D. Fudenberg and J. Tirole, *Game Theory*, MIT Press, Cambridge, MA, 1991.
13. P. Grim, Operators in the paradox of the knower, *Synthese* **94** (1993), 409–428.

14. J. M. Holtzman, A note on Schrödinger's cat and the unexpected hanging paradox, *British J. Phil. Sci.* **39** (1988), 397–401.
15. D. Kaplan and R. Montague, A paradox regained, *Notre Dame J. Formal Logic* **1** (1960), 79–90.
16. R. L. Kirkham, On paradoxes and a surprise exam, *Philosophia* **21** (1991), 31–51.
17. R. C. Koons, Doxastic paradox and reputation effects in iterated games, *Theoretical Aspects of Reasoning About Knowledge: Proceedings of the Fourth Conference (TARK 1992), March 22–25, 1992, Monterey, California*, ed. Y. Moses, Morgan Kaufmann, San Mateo, California, 1992, pp. 60–72.
18. I. Kvart, The paradox of surprise examination, *Logique et Analyse* **21** (1978), 337–344.
19. J. G. Loeser, Three perspectives on Schrödinger's cat, *Amer. J. Physics* **52** (1984), 1089–1093; letters and replies, **53** (1985), 937 and **54** (1986), 296–297.
20. A. Margalit and M. Bar-Hillel, Expecting the unexpected, *Philosophia* **13** (1983), 263–288.
21. T. H. O'Beirne, Can the unexpected *never* happen? *New Scientist* **10** (1961), 464–465; letters and replies, 597–598.
22. D. Olin, The prediction paradox resolved, *Phil. Studies* **44** (1983), 225–233.
23. D. Olin, Predictions, intentions, and the prisoner's dilemma, *Phil. Quart.* **38** (1988), 111–116.
24. W. V. O. Quine, On a so-called paradox, *Mind* **62** (1953), 65–67.
25. R. Shaw, The paradox of the unexpected examination, *Mind* **67** (1958), 382–384.
26. J. W. Smith, The surprise examination on the paradox of the heap, *Phil. Papers* **13** (1984), 43–56.
27. R. Smullyan, *Forever Undecided: A Puzzle Guide to Gödel*, Knopf, New York, 1987, parts I–V, particularly chapter 2.
28. R. A. Sorensen, Recalcitrant versions of the prediction paradox, *Australasian J. Phil.* **69** (1982), 355–362.
29. R. A. Sorensen, Conditional blindspots and the knowledge squeeze: a solution to the prediction paradox, *Australasian J. Phil.* **62** (1984), 126–135.
30. R. A. Sorensen, *Blindspots*, Clarendon Press, Oxford, 1988.
31. T. Williamson, Inexact knowledge, *Mind* **101** (1992), 217–242.
32. P. Windt, The liar in the prediction paradox, *Amer. Phil. Quart.* **10** (1973), 65–68.

TIMOTHY CHOW received his Ph.D. from M.I.T. in 1995 under Richard Stanley and is now an NSF postdoc at the University of Michigan. He is interested in most areas of combinatorics, particularly those that overlap with algebra and number theory. His interest in the surprise examination paradox is part of a larger fascination with truth in all its forms, a fascination that explains his love of God, philosophy, and beauty in addition to mathematics. He also laughs a lot and is a huge fan of Charles Schulz's *Peanuts* comic strip.
Department of Mathematics, University of Michigan, Ann Arbor, MI 48109-1109
tchow@umich.edu