Authors Queries

Journal: **Journal of Economic Methodology**
Paper: **332312**
Title: **The surprise exam paradox, rationality, and pragmatics: a simple game-theoretic analysis**

Dear Author
During the preparation of your manuscript for publication, the questions listed below have arisen. Please attend to these matters and return this form with your proof. Many thanks for your assistance

Query Reference	Query	Remarks
1	Please provide Keywords.	
2	Fitch (1964): References section has 1972; if that is not the source intended here, please supply details of 1964 source for References and delete the 1972 entry.	
3	Smullyan (1987): References section has 1978 but 1987 is correct?	
4	Gardner (1962): not in References section; please supply.	
5	Popper (1962): not in References section; please supply.	

332312
Journal of Economic Methodology jec160296.3d 30/7/08 14:52:43
The Charlesworth Group, Wakefield +44(0)1924 369598 - Rev 7.51n/W (Jan 20 2003)

7	'The assumption that an exam must be given is Definitions 1-3 next give different versions of the meaning of a surprise exam.' – does not seem to make sense; please clarify.	
8	The second table was given the same table number and title as the first. I have changed it to Table 2 but assume the title should be different?	
9	Is coordination correct here rather than cooperation?	
10	Olin (1982): not in References section; please supply.	
11	'Backward induction … surprise exam paradox' – You styled this as a quotation: please give source, or if not a quotation please instruct to change style to ordinary text.	
12	Again, 'If when analyzing … that subgame' is styled as a quotation: please give source, or if not a quotation please instruct to change style to ordinary text	
13	Cohen 1950: References section has 1959; if that is not the source intended here, please supply details of 1950 source for References and delete the 1959 entry.	

Journal of Economic Methodology **jec160296.3d** 30/7/08 14:52:58
The Charlesworth Group, Wakefield +44(0)1924 369598 - Rev 7.51n/W (Jan 20 2003)

Journal of Economic Methodology,
Vol. 15, No. 3, September 2008, 285–299

Routledge
Taylor & Francis Group

The surprise exam paradox, rationality, and pragmatics: a simple game-theoretic analysis

José Luis Ferreira[a] and Jesús Zamora Bonilla[b]*

[a]*Universidad Carlos III de Madrid, Spain;* [b]*UNED, Madrid, Spain*

The surprise exam paradox has attracted the attention of prominent logicians, mathematicians and philosophers for decades. Although the paradox itself has been resolved at least since Quine (1953), some aspects of it are still being discussed. In this paper we propose, following Sober (1998), to translate the paradox into the language of game theory to clarify these aspects. Our main conclusions are that a much simpler game-theoretic analysis of the paradox is possible, which solves most of the puzzles related to it, and that this way of analysing the paradox can also throw light on our comprehension of the pragmatics of linguistic communication.

Keywords:

Tú juegas a engañarme,
yo juego a que te creas que te creo.
Luz Casal, ''No me importa nada.''

1 Introduction

Pragmatics, as a branch of the study of language, has been inspired by the fact that communication is a type of action. This entails that the intentions of the agents, as well as the assumption that speakers are rationally pursuing the fulfilment of their goals, will play a necessary role in the explanation of their communicative actions. Furthermore, being a process in which at least two agents intervene, communication can best be understood from the point of view of game theory. In the past decade, a lot of research has been done in this direction, of which this monograph is an example, but in our paper we shall concentrate on a deep tension existing between some basic presuppositions underlying pragmatics and game theory, respectively. We are referring to the idea, originally advanced by Paul Grice (1975) that the pragmatic aspects of conversation are essentially guided by norms of cooperation: the hearer can understand the speaker's message not only thanks to the former guessing the latter's intentions, but also thanks to the assumption that the speaker is respecting the 'conversational maxims' (as 'be as informative as required, but not more', 'don't say what you believe to be false', 'be relevant', 'be clear', and so on). From the point of view of game theory, this cooperation principle, which underlies most of the mainstream research in pragmatics, is problematic, because of at least the following three reasons: first, whether cooperation is rational or not will depend on the structure of each game, and cannot be assumed as a universal strategy or constraint on individual actions (this connects with the general problem of how to

*Corresponding author. Email: jpzb@fsof.uned.es

ISSN 1350-178X print/ISSN 1469-9427 online

DOI: 10.1080/13501780802321459
http://www.informaworld.com

make any normative principle relevant in a game-theoretic account of individual action); second, as a matter of fact, speakers often fail to respect one or other of the conversational maxims that specify the content of the cooperation principle, usually because its disobedience is in their interest; third, a merely commonsense, casual justification is given for the maxims, as they are not inferred from some more general constraints (i.e. an explanation is not offered of why the game of language – as actually observed, or in general – could not proceed according to different rules). On the other hand, the attempt to introduce verbal communication within the structure of mathematical games, though quite successful and productive (e.g. signalling games), has also proved to lead to numerous difficulties, both conceptual and technical (see Mailath, Okuno-Fujiwara, and Postlewaite (1993) for a discussion on the logical foundations of equilibrium selection in signalling games), and this has suggested to many the idea that game-theoretic models of knowledge and reasoning have to be modified by, or at least complemented with, some types of non-classical logic.

Our goal in this paper is not to analyse these tensions between mainstream pragmatics and game theory (see Parikh 2000, and Glazer and Rubinstein 2006 for a similar criticism), nor to defend the sufficiency of classical game-theoretic logic in general, but simply to illustrate, by means of a philosophically relevant example, the possibility of interpreting linguistic communication with the assumption that it is not necessarily based on cooperative principles. This example is the well-known 'surprise exam paradox'. We will show that this paradox can be interpreted from the viewpoint of game theory (following Sober 1998) without the need of introducing more-than-usual complex considerations about knowledge and reasoning, and also in a way that allows us to derive some interesting conclusions about the pragmatics of communication.

The most common version of the surprise exam paradox is as follows. A teacher announces to her students that she will give them an exam during one of next week's classes, and that the exam will be a surprise. The students reason that she cannot possibly fulfil her intention. The exam cannot be held on Friday: if it were, they would expect it after Thursday's class (having noted that no exam had yet been given). Once Friday is ruled out, the exam must take place on one day of Monday through Thursday. But then, for exactly the same reason, it cannot be held on Thursday or, by backward induction, any other day. Arriving in class next Tuesday, the students discover that they are to take an exam that day. None of them, of course, expected it. What was wrong with the students' reasoning?

This paradox has attracted the attention of prominent logicians, mathematicians and philosophers for decades (Popper, Quine, Good, Gardner and Smullyan among them), and the number of articles about its resolutions in serious philosophical and mathematical journals passes the 100 mark.

Although the paradox itself has been resolved at least since Quine (1953), some aspects of it are still being discussed. Until recently, all discussions have taken place within the realm of logic. However, in recent years, a few authors have found game theory as a more natural model to study the paradox. Although the use of game theory has already clarified important issues regarding the paradox, we think that there are others that still have not been addressed within its framework. Also, some of the authors who use game theory open new issues that also deserve our attention. The present work is also about the discussion of some of these old and new issues.

Timothy Chow (1998) provides a good perspective about all serious solutions to the paradox, and the reader is referred to this work and references within for the formal arguments. Here we will summarize (borrowing from Chow) how the classical resolutions work.

The logical school shows that the students' assumption (that there will be a surprise exam next week) is not sound. In fact, most formalizations easily find that the statement of the teacher is self-contradictory and, therefore, that any form of reasoning by the students based in the truth of the statement is doomed. This is the approach in Shaw (1958), Fitch (1964), Bosch (1972) and Windt (1973), to name a few of the earliest works. It is the interpretation of this contradiction that still causes some confusion.

Smullyan (1987) provides an example of the resolution of the paradox in words:

> Actually, the professor said two things: (1) You will get an exam someday this week; (2) You won't know on the morning of the exam that this is the day. I believe it is important that these two statements should be separated. It could be that the professor was right in the first statement and wrong in the second. On Friday morning, I couldn't consistently believe that the professor was right about both statements, but I could consistently believe his first statement. However, if I do, then his second statement is wrong (since I will then believe that I will get the exam today.) On the other hand, if I doubt the professor's first statement, then I won't know whether or not I'll get the exam today, which means that the professor's second statement is fulfilled (assuming he keeps his word and gives me the exam). So the surprising thing is that the professor's second statement is true or false depending respectively on whether I do not or do believe his first statement. Thus the one and only way the professor can be right is if I have doubts about him; if I doubt him, that makes him right, whereas if I fully trust him, that makes him wrong!

A logician will find a version of the *liar paradox* in these words. In fact, authors like Gardner (1962) and Popper (1962) reduce the paradox of the unexpected exam to this most famous liar paradox. A game theorist will detect an argument similar to the ones shown to question backward induction – like in Binmore (1987, 1988). This is just one example of the issues that deserve clarification. Even if we do not question backward induction and reject Smullyan's claim that 'On Friday morning, I couldn't consistently believe that the professor was right about both statements', still his overall argument is basically correct. In other words, it is not necessary to reject backward induction to get the contradiction.

Another source of confusion is the fact that if the teacher's statement is contradictory, then we still have to show how it was, after all, vindicated by giving a surprise exam on Tuesday. Other aspects include the role that the act of announcing the exam plays in the contradiction of the teacher's statement, and the implicit assumption of the students' memory.

The epistemological school tries to overcome some of the previous problems by defining a knowledge operator (these operators are familiar to game theory) along with a series of properties that are added to the rules of inference in logic. This approach can be seen in Sorensen (1984). This school succeeds in providing a perhaps more rigorous approach by making explicit assumptions, and also finds a contradiction between the axioms of knowledge and the fact that the students know (as true) the teacher's statement. Still some problems remain (why the teacher seems to be right after all?) and new ones arise, like which of the assumptions one should drop to resolve the paradox (some of the axioms or the knowledge of the students?) – see, e.g. Quine (1953) and Olin (1983) – or what does it mean that the students know

the teacher's statement. The statement that the students are supposed to know is a statement about the students' inability to know certain things. In Chow's words:

> Consider the statement, 'It is raining and John Doe does not know it.' 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.

It may be too early to talk of a game-theoretical school of the resolution of the surprise exam paradox. There are only a few works that use game theory to resolve or illustrate some aspects of the paradox. Sorensen (1988) argues that there is an analogy between the surprise exam paradox and the finitely repeated prisoner's dilemma. Olin (1988), in contrast, argues against this analogy. The most fruitful use of game theory can be found in Sober (1998). Sober provides a neat solution to the paradox, modelling it as a repeated matching pennies game, and then goes on to discuss issues such as the assumption that the students believe (and use) the same distribution as the teacher, offering a distinction between prudential and evidential prediction. As with other solutions of the paradox, we find this one basically correct (in the sense of providing a logical model to explicitly show the contradiction in the students' assumptions), but also find that its framing provokes a confusing discussion, like the mentioned distinction between prudential and evidential prediction. Gilboa and Schmeidler (1988) provide a different approach: they model the surprise exam paradox as an example of information-dependent games. These games constitute a new model, different from normal form games in that a new element is added, namely the set of possible prediction profiles, and in that the payoff function depends not only on the chosen strategies but also on the chosen prediction. For a suitable definition of the term, they show that there is no informationally consistent way of playing the surprise exam game, thus reducing the paradox to an impossibility theorem. We will argue that this approach leaves some questions open.

Section 2 provides the game-theoretical analysis of the paradox. Section 3 connects our analysis of the paradox with the Gricean approach to pragmatics. Section 4 shows how the approach in section 2 clarifies many of the logical and epistemological issues addressed in the literature of the paradox. Section 5 concludes.

2 A simple game-theoretic approach to the surprise exam paradox

We will illustrate the paradox and its solution in a game-theoretical framework that includes Sober's as a special case. We will then show that many questions can be easily expressed and clarified in this language. The model will formalize the following exposition of the students' reasoning and of the resolution of the paradox.

- *The students' argument*. (1) The students assume the teacher's statement to be true; (2) the students rule out Friday; (3) by backward induction they rule out any other day; and (4) they conclude that there cannot be any exam at all.
- *The mistakes in the students' argument*. Assuming (1)–(3) above, the students' correct conclusion should have been that either (4) there will not be an exam at all or that (5) if there is an exam, it will not be a surprise. In order to conclude (4) they implicitly reject (5) on the basis of the teacher's statement, but they may have as well rejected (4) to conclude (5). In any case, assuming

332312
Journal of Economic Methodology jec160296.3d 30/7/08 14:53:01
The Charlesworth Group, Wakefield +44(0)1924 369598 - Rev 7.51n/W (Jan 20 2003)

the teacher's statement is true leads to a contradiction. This means that there are no sound bases to rule out Friday and start the backward induction argument (Sober (1998) discusses this point and gives nice intuitions to it.)

- *Why there is a surprise exam after all.* If the teacher's statement is false then there may or may not be an exam, and if there is an exam, it may or may not be a surprise. For any sensible meaning of the term 'surprise', we will see that the negation of the teacher's statement does not mean that there can never be a surprise exam. The fact that on Tuesday she is able to give a surprise exam is just one possible realization than cannot occur with probability one.

2.1 The game form

We will restrict the paradox to a week of only two days, 1 and 2, and display the game forms of different interpretations and assumptions. The students act as one player.

- *Game form 1.* On day 1 the teacher may give an exam (E) or not (NE), and the students may anticipate it (A) or not (NA). Only if the teacher chooses NE and the students choose NA does the game go to the second day. On the second day the teacher must give an exam, and the students may anticipate it (A) or not (NA), as on the first day. The game ends.
- *Game form 2.* On day 1 the teacher may give an exam (E) or not (NE), and the students may anticipate it (A) or not (NA). Only if the teacher chooses NE and the students choose NA does the game go to the second day. On the second day the teacher may give an exam (E) or not (NE), and the students may anticipate it (A) or not (NA), as on the first day. The game ends.
- *Game form 3.* On day 1 the teacher may give an exam (E) or not (NE), and the students may anticipate it (A) or not (NA). If the teacher chooses NE the game goes to the second day. On the second day the teacher must give an exam, and students may anticipate it (A) or not (NA), as on the first day. The game ends.
- *Game form 4.* On day 1 the teacher may give an exam (E) or not (NE), and the students may anticipate it (A) or not (NA). If the teacher chooses NE the game goes to the second day. On the second day the teacher may give an exam (E) or not (NE), and students may anticipate it (A) or not (NA), as on the first day. The game ends.

In game forms 1 and 2 the game ends before the last day after an exam has been given or after it has been anticipated, whereas in game forms 3 and 4 the game ends before the last on the first day if an exam was given. The assumption that an exam must be given is definitions 1–3 next give different versions of the meaning of a surprise exam.

- *Definition 1.* In the Game form i (i=1,2,3,4), it is said that a surprise exam occurs if the actions (E,NA) are chosen on some day with positive probability conditioned on being on that day.
- An alternative and stronger definition is:
- *Definition 2.* In the Game form i, it is said that a surprise exam occurs if the actions (E,NA) are chosen on some day with positive probability.
- An even stronger definition can be given:

- *Definition 3*. In the Game form i, it is said that a surprise exam occurs if the actions (E,NA) are chosen on some day with probability one.

Since both the teacher (T) and the students (S) only care about the final outcome, and all game forms end after the first day on which an exam is given, there are four utility levels to consider for each player: $u_i(E, A)$, $u_i(E, NA)$, $u_i(NE, A)$, and $u_i(NE, NA)$. We will assume:

(1) $u_T(E, NA) > \max\{u_T(E, A), u_T(NE, A), u_T(NE, NA)\}$;
(2) $u_T(E, A) < \min\{u_T(E, NA), u_T(NE, A), u_T(NE, NA)\}$;
(3) $u_S(E, A) > \max\{u_S(E, NA), u_S(NE, A), u_S(NE, NA)\}$; and
(4) $u_S(E, NA) < \max\{u_S(E, A), u_S(NE, A),, u_S(NE, NA)\}$.

Clearly the teacher derives her highest utility after a surprise exam, and her lowest after an exam has been anticipated by the students. For the students the preferences are reversed. These preferences are consistent with the story. If the teacher's announcement is going to have any impact on the students, it must be understood that the teacher actually wants to give a surprise exam. Otherwise, the students cannot start any reasoning and can conclude nothing, from which no paradox arises. Tables 1–4 show the first and second days for all game forms.

Now, to study whether a surprise exam can occur in a subgame perfect equilibrium in any of the game forms according to any of the definitions is a simple exercise. This is the content of Proposition 1.

Proposition 1: In all game forms, a surprise exam occurs according to definitions 1 and 2, but not according to definition 3.

Table 1. Game forms 1 and 2.

Day 1	A	NA
E	$u_T(E, A)$, $u_S(E, A)$	$u_T(E, NA)$, $u_S(E, NA)$
NE	$u_T(NE, A)$, $u_S(NE, A)$	Go to day 2

Table 2. Game forms 1 and 2.

Day 1	A	NA
E	$u_T(E, A)$, $u_S(E, A)$	$u_T(E, NA)$, $u_S(E, NA)$
NE	Go to day 2	Go to day 2

Table 3. Game forms 1 and 3.

Day 2	A	NA
E	$u_T(E, A)$, $u_S(E, A)$	$u_T(E, NA)$, $u_S(E, NA)$

Table 4. Game forms 2 and 4.

Day 1	A	NA
E	$u_T(E, A)$, $u_S(E, A)$	$u_T(E, NA)$, $u_S(E, NA)$
NE	$u_T(NE, A)$, $u_S(NE, A)$	$u_T(NE, NA)$, $u_S(NE, NA)$

Proof:

- *Game form 1*. On day 2, the teacher must choose E (actually, this is not even a choice, as E is the only action). The students know this and choose A (i.e. a surprise exam cannot take place on day 2). On day 1, the only equilibrium occurs in mixed strategies. This means that (E,NA) occurs on day 1 with a positive probability, but not with probability one.
- *Game form 2*. Only a mixed strategies equilibrium exists in the subgame on day 2. Thus, payoffs after (NE,NA) on day 1 are the payoffs of this subgame, that must be between the highest and the lowest for both players. This implies again a mixed strategies equilibrium on day 1.
- *Game form 3*. In both subgames on day 2, the teacher must choose E. The students know this and choose A (i.e. a surprise exam cannot take place on day 2.) Substituting these subgames with their payoffs, we find that, as before, only a mixed strategies equilibrium exists on day 1.
- *Game form 4*. Only mixed strategies equilibria exist in the two subgames on day 2. Thus, payoffs after (N,N) and after (N,E) on day 1 are the payoffs of this subgame, that must be between the highest and the lowest for both players. This implies again a mixed strategies equilibrium on day 1.

The fact that no pure strategies equilibrium exists on day 1 (and on day 2 in game forms 2 and 4) implies that a surprise exam occurs according to definitions 1 and 2, but not according to definition 3. QED.

In all game forms, the conclusion is the same: a surprise exam occurs according to definitions 1 and 2, but not according to definition 3. Further, it is impossible to give a surprise exam with probability one or with probability zero in a subgame perfect equilibrium. The analysis can easily be extended for weeks of more than two days.

It is easy to see that no other way to choose strategies (e.g. as in correlated strategies via a public or private signal) can make a difference. Both the teacher and the students will select private randomizations and end up with a surprise exam with a probability strictly between zero and one.

The analysis in Sober (1998) uses a matching pennies game as the stage game, that goes to the second stage if the professor didn't give any exam the first day. In this second stage, the professor must give an exam. This corresponds to our game form 3 when utilities are $u_T(E, NA) = u_T(NE, A) = u_S(E, A) = u_S(NE, NA) = x$, and $u_S(E, NA) = u_S(NE, A) = u_T(E, A) = u_T(NE, NA) = -x$. That utilities take these particular values is irrelevant for the discussion as long as assumptions (1)–(4) on utilities are satisfied. Finally, Sober's definition of a surprise is equivalent to our definition 1 (although his terms are that a surprise occurs with a given probability, not that a surprise simply occurs as in ours). Then, all discussions that in the sequel use game form 3 to make a point could also use Sober's model.

3 The surprise exam and the maxims of conversation

What does our analysis of the surprise exam tell us about the possibility of explaining communication through the Gricean cooperation principle? First and foremost, it is clear from the students' reasoning that the proposition the teacher

utters is self-contradictory, and this apparently constitutes a blatant violation of the Gricean maxims. After hearing the teacher's assertion, the students are faced with the following dilemma: either the teacher (wrongly) believes that her claim is logically consistent, or she knows it is inconsistent. In both cases, the students have to guess the intentions of the teacher, and from a game-theoretic point of view this amounts to guessing what is the game are they going to play (as, e.g. the games depicted in Tables 1 to 4). If the teacher has been inadvertently inconsistent, then the problem for the students is to learn what logical mistake on the part of the teacher was the most likely one: either she wants to make a surprise exam, but has ignored the logical incompatibility of her intentions by announcing them, or has committed a different shortcoming in her reasoning. Since it is difficult to imagine what other logical mistake may have led to the teacher's claim (while assuming the overall consistency of her thought), the students will conclude that the teacher plans to make a surprise exam and will proceed according to some of the game forms discussed above. On the other hand, if the teacher has made the inconsistent claim deliberately (which is the default assumption under the hypothesis that agents are rational), more options are available: perhaps she wanted the students to doubt her claim, so as to make the surprise more likely (but why through this Machiavellian strategy, instead of by not announcing the exam at all?), or perhaps she had no intention of making any exam, and only wanted the students to engage in some kind of philosophical reflection (imagine the teacher is a Zen monk uttering a koan; a possibility like this one will lead to the inclusion of a new game in our analysis of the paradox, but one in which this paradox does not arise). In any case, what the example makes clear is that there is no direct way of interpreting the teacher's intentions from her bare claim, nor even by making use of 'conversational practices': the students have to take her utterance as a datum to guess the likelihood of the possible intentions of the teacher. Of course, information about the linguistic conventions (e.g. what specific conversational rules) that are followed in the community is useful in the students' reasoning, but this amounts to the same problem that rational players face in any coordination game (as understood, at least, since Lewis 1969), and the fact that in this case coordination is 'linguistic' makes little difference. 9 9

Second: as noted particularly by Glazer and Rubinstein (2006), the hypothesis that the agents follow the cooperation principle is particularly unlikely in those contexts where the interaction between speakers and hearers is one of competition or conflict. Those authors concentrate on cases of persuasion, whereas the surprise exam paradox points to a different type of context: such as where one player may have an advantage from providing false or confusing information, or by making the smooth process of reasoning of the other players more difficult in some way. It is true that language serves the purpose of coordination, but it is not less true that language can also serve to deceive or mislead other people. Arguments have been developed from Kant to Habermas trying to show that truthfulness is a constitutive 'condition of possibility' of communication, and that this has both normative and explanatory consequences (see Origgi 2004 for a clear and recent statement of this position). But any theory about the use of language must cope with the fact that the deliberate transmission of irrelevant, false, or confusing information is by no means a marginal phenomenon. The fact that these phenomena abound not only needs an explanation in statistical terms (e.g. what 9

degree of compliance with the existing conversational maxims constitutes an equilibrium), but also requires that we recognize the possibility of analysing the behaviour of language users under the assumption that they can expect some misbehaviour on the part of the others.

4 The classical aspects of the paradox

In this last section we discuss most of the questions that have been discussed at length in the literature in relation to the paradox (in general, not directly in connection with language pragmatics), and show that these are easily resolved thanks to our simple game-theoretical approach.

The teacher's statement is self-contradictory. As we saw in section 2, both the logical and the epistemological school have established this fact a long time ago. In the game-theoretical framework, this contradiction is translated into the fact that the situation described by the teacher cannot occur with probability one. There are two ways to contradict the teacher's statement: there is no exam or, if there is an exam, it will not be a surprise. Of these two possibilities, only the first is a control variable in the hands of the teacher in game forms 1 and 2. For the second we need the interaction with the students.

The teacher's statement is vindicated. The teacher's statement seems to be vindicated, the story goes, when Tuesday comes and the exam is given. But this is just because the story tells us about one possible realization of the chosen actions. We have seen that the probability that this happens cannot be one. The teacher must choose to wait until the last day with probability greater than zero. That is, another realization of the story, which must occur with positive probability, but which the paradox teller does not mention, is that the professor waited until the last day to give the exam. The earliest work to point out this fact was, perhaps, Meltzer and Good (1965). In a more recent study, Borwein, Borwein, and Maréchal (2000), recognizing that there cannot be a surprise with probability one, compute the distribution that maximizes the surprise the teacher can create, where the measure of surprise is a suitable definition of entropy.

This also suggests that a possible interpretation of the working of conversation maxims is that speakers do not simply have an obligation of fulfilling them, but there can be a trade-off between the other goals of the speakers and the constraint of fulfilling the maxims 'as much as possible'. So the students can reason that the teacher is going to take that decision which makes her statement 'as close to the truth as possible'. In connection with that, an interesting line of research would be that of trying to connect strategic signalling games with the abundant literature on counterfactuals and on truthlikeness in philosophy of science, which have systematically explored the intuition that false, but close to the truth, statements are essential to scientific knowledge. (Two classical references are Lewis 1973 and Niiniluoto 1987.)

The statement of the teacher changes everyone's life. Without it, the teacher and the students may face surprise exams, but they can take place any day (not only next week). The announcement makes the game common knowledge.

The statement of the teacher is self-referential. The logical school finds that the paradox after the teacher's statement can be traced to its self-referential nature. As Shaw (1958) puts it, the surprise can be defined in this way:

1. The exam will take place next week.
2. It will take place on a day that cannot be deduced in advance from the preceding axiom plus this axiom.

Others argued that not all self-referential statements are forbidden. However, in the language of game theory, there is no need to talk about self-referential sentences to explain the paradox. We do not have an illegitimate use of language creating a paradox (as in the liar's paradox). What we have is a paradox created by (1) the fact that the students develop an illegitimate backward induction argument after assuming a false premise (that a surprise exam with probability one is possible); and (2) a false sense that the premise is true (after all, there is a surprise exam in the story). Recall Smullyan's words:

> … the one and only way the professor can be right is if I have doubts about him; if I doubt him, that makes him right, whereas if I fully trust him, that makes him wrong.

We have seen that both the teacher and the students are wrong: the teacher, because she cannot give a surprise exam with probability one, and the students because they cannot start their backwards induction argument. And this is the end of it; i.e. it is not the case that the truth value of a group of statements is not well established (e.g. if A is true, B is false, but if B is false A is false, but if A is false B is true, …) as required in liar's paradoxes.

What does it mean that the students know the teacher's announcement?. The students cannot know something that is not true. The only reasonable interpretation is that what the students know is the fact that the teacher has made the announcement. This is translated into the fact that the teacher prefers to give a surprise exam rather than to give an exam that is anticipated by the students. Recall that the teacher's announcement says two things: (1) there will be an exam next week; and (2) the students will not be able to predict the day of the exam. Game forms 1 and 3 assume (1) (which is then common knowledge) and conclude that (2) cannot occur with probability one. Game forms 2 and 4 do not assume (1), and also conclude that (2) cannot occur with probability one. The second part of the statement cannot happen with probability one, therefore the students cannot know (or assume, without reaching a contradiction, that they know) both (1) and (2).

The notion of surprise. Typically, a surprise exam is defined as a situation in which the teacher gives the exam on one day, and the students do not know (or are not able to deduce) it. Other authors try a probabilistic approach, and say that students are not surprised if, when the professor gives the exam, they are able to determine that the probability is higher than a threshold. Our analysis shows that, in equilibrium, students are able to deduce the probabilities used by the teacher. This is because the equilibrium can be calculated by anyone that analyses the game, including all players. Do we want this to mean that the students are not surprised? If we do, then there is never a surprise exam in our models as long as the game is there for anyone to analyse. The fact that an exam is given and students are not prepared is just a possible realization that occurs with a known probability, hence there is never a surprise with this definition. But, on the other hand, if this is our definition of surprise, we solved the wrong game. The game that the teacher and the students will be playing would be one in which both the teacher and the students choose a probability distribution at random. It they choose the same, there is no surprise, while if they choose differently, there is a surprise. In this game, if the teacher

332312
Journal of Economic Methodology jec160296.3d 30/7/08 14:53:02
The Charlesworth Group, Wakefield +44(0)1924 369598 - Rev 7.51n/W (Jan 20 2003)

chooses his distribution randomly, the probability of the students guessing it is zero, and we are back to a situation with surprise exams even with this definition.

In any case, the analysis above started with a semantic discussion, and the wording of the paradox clearly indicates that the ability of the students to anticipate the probability distribution chosen by the teacher is not what constitutes the surprise. Therefore, surprise must be defined as the lack of knowledge of the students about the teacher's action or, as in definitions 1–3, as the observation of a behaviour that is impossible (in equilibrium) if the students know the teacher's action. Since the game-theoretical analysis shows that the teacher chooses randomly between performing or not the exam the first day, it follows that the students can never know for sure when the exam takes place (except on the last day, but only conditional on reaching that day, and only in game forms 1 and 3). The probabilistic approach that is permitted in this framework does not imply that we must select a threshold to define a surprise. It could be done, but it is not necessary to model the paradox or to explain it.

Students may rule out the last day, but no other day. This conclusion has been reached in some articles in the literature (typically the ones that assume that an exam must take place). Here, this is not the case. The last day can never be ruled out (in the sense that the exam will never take place that day), even if we accept that an exam must occur. There is always a probability that the game goes to the last day. The students cannot even start the backward induction argument that, in the paradox, rules out all days.

The last day is special. Some works argue that the last day is special in the sense that a surprise exam cannot take place on it, even if the last day is reached. These are the works that allow for an exam not to take place as a way out of the paradox. We see that it is special in game forms 1 and 3, but not in game forms 2 and 4. Notice that even if it is indeed special, it does not mean that it can be ruled out.

The students' knowledge is special. Olin (1982) develops the concept that Sorensen (1984) later denotes an epistemic blindspot:

> A proposition p is an epistemic blindspot for a person a if and only if p is consistent while $K_a p$ (a knows p) is inconsistent.

After this distinction some authors argue that the students are forced to accept the teacher's statement and its conclusion as two contradictory statements from which they can deduce nothing. An external observer, however, sees no contradiction and anticipates a surprise. In a single day scenario, the observer sees that the exam will take place and that the students will not be able to deduce it (Wischik 1996). The game-theoretical approach shows that these concepts are unnecessary. In fact, most game theory, and our models in particular, are based on the assumption that the players are as intelligent as the modeller and can deduce everything the modeller can. Even with this strong and standard assumption about the students' knowledge the paradox is naturally resolved.

Students' memory. Some authors of the epistemological school have discussed whether to assume that one agent knows a proposition at time t should imply that he also knows it at time t+1, and whether this distinction could solve the paradox. In our framework we assume that the students have the same knowledge of the structure of the game at all times. The students and the teacher have perfect recall, again an implicit and standard assumption in most of the game-theoretical models.

The backward induction paradox 1. Some authors have suggested that this paradox is similar to the finitely repeated prisoners' dilemma. The argument is:

> Backward induction says that cooperation is impossible in the finitely repeated prisoner's dilemma, yet cooperation is observed (and viewed as more sensible than the non cooperative behavior), thus we should question the backwards induction argument. Introduction of irrational behavior with a small probability (irrational according to backwards induction) may bring the theoretical analysis closer to observed behavior and common sense. The same can be done in the surprise exam paradox. 11

It is true that the introduction of a small probability of being irrational may reconcile the logical analysis and the fact that the surprise exam is given, but this is not necessary, as we saw. The comparison with the prisoners' dilemma is not granted. In the prisoner's dilemma the paradox lies in the difference between the prediction of the theoretical model and the observed behaviour. The solution of the theoretical model by itself does not lead to a paradox, as in the case of the surprise exam. Chow (1998) makes the same point. Luc Bovens (1997) and Priest (2000) also provide a discussion on the similarities and differences between the finitely repeated prisoner's dilemma and the surprise exam paradox.

The backward induction paradox 2. In game theory there is a discussion about whether backward induction can be deduced from the rationality assumptions (e.g. Rosenthal 1981, and Ben-Porath 1997; in addition to the already mentioned Binmore 1987, and 1988). The question is:

> If when analyzing a particular subgame I reach the conclusion that I will never be there for I know that my opponent will choose actions that lead the course of the game in another direction, what should I conclude if I happen to be in that subgame? 12

As seen previously, when this situation appears in the surprise exam paradox, Smullyan concluded that he can no longer trust the teacher, and provides his resolution of the paradox after this fact. However, we see that these problems concerning backward induction have nothing to do with the paradox: since the last day must be reached with positive probability there is no discussion on how to interpret a non-expected action in the game. In other words, there is no need to solve the backward induction paradox in order to solve the surprise exam paradox.

Not necessarily an exam vs. not necessarily a surprise. In noticing that these two predicaments of the teacher's statement cannot be true at the same time, some authors (e.g. Alexander 1950; Quine, 1953) opt to explain the paradox by denying that there must be an exam, while others (e.g. Cohen 1950; criticized by Nerlich 1961) deny the necessity of the surprise. Our analysis shows that either way the paradox is easily explained. Moreover, it shows that it even if we do not regard the exam as necessary, still the surprise cannot be obtained with probability one. 13

Prediction before the week begins vs. prediction before the day begins. Ayer (1973) and Janaway (1989) discuss this distinction as relevant. Only in game forms 1 and 3 can students make the prediction before the last day, provided that they reach that day. Even then, this day is reached with a probability lower than one.

Prediction as an action vs. prediction as a logical deduction. Perhaps this is the conceptually most radical departure from the classical analysis of the paradox. Both the logical and the epistemological schools implicitly or explicitly assumed that the proper formalization of the word 'prediction' ought to be that the students should be able to logically deduce (or know) when the exam would take place. As the game-theoretical approach shows (after Sober 1998), the teacher does not select the day of

the exam deterministically based on her assertion, which implies that the students cannot logically deduce (with probability one) the day of the exam. Prediction as a logical deduction is, thus, an impossibility. The only way to make sense of the fact that students want to anticipate the exam and that the teacher does not want the students to anticipate it is in ranking preferences the way we did in the game. This is different from 'knowing' what the other party will do, which is discussed next.

Prudential vs. evidential prediction. Sober (1998) suggests this possible distinction. The prudential prediction is just the interpretation of the solution to the game when prediction is interpreted as actions. Since the probability distribution used by the teacher to select each day between E and NE may be different from the distribution used by the students to select between A and NA, the prudential interpretation has the awkward feature that the knowledge of the students is different from the teacher's action. Sober suggests a different analysis based on an alternative, the evidential prediction, that requires a game in which the teacher selects a probability distribution over the days to give an exam, and in which the students also select a probability distribution with the aim of guessing the teacher's. According to Sober, in equilibrium, both the teacher and the students will select the uniform distribution. However, this is not the case; in equilibrium, the teacher will select her distribution at random (in the two-day week, she will choose the probability of giving the exam on the first day to be p, and the value of p will be selected as the realization of a uniform distribution over values in the interval [0,1]). Students will do the same, and the probability of 'guessing' (the probability that both choose the same value) would be zero. Sober argues that even if the students chose the same distribution as the teacher, this still does not mean that the students are not surprised (to know the distribution does not mean to know with probability one the day of the exam, otherwise we may be forced to define surprise as a probability of anticipating the exam with a probability higher than a threshold). What we say is even stronger: that this game of anticipating the teacher's distribution (the evidential prediction) is quite pointless and has nothing to do with any reasonable interpretation of the paradox: to know the way in which the teacher chooses does not, *per se*, say anything about the surprise. However, if we wish to call an anticipation of the exam (the outcome (E,A)) that occurs with probability higher than a threshold (say higher than 0,5) a surprise, we can still do it in the standard game (Sober's prudential prediction). Next follows a related discussion.

Knowledge as an action vs. knowledge as a belief. In the subgame perfect equilibrium, the students must assign beliefs that correspond precisely to the updated probability distribution chosen by the professor. This is the students' knowledge. However, this knowledge does not preclude the surprise. This can be made clear with the following example. Suppose that the professor does not choose her equilibrium distribution and that, on a given day, she assigns a probability higher than in equilibrium (but not one) to, say, the action of giving the exam (E), and that the students know this. The students still do not know for sure that the exam will take place (belief), but they will choose A with probability one (action). We may or may not want to call this a surprise. In the first case the notion of surprise is linked to the notion of knowledge, in the second case it is not. In the equilibrium (the resolution of the paradox) there is no need to make the choice, as both beliefs and actions are non-degenerate probability distributions (except in the trivial and non-problematic last

day in game forms 3 and 4). Only if we insist on identifying knowledge with actions (and then, with the definition of surprise) we find a problem where there is none.

The surprise exam is an information-dependent game. This new class of game is defined in Gilboa and Schmeidler (1988). In addition to the set of players and of the strategy sets, it requires a set of possible predictions for each player, and a utility function that depends on both the strategies and the predictions. For a suitable definition of the term, they show that there is no informationally consistent way of playing the surprise exam game. We find this a very interesting formalization of the interaction between the students and the teacher, but it leaves a few issues in the air. If the teacher has no informationally consistent way of playing, what action should she take? A non-equilibrium situation provides no answer. In this context there is no clear way to interpret the fact that the exam is finally given and that the students are surprised after all. If mixed strategies were allowed, and the teacher had an informationally consistent way of playing, then the students' prediction would be about the probability distribution of this strategy, and we already argued that this does not correspond to the notion of surprise. Finally, our analysis shows that there is no need to develop a whole new class of games to deal with the paradox. Standard game theory and simple games are enough.

5 Conclusion

We have translated the surprise exam paradox into simple and standard games. By examining the elements of the models and their solutions we believe we have clarified many of the issues that have been, and still are, discussed about the paradox. It is interesting to note that, although the academic discussions of the paradox may have helped to develop new areas of research and new concepts in logic, the analysis of the paradox itself may not need them after all. On the other hand, our analysis shows that the study of the pragmatics of communication cannot be based exclusively on the assumption that language users are bound by a cooperative principle; what is really essential for their understanding of the utterances of others is their knowledge of the game to which those utterances refer.

Acknowledgements

Financial support from DGI grant SEJ2005-08633 (Ministerio de Educación y Ciencia) is gratefully acknowledged. The second author acknowledges financial aid through projects HUM2005-01686/FISO and HUM2005-25447-E.

References

Alexander, P. (1950), "Pragmatic Paradoxes," *Mind*, 59, 536–638.

Ayer, A.J. (1973), "On a Supposed Antinomy," *Mind*, 82, 125–126.

Ben-Porath, E. (1997), "Rationality, Nash Equilibrium and Backwards Induction in Perfect-Information Games," *Review of Economic Studies*, 64, 23–46.

Binmore, K. (1987), "Modelling Rational Players: Part I," *Economics and Philosophy*, 3, 179–214.

Binmore, K. (1988), "Modelling Rational Players: Part II," *Economics and Philosophy*, 4, 9–55.

Borwein, D., Borwein, J.M., and Maréchal, P. (2000), "Surprise Maximization," *American Mathematical Monthly*, 107, 527–537.

Bosch, J. (1972), "The Examination Paradox and Formal Prediction," *Logique et Analyse*, 15, 505–525.
Bovens, L. (1997), "The Backward Induction Argument for the Finite Iterated Prisoner's Dilemma and the Surprise Exam Paradox," *Analysis*, 57, 179–186.
Chow, T.Y. (1998), "The Surprise Examination or Unexpected Hanging Paradox," *American Mathematical Monthly*, 105, 41–51.
Cohen, L.J. (1959), "Mr O'Connor's Pragmatic Paradoxes," *Mind*, 59, 85–87.
Fitch, F. (1972), "A Goedelized Formulation of the Prediction Paradox," *American Philosophy Quarterly*, 1, 161–164.
Gilboa, I., and Schmeidler, D. (1988), "Information Dependent Games: Can Common Sense Be Common Knowledge?" *Economics Letters*, 27, 215–221.
Glazer, J., and Rubinstein, A. (2006), "A Study in the Pragmatics of Persuasion: A Game Theoretical Approach," *Theoretical Economics*, 1, 395–410.
Grice, H.P. (1975), "Logic and Conversation," in *Syntax and Semantics: Speech Acts* (Vol. 3). New York: Academic Press, pp. 41–58.
Janaway, C. (1989), "Knowing About Surprises: A Supposed Antinomy Revisited," *Mind*, 98, 391–410.
Lewis, D. (1969), *Convention: A Philosophical Study*, Cambridge, MA: Harvard University Press.
Lewis, D. (1973), *Counterfactuals*, London: Blackwell.
Mailath, G., Okuno-Fujiwara, M., and Postlewaite, A. (1993), "Belief-Based Refinements in Signalling Games," *Journal of Economic Theory*, 60, 241–276.
Meltzer, B., and Good, I.J. (1965), "Two Forms of the Prediction Paradox," *British Journal for the Philosophy of Science*, 16, 50–51.
Nerlich, G.C. (1961), "Unexpected Examinations and Unprovable Statements," *Mind*, 70, 503–513.
Niiniluoto, I. (1987), *Truthlikeness*, Dordrecht: D. Reidel.
Olin, D. (1983), "The Prediction Paradox Resolved," *Philosophy Studies*, 44, 225–233.
Olin, D. (1988), "Predictions, Intentions, and the Prisoner's Dilemma," *Philosophy Quarterly*, 38, 111–116.
Origgi, G. (2004), "Is Trust an Epistemological Notion?" *Episteme*, 1, 61–72.
Parikh, P. (2000), "Communication, Meaning, and Interpretation," *Linguistics and Philosophy*, 23, 185–212.
Priest, G. (2000), "The Logic of Backwards Induction," *Economics and Philosophy*, 16, 267–285.
Quine, W.V.O. (1953), "On a So-Called Paradox," *Mind*, 62, 65–67.
Rosenthal, R. (1981), "Games of Perfect Information, Predatory Pricing and the Chain-Store Paradox," *Journal of Economic Theory*, 4, 25–55.
Shaw, R. (1958), "The Paradox of the Unexpected Examination," *Mind*, 67, 382–384.
Smullyan, R. (1978), *Forever Undecided: A Puzzle Guide to Göedel*, New York: Knopf.
Sober, E. (1998), "To Give a Surprise Exam, Use Game Theory," *Synthese*, 115, 355–373.
Sorensen, R.A. (1984), "Conditional Blindspots and the Knowledge Squeeze: A Solution to the Prediction Paradox," *Australasian Journal of Philosophy*, 62, 126–135.
Sorensen, R.A. (1988), *Blindspots*, Oxford: Clarendon Press.
Windt, P. (1973), "The Liar in the Prediction Paradox," *American Philosophy Quarterly*, 10, 65–68.
Wischik, L. (1996), "The Paradox of the Surprise Examination," mimeo, Queens College.