On the Surprise Examination Paradox and Its Surprising Ramifications[a]

Steven Orla Kimbrough

kimbrough@wharton.upenn.edu

May 2002

---

[a] File: surprise-exam-slides-poland.txt. Revision: 5/16/02.

The Surprise Examination Paradox

aka: Surprise Hanging Problem

“There will be a surprise exam given in one of the next 6 meetings of the class.”

Reasoning by backwards induction...

First question on the exam

1. Explain the fallacy in the reasoning that led you to believe it impossible for me to give you a surprise exam as announced.

Elementary confusion by the students: Speech alone (in this context) does not have the power to prevent an exam from being given, even a surprise exam.

The real question: Can the teacher give a surprise exam *and* speak truly in saying that there will be a surprise exam?

Will try to reconstruct and present a proper way of reasoning about this.

Begin with the one-shot problem: “There will be a surprise exam tomorrow."

Version 1 Applied to the One-Shot Surprise Exam Problem

1. $E \vee \neg E$

   There will be an exam tomorrow or not, a tautology.

2. $l > \frac{2}{3}$

   Arbitrary threshold ($\underline{\text{l}}$ine or $\underline{\text{l}}$evel); may be changed without loss of generality.

3. $P(E) \geq l \rightarrow \neg S$

   If the probability of an exam tomorrow is greater than or equal to the stipulated threshold, then no surprise.

4. $V(a) \rightarrow (E \wedge S)$

   There will be an exam tomorrow and it will be a surprise, asserted by the teacher. $a =$ the assertion by the teacher that there will be an exam on the next class (E) and that it will be a surprise (S). If the assertion is truthful or veridical, $V(a)$, then $(E \wedge S)$.

5. $E \rightarrow P(E) = 1$

   If there is an exam tomorrow, then the probability of that there is an exam tomorrow is 1.

6. $P(E) = 1 \rightarrow P(E) \geq l$

   A simple mathematical truth.

$\vdash$ $(\neg S \wedge \neg V(a)) \vee \neg V(a)$

Comment: Valid, but unsound. Premise (5) is the problem.

Version 2 of the Students' Reasoning Applied to the One-Shot Surprise Exam Problem

1. $E \vee \neg E$
2. $l > \frac{2}{3}$
3. $P(E) \geq l \rightarrow \neg S$
4. $V(a) \rightarrow \Box(E \wedge S)$
5. $\Box E \rightarrow P(E) = 1$
6. $P(E) = 1 \rightarrow P(E) \geq l$

$\vdash \neg V(a)$

Comment: Valid, but now premise (4) is problematic.

Version 3 of the Students' Reasoning Applied to the One-Shot Surprise Exam Problem

---

1. $E \vee \neg E$
2. $l > \frac{2}{3}$
3. $P(E) \geq t \rightarrow \neg S$
4. $V(a) \rightarrow (P(E) = 1 \wedge S)$
5. $P(E) = 1 \rightarrow E$
6. $P(E) = 1 \rightarrow P(E) \geq l$

$\vdash \neg V(a)$

Comment: Now premise (4) is problematic. If this *is* what our teacher meant, then we'll simply get another teacher, who will mean something else. The real question is whether when we find such a teacher she can speak truly and give the surprise exam.

Version 4 of the Students' Reasoning Applied to the One-Shot Surprise Exam Problem

1. $E \vee \neg E$
2. $t > \frac{2}{3}$
3. $P(E) \geq t \rightarrow \neg S$
4. $V(a) \rightarrow (E \wedge S)$
5. $P(E) = 1$
6. $P(E) = 1 \rightarrow P(E) \geq t$

$\vdash \neg V(a)$

Again, the argument is valid and the key premise is (5). Its justification is that there's no where else to put the probability mass. There will be an exam and there is only one day available for it, so all the probability has to be on that day. But this is wrong. $P(E)$ can in principle be anything at all.

Version 5 of the Students' Reasoning Applied to the One-Shot Surprise Exam Problem

1. $E \vee \neg E$
2. $l > \frac{2}{3}$
3. $(P(E) < l \wedge E) \rightarrow S$
4. $V(a) \leftrightarrow (E \wedge S)$
5. $P(E) < l$

$\vdash$ $(E \wedge S \wedge V(a)) \vee (\neg E \wedge \neg V(a))$

Bingo!

Notice that assumption (4) has been strengthened to a biconditional. This is harmless and could have been done for the earlier versions. The strengthening amounts to accepting a rule that credits the teacher with speaking truthfully, $V(a)$, if what she said—that $(E \wedge S)$—is in fact true.

The upshot

In the one-shot surprise exam problem, the teacher must either speak falsely (e.g., by making a self-contradictory statement) or speak truly but with a probability no larger than l. Only by putting herself at risk of falsehood is it possible for her to speak truly in this case. By taking a risk (of speaking falsely) the teacher expands her scope of action.

The teacher can trade risk (of falsehood) for reward (being able to give a surprise exam).

Now the 6-shot (n-shot) Surprise Exam Problem

If the teacher is willing to run a risk, $r > 1 - l$ of speaking falsely, then it is not true that surprise could not lurk on the last day. This suffices to block the backwards induction and to undue the students' reasoning in the n-shot case.

But…

The soundness of version 5 relies on the teacher being willing to accept a risk of at least $1 - l$ of speaking falsely. Given this, many would choose not to utter the one-shot surprise exam assertion. Honesty, integrity, prudence, or whatever may well prevent a reasonable person from saying something they know to have a chance higher then $\frac{1}{3}$ of being false. Better to keep silent.

What if the students know this?

Students' Reasoning Applied to the Augmented One-Shot Surprise Exam Problem

1. $E \vee \neg E$
2. $l > \frac{2}{3}$
3. $(P(E) \geq l \wedge E) \rightarrow \neg S$
4. $V(a) \leftrightarrow (E \wedge S)$
5. $P(E) \geq (1 - r)$
6. $r < (1 - l)$.
7. $(P(E) \geq (1 - r) \wedge r < (1 - l)) \rightarrow P(E) \geq l$

$\vdash$ $(\neg V(a) \wedge E \wedge \neg S) \vee (\neg E \wedge \neg V(a))$

Teacher's falsehood, validly deduced.

A deeper lesson lurks

More shots attenuate the teacher's verisimilitude scruples. Sufficient is:

1. In each period, initially the probability of the exam is less than l, and
2. The total probability of having the exam is greater than or equal to $(1 - r)$

This is trivially arranged for any l, and for any $(1 - r) < 1$, provided enough periods are available. Simply decide to give the exam with equal probability to every period. Again concretely with the l and $(1 - r)$ values given above, give a probability of $\frac{1}{5}$ of holding the exam on each of 5 days. The exam is held. If the exam is held on the fifth day, that morning the students will know that the exam will be held that day (assuming they are certain the exam will be held at all). There is only a 1 in 5 chance this will happen, which is above $(1 - r)$ and acceptable to the teacher. If the exam occurs on day 4, the students have a 50% certainty that morning, which is below l. And earlier is even better for the teacher's veracity. Further, with enough periods the teacher can set the probability of giving the exam to 1 and still have a probability $< r$ of not surprising the students, as we have just seen in the example.

In sum on the Surprise Exam Paradox

The n-period case generalizes the 1-shot case. The teacher can speak truly in this form provided the teacher is willing to undertake some risk, $r > 0$, of speaking falsely and providing n is large enough (given r and l). The teacher cannot be certain of speaking truly, but in this respect the case is like most. Usually, when we assert we take some chance of speaking falsely, even with the best of intentions. What is odd is to interpret a speaker otherwise. The only way the teacher could have spoken truthfully and given the surprise exam was to have spoken with some chance of speaking falsely. The students erred in interpreting the teacher uncharitably.

Question 2 on the Exam

2. In a 100-shot Repeated Prisoner's Dilemma game, played between the teacher and an unknown, but fully competent human subject, the teacher announces that she will gain the reward from mutual coöperation at least 2 times, net. That is, if P is the penalty for mutual defection and R is the reward for mutual coöperation, the teacher is asserting that she will get at least $98 \cdot P + 2 \cdot R$ points from the 100 trials. Can this assertion be plausibly justified? Why or why not?

The one-shot Prisoner's Dilemma game

The (one-shot) Prisoner's Dilemma game involves two players each with two strategies: C (coöperate) and D (defect). In strategic form the game is:

	C	D
C	R R	T S
D	S T	P P

with the requirement that $T > R > P > S$ and that $2 \cdot R > T$. Typically, even usually, in experiments $T = 5, R = 3, P = 1$, and $S = 0$. Since $T > R$ and $P > S$, there is only one equilibrium point (EP): both players play D. The dilemma, of course, is that if both players could play C, both would be better off, since $R > P$.

Game theory says the teacher is wrong

John Nash proved that, allowing pure and mixed strategies, every finite n-person game has at least one equilibrium point (EP).

The *Nash equilibrium solution concept* holds that the "solution" or predicted outcome of any game (among rational players) will be an EP. Since the one-shot prisoner's dilemma has only one EP, the Nash equilibrium solution concept predicts that both players will defect.

Further, for any fixed number, n, of iterations of Prisoner's Dilemma, a backwards induction argument suffices to prove that the n-period Iterated Prisoner's Dilemma (IPD or RPD) game still has exactly on EP: both players play D in each game.

Experiments say the teacher is right

It is interesting, and significant, that in the first human experiment with repeated prisoner's dilemma the human subjects were asked to record their thoughts as the game was being played [**?**].[a]
Comments such as

- "Perverse!"
- "Oh ho! Guess I'll have to give him another chance."
- "In time he could learn, but not in ten moves so:"
- "What's he doing?!!"
- "I'm completely confused. Is he trying to convey information to me?" and
- "This is like toilet training a child—you have to be very

---

[a] The relevant data are most conveniently reproduced in Poundstone's excellent and easily available popular treatment of the prisoner's dilemma game: [**?**, pages 106–123].

patient.”

appear throughout the 100 iterations of the game. Even so, the two subjects jointly cooperated in 60 of the 100 iterations. By the lights of classical game theory this was a remarkably rewarding triumph of irrational behavior.[b]

---

[b] These results are not inconsistent with subsequent empirical findings.

Consider the one-shot PD

By defecting, the teacher can guarantee that she will receive a payoff greater than S. Similarly, by not announcing a surprise exam as she did, the teacher can avoid uttering a falsehood (in that case). If the teacher/player is willing to undertake some risk, however, there is also a chance that the teacher can do better than to receive P or to do without a surprise exam. Suppose the teacher has the following policy: play C with probability r and play D with probability $(1-r)$. If the other player plays the same strategy, then the expected return for each player, $E(r)$, is:

$$R \cdot r^2 + T \cdot (r - r^2) + P \cdot (1-r)^2 \tag{1}$$

on the harmless assumption that $S = 0$. Rearranging, the gamble pays off (in expectation) if

$$\frac{R \cdot r^2 + T \cdot (r - r^2)}{(1 - (1-r)^2)} > P \tag{2}$$

or

$$\frac{R \cdot r + T \cdot (1 - r)}{(2 - r)} > P \tag{3}$$

Obviously, setting $r = 1$ (for both players) yields an expected value of R, and for fixed T, R, and P (with $S = 0$) this maximizes the expected value.

More interestingly, note that on the left-hand side both the numerator and the denominator are positive, so fixing r and P, it is always possible to increase R and T sufficiently to make the inequality hold. Of course, on the standard assumptions of game theory, this should not matter.

This is but a crude model of how a player might reasonably deal with risk in the PD game. Still it tells us something. Let $d = T - R$. Prisoner's Dilemma requires that $0 < d < R$. Fix d at some small value, say $d = 2$ in:

$$\frac{R \cdot r + (R + d) \cdot (1 - r)}{(2 - r)} > P \tag{4}$$

In the usual PD problem, $R = 3, T = R + d = 3 + 2 = 5, P = 1, S = 0$. Using the $R = R + d$ formulation, fix d, P, S. We see from expression (4) that $E(e)$ (the expected return for each player, assuming an independent probability of r of coöperating) increases as R increases, for any fixed value of r. Suppose the outcome values are dollar amounts. Let $R = 100$ ($R = \$100, T = \$102$ etc.).

Consider now our rational human teacher

and her rational human counter-player. Each sees the situation; each understands that the goal is to maximize dollars captured individually, not to get more dollars than the other player. It remains true that each player can with certainty avoid the sucker's payoff S by defecting. . . . Surely, many players would reason that jointly they have much to lose by not coöperating and little extra to gain by individually defecting. Given that the other player has a coinciding interest in mutual coöperation, why not take a chance an play C? Surely as well, the strength of such sentiments increases with R. Suppose $R = \$1,000,000$. Suppose it is much larger than that.

Believe it? OK. Don't? OK.

Consider now the teacher's 100-play IPD game. Even if the teacher and her counter-player both find the one-shot PD reasoning above unconvincing, they surely would give pause when faced with 100 plays, each with a reward of \$1,000,000 for mutual coöperation. Do they both really want to follow a policy of defecting every time, no matter what? Must they conclude that their actions cannot have an effect—positive or negative—on the other player? Surely that is an awfully strong and unduely pessimistic assumption about a supposedly rational player. Why not try coöperating and if it is reciprocated continue to do so?

If these numbers aren't convincing for these arguments, increase the number of plays to a million and R to a billion. Increase them all you want. If the teacher's r is large enough given T, R, P, S, the chance she will get her points is, I think we should agree, an excellent one.

The point may be summarized in the following manner.

Suppose each player reasons that there is some chance, r, that the counter-player can be induced during the 100 iterations to play C fairly often. Given the counter-player's interest in maximizing returns (as opposed to gaining relative points only), is it reasonable to assume, without probing, that $r = 0$? Let there be exactly n iterations of the game, known to players. Let $P = \$\frac{1}{n}$. Let $d = T - R$ be small, as above. Are there no large values of n and R for which it would not be folly not to probe the counter-player for mutual coöperation? This obviously rhetorical question answers itself in the negative.

Comments

- “Birds do it, bees do it. . . ”
- People do it.
- Horses do it.

The Nash equilibrium is a solution *concept* for n-person games. The concept is that the games are solved by finding an EP (equilibrium point). That is where the players will end up. Long-established experience has shown that humans in (definitely) Iterated Prisoner’s Dilemma games consistently do better than mutual all defect. Assuming no pervasive flaws in experiments conducted over a 50-year period, either the human subjects are consistently, egregiously irrational, or the Nash equilibrium concept is flawed (or both).

Contra Nash

The fundamental error on the part of the students was to assume that the teacher's r is, or even must be, 0. The students assumed that in making the tradeoff between risk (of speaking falsely) and reward (being able to offer a surprise exam), the teacher would place no value (or no sufficiently large value) on the reward, at the expense of taking some risk. Similarly, I have argued, in the definitely IPD both players face a tradeoff between risk (of getting the sucker's payoff, S) and reward (achieving P very often during the iterations). The flaw in the Nash equilibrium solution concept (at least for IPD) is to impose the assumption that both players are unwilling to trade any risk at all (however small) for any reward at all (however large).

On Beyond Nash (Something positive.)

Suppose we have a game with n players (in IPD, this n is 2). At the conclusion of play, each player i has played some strategy, s_i. This constitutes the *strategic configuration* or SC.

$$SC = (s_1, s_2, \ldots, s_n) \tag{5}$$

Let $H_i(s_1, s_2, \ldots, s_n) = H_i(SC)$ be the payoff to player i given the strategies, including i's, played in SC. An *equilibrium point* or EP is any SC^* such that for each $i = 1, 2, \ldots, n$

$$H_i(SC^*) = \max_{s_i} H_i(s_1^*, s_2^*, \ldots, s_i, \ldots, s_n^*) \tag{6}$$

In other words, an EP (equilibrium point) is an SC (strategic configuration) such that no individual player can (or could) unilaterally do better by picking a different strategy than the one the player has in SC. The Nash equilibrium solution concept for games is that among rational players every game will conclude at an EP.

Generalizing the Nash solution concept

Given s, a particular SC, define the *improvement vector* for s or IV_s, as

$$IV_s = (c_1, c_2, \ldots, c_n) \tag{7}$$

where c_i is the *count* or number of ways i distinct players can jointly alter their strategies in such a way that each of the i players does equally well or better. More carefully, if $c_3 = 2$ then there are two distinct groups of 3 players who collectively have strategies that if taken, while all the strategies outside the group remain as in s the relevant SC, would make everyone in the group no worse off. Now a single group of 3 (or whatever) might have many ways to do this. IV_s ignores this and only counts the number of groups of a given size with 1 or more opportunities. (We can also define a *strict improvement vector* or SIV, as an IV in which everyone in every group is strictly better off.)

Points arising

1. The IV concept is a generalization of the EP concept. $IV_{SC} = (0, c_2, \ldots, c_n)$ if and only if SC is an EP.
2. The payoff vector, H_s, of a strategic configuration s is denoted:

$$H_s = (H_1(s_1), H_2(s_2), \ldots, H_n(s_n))$$

3. In the one-shot PD, the *table of space of improve vectors* or the *improvement space table* or IS is:

SC	IV	H_{SC}	target Hs
(D,D)	$(0,1)$	(P,P)	(R,R)
(D,C)	$(1,0)$	(T,S)	(P,P)
(C,D)	$(1,0)$	(S,T)	(P,P)
(C,C)	$(2,0)$	(R,R)	(T,S)
			(S,T)

where the target Hs are the payoffs if the associated IVs are realized.

4. We might similarly define how the players might act to make things worse. This would result in a disimprovement space table.
5. We say that SC_x *strictly dominates* SC_y iff for all $i = 1, 2, \ldots, n$,
   $$H_i(s_i^x) > H_i(s_i^y)$$
   or equivalently
   $$H_x > H_y$$
   Note that (C, C) strictly dominates (D, D) in the one-shot PD. (Hence the dilemma for the Nash solution concept.)
6. In the n-shot IPD, there is a correspondingly larger improvement space, generated by the Cartesian product of the one-shot spaces.

7. Any well-defined game will have well-defined improvement and disimprovement spaces. In learning to play the game, or in rational deliberation about it, players should be thought of as exploring these spaces.

   At risk of being wrong, but with potential for reward, players can form hypotheses about their counterparts, as play unfolds (actually or prospectively).

8. Given a well-defined game, with its improvement/disimprovement spaces and attendant payoff vectors, how play actually unfolds will depend upon the strategy formation and selection algorithms employed by the players.

9. In general and under realistic assumptions regarding strategy formation and selection algorithms, ending at an EP will be unusual in games with IVs having non-zero terms past the first position with attendant large payoffs.

/* Note: This is what the now large literature on computational dynamics in games has yielded—well something consistent with this. */

10. We can look for and expect robustness results for classes of learning algorithms in games. Think: replicator dynamics, reinforcement learning. Convergence (and speed of it) to SCs with particular properties. Example: in Q-learning, with potential reward R and risk r, the players will achieve the R with such-and-such characteristics.

    And that is how we can improve predictions and better understand decision processes (human or not).