tripleC 7(2): 194-201, 2009
ISSN 1726-670X
http://www.triple-c.at

Knowledge, Information and Surprise

Margarita Vázquez

Facultad de Filosofía. Universidad de La Laguna. Campus de Guajara s/n, 38201 La Laguna, Spain.
E-mail: mvazquez@ull.es. Web page: http:// webpages.ull.es/users/mvazquez

Abstract: *In this paper, I analyse the paradox called "The surprise exam paradox" or "The unexpected hanging paradox". I study some interpretations of this paradox, like Quine and Ned Hall ones, and give my own view about its solution, making some approaches from classical logic and from temporal or epistemic logic.*

Keywords: Knowledge, Information, Paradox, Surprise exam, logic.

Acknowledgement: This work was supported by Research Projects HUM2005-03848/FISIO and FF12008-01205 from the Spanish Government. This material was presented in the *I International Meeting of Experts in Theories of Information*, held in Leon, November 2008. I want to thank Paco Salto and José Mª Díaz Nafría for all their help.

In what follows, I will try to anlyze and give a solution to the surprise exam paradox, also known as the "unexpected hanging paradox"[1] or the "surprise exam paradox". My solution will follow, in a certain sense, Quine's proposal of solving this paradox trying to simplify it. In order to achive my goal I will use classical and temporal logic (in particular, ockhamist indeterminist temporal logic).

The surprise exam paradox was introduced in the early 1940s. At the beginning, it took the form of the "civil defence exercise paradox", although it has adopted many different versions[2]. John O'Connor published

[1] Martin Gadner explains, in the following way, this paradox in Gadner (2001).

The man was sentenced on Saturday. "The hanging will take place at noon," said the judge to the prisoner, "on one of the seven days of next week. But you will not know which day it is until you are so informed on the day of the hanging."

The judge was known to be a man who always kept his word. The prisoner, accompanied by his lawyer, went back to his cell. As soon as the two men were alone the lawyer broke into a grin. "Don't you see?" he exclaimed. "The judge's sentence cannot possibly be carried out."

"I don't see," said the prisoner.

"Let me explain. They obviously can't hang you next Saturday. Saturday is the last day of the week. On Friday afternoon, you would still be alive and you would know with absolute certainty that the hanging would be on Saturday. You would know this before you were told so on Saturday morning. That would violate the judge's decree."

"True," said the prisoner.

"Saturday, then is positively ruled out," continued the lawyer. "This leaves Friday as the last day they can hang you. But they can't hang you on Friday because by Thursday afternoon only two days would remain: Friday and Saturday. Since Saturday is not a possible day, the hanging would have to be on Friday. Your knowledge of that fact would violate the judge's decree again. So Friday is out. This leaves Thursday as the last possible day. But Thursday is out because is you're alive Wednesday afternoon, you'll know that Thursday is to be the day."

"I get it," said the prisoner, who was beginning to feel much better. "In exactly the same way I can rule out Wednesday, Tuesday, and Monday. That leaves only tomorrow. But they can't hang me tomorrow because I know it today!"

[2] A popular version is using numbered boxes. The following version is taken from Gerbrandy (1999):

A series of n numbered boxes is opened in sequence by the quiz master, starting from number 1, then number 2, etcetera. One of the boxes contains an enormous amount of money, and the quiz master knows which box it is. A player, b, gets the money if he knows, just before the box containing the money is opened, that this box is the one with the money in it. Player b is not allowed to guess;

a first version in *Mind* in 1948[3], and it was followed, at once, by many other papers in the same journal[4]. Since then, many solutions have been offered, analysing the paradox from many different points of view.

One of the most popular variations of this paradox is the known as "the surprise exam paradox". It goes as follows:

At the end of class one Friday afternoon, the professor announces to her students that she will give them an exam during one of next week's classes. (Class meets every day during the week.) She adds that the exam will be a surprise, in that the students won't expect, on the morning of exam day, that the exam will be that day. One of her cleverer students pipes up, saying that she cannot possibly fulfill her intention to give such an exam. "For it cannot be held on Friday: if it were, we would expect it on Friday morning (having noted that no exam had yet been given). So Friday is ruled out; the exam must take place on one of Monday through Thursday. But then, for exactly the same reason, it cannot be held on Thursday. But then, for exactly the same reason, it cannot be held on Thursday, else we would know that fact ahead of time (having noted that no exam had yet been given, and having ruled out Friday). And so on: It's really just a simple use of mathematical induction to show that your statement is inconsistent." The professor beams at her bright young student, and says nothing.

Arriving in class next Tuesday, the students discover that they are to take an exam that day. None of them, of course, expect it. The exam consists of one question: "What was wrong with the clever student's reasoning?" (Hall, 1999)

1. Three Questions

The paradox gives us several questions. The first one is if the surprise would be the same any day of the week: Is it the same surprise if the exam takes place on Monday, on Wednesday or on Friday? We can also ask if the paradox works independently of the number of days of the week: Would it be the same with only a day in the week? With 3 days in a week? With 7 days in a week? Would it be the same in a week 100 days long? And the last and most important question: Is there, in fact, a paradox?

1.1. First Question: Is it the same surprise if the exam takes place on Monday, on Wednesday or on Friday?

There is a stress between the confidence in the person that says the sentence and the reasoning. By the way of reasoning we arrive to the conclusion that the fact is not going to happen, because we trust in the person that

he must have a convincing argument that the money is in the box to be opened.

Suppose that the box with the money in it is somewhere in the middle –say there are five boxes, and the money is in the fourth. In that case, player b will never win the game, because at the moment that the fourth box is opened, he has no reason to assume that box 4 is not an empty box. Since the quiz master knows which box contains the money, she knows that b cannot win the game.

The 'paradox', now, is the following. Suppose the quiz master say to b: "You cannot win the game." As we have seen, this is true. Now b reasons as follows. "Suppose the money is in the last box. In that case, I would know that the money is in that box at the moment when all other box es were opened, and I would win the game. So, if the quiz master tells the truth, the last box is empty. But if this is true, the money cannot be in box 4 either, because I know (now) that the last box is empty, and so, if boxes 1 to 3 were opened, the money had to be in the fourth box, and I would win the game as well. I can repeat this proof for all boxes. But then I have to conclude that all the boxes are empty. This is in contradiction with what I know of the game. Therefore the quiz master must be lying to me."

[3] It was first published by John O'Connor (1948):

Consider the following case. The military commander of a certain camps announces on a Saturday evening that during the following week there will be a "Class A blackout". The date and the time of the exercise are not prescribed because a "Class A blackout" is defined in the announcement as an exercise which the participants cannot know is going to take place prior to 6.0 p.m. on the evening in which it occurs. It is easy to see that it follows that the exercise cannot take place at all. It cannot take place on Saturday because if it has not occurred on the first six days of the week it must occur on the last. And the fact that the participants can know this violates the condition which defines it. Similarly, because it cannot take place on Saturday, it cannot take place on Friday either, because when Saturday is eliminated Friday is the last available date and is, therefore, invalidated for the same reason as Saturday. And by similar arguments, Thursday, Wednesday, etc., back to Sunday are eliminated in turn, so that the exercise cannot take place at all.

[4] Cohen (1950), Alexander (1950), Scriven (1951), etc.

states it. However, if it is not possible to happen, it happens, and the person told the truth. No matter if we believe or not what the person states, the fact fulfils (or at least, it can fulfil). If we do not trust the person, the reasoning does not go. If we trust, the reasoning goes, but reality will contradict it.

There is a stress between the confidence in the person (the professor, the judge, the quiz master) that says the sentence and the reasoning, between the confidence and the surprise. Someone, who is wholly trustfully for us, say something important. We believe her. We have no doubts. We know she is not going to lie. This person says that we are going to have a surprise. If she says we are going to have a surprise, we must have a surprise. However, how can we have a surprise knowing that surprise?

Reasoning we arrive to the conclusion that the fact is not going to happen, because we trust in the person that states it. Nevertheless, if it is not possible to happen, if we are sure that it is not going to happen, it happens, and the person told the truth. We have the fact and we have the surprise. We have even a bigger surprise, because we were sure that it was not going to happen.

Imagine that the person is not trustfully, or alt least no wholly trustfully. In that case, we cannot make the reasoning. Here there is not stress between the confidence and the surprise. We can have a surprise because we do not know whether the person is lying or not. We are not sure if we are going to have an exam, the prisoner does not know if he is going to be hanged, etc. Therefore, if it happens, it will be something a little bit unexpected.

This is a strange situation. No matter if we believe or not what the person states, the fact fulfils (or at least, it can fulfil). If we do not trust the person, the reasoning does not go. If we trust, the reasoning goes, but reality will contradict it.

1.2. Second Question: Would it be the same with only a day in the week? With 3 days in a week? With 7 days in a week? Would it be the same in a week 100 days long?

Quine studies the paradox in the version of the unexpected hanging (Quine, 1953). He explains that the man arrives at the conclusion that the announcement is not going to be fulfilled. On Thursday he is hanged, when he thought that it was not possible. He was wrong in his argument that hanging should have been before Thursday. The problem is that, at time x, the man only saw two alternatives:

1) The event will have occurred at or before that time.
2) The event will occur at last chance, Friday, and the man will know it the day before.

The man rejects two, so he chose one.

For Quine, the man has not two, but four alternatives:

1) The event will have occurred at or before that time.
2) The event will occur at last chance, Friday, and the man will know it the day before.
3) The event does not happen last day (and violates the announcement).
4) The event happens last day and the man will remain ignorant, because he does not know if the announcement was going to be fulfilled or not.

Quine thinks that what is wrong with the argument of the man is that he does not see the possibility of three and four, and that one and four are compatible with the announcement.

To show more clearly his argument, Quine writes the story in a 1 day one that goes, more or less, as follows:

The judge tells the man on Sunday afternoon that he will be hanged the following noon and will remain ignorant of the fact till the intervening morning. It would be like the man to protest at this point that the judge was contradicting himself. And it would be like the hangman to intrude upon the man

complacency at 11:55 next morning, thus showing that the judge has said nothing more self-contradictory than the simple truth. If the man has reasoned correctly, Sunday afternoon, he would have reasoned as follows:

"We must distinguish four cases:
1) That I shall be hanged tomorrow noon and I know it now (but I do not).
2) That I shall be unhanged tomorrow noon and know it now (but I do not).
3) That I shall be unhanged tomorrow noon and do not know it now.
4) That I shall be hanged tomorrow noon and do not know it now."

The latter two alternatives are the open possibilities, and the last of all would fulfill the announcement. The man should have thought that better than charging the judge with self-contradiction, he should suspend judgment and hope for the best."

Quine clearly analyses this problem as a contingent future one, in the way of Aristotle's sea battle (*Peri Hermeneias*, chapter 9).

While Quine uses a week one day long to show his solution, Ned Hall (Hall, 1999) uses a "week" long enough (let us say, for example, a hundred days week) and the example of the student and an exam. For him the reasons that operate in the 1-day case are very different from the multi-day case. He says that the student is justified in believing the announcement in a week long enough. For him the problem is a problem of believes, and the student has a dilemma between the justification in a week long enough and a "confidence" principle that states if, at the outset, the student is justified in believing some proposition, then he is also justified in believing that he will continue to be justified in believing that proposition. So, the student is not justified in believing the announcement, regardless of the number of days in the week.

Ned Hall solves himself the dilemma by means of degrees of belief.

What I want to analyse from Ned Hall's paper is his criticism to Quine, which I think is wrong. Ned Hall focuses on belief and sees the problem of the paradox in the concept of belief, so he tries to solve it changing the concept of justified belief (for justified degree of belief).

Ned Hall shows different variants of the announcement:

1. The professor announces, not to the students, but to a colleague that she will set a surprise exam in the following week. Then the student cannot rule out Friday.
2. The student has heard a lot about this professor and knows that whenever she announces a surprise exam, she invariably gives it on the last available day. So, the student is not only justified in believing the announcement, he is also justified in believing it is false.

According to Ned Hall, Quine thinks that the student is not justified in believing the professor announcement. However, Quine does not speak about the student's justification. Quine talks about student's knowledge. In addition, what the student knows before the event happens is the space of possibilities. The problem for Ned Hall is a problem of beliefs. The problem for Quine is about events and future events.

It is not, as Ned Hall affirms, that in Quine diagnosis the student learns nothing relevant from the announcement. He learns something relevant. He learns that the exam can be one day of the week. Nevertheless, there are also other possibilities.

The different solutions given by Ned Hall and Quine keep relation with the solution to the first question, with the stress between the confindence and the surprise. While Quine focuses in the surprise, because the event can happen or not, Ned Hall prefers the confidence. If there is confidence, there is justified belief.

1.3. Third Question: Is there, in fact, a paradox?

In the so-called paradox there is a disjunction among the days of the week: "or it is the first day or it is the second day or … or it is the last day" (or Monday or Tuesday or Wednesday or Thursday or Friday). If it is Friday and I did not get an exam, that means that the exam took place neither on Monday, nor on Tuesday, nor on Wednesday, nor on Thursday, so my only solution is that the exam will be held on Friday. On Thursday the situation is completely other, because I do not have a negation on Friday and I still have two

options, it can be on Thursday or it can be on Friday. Therefore, the deduction that it must be on Thursday does not go.

In any case, the stress between confidence and surprise only has whole strength the last day of the week. There is a disjunction among the days of the week: “or it is the first day or it is the second day or … or it is the last day” (or Monday or Tuesday or Wednesday or Thursday or Friday), or among the ten boxes.

When we arrive to the last day raises the problem of the confidence in the person that states the sentence, if we trust the person we know the fact is going to happen this day. However, what happen the other days?

- If it is Friday and I did not get an exam, that means that the exam took place neither on Monday, nor on Tuesday, nor on Wednesday, nor on Thursday (no Monday and no Tuesday and no Wednesday and no Thursday), so, by classical propositional logic, my only solution is that the exam will be held on Friday.
  I know in advance that
  $M \vee T \vee W \vee Th \vee F$
  and I know now (Friday)
  $\neg M \wedge \neg T \wedge \neg W \wedge \neg Th$,
  Therefore, my only possible conclusion is
  F.
- On Thursday the situation is completely other, because I do not have a negation on Friday and I still have two options, it can be on Thursday or it can be on Friday. Therefore, the deduction that it must be on Thursday does not go.
  I know in advance that
  $M \vee T \vee W \vee Th \vee F$
  and I know now (Thursday)
  $\neg M \wedge \neg T \wedge \neg W$,
  Therefore, my conclusion is
  $Th \vee F$.
- On Wednesday the situation is similar to Thursday, because I do not have negations on Friday and Thursday, I have three options, it can be on Wednesday, it can be on Thursday or it can be on Friday. Therefore, the deduction that it must be on Wednesday does not go.
  I know in advance that
  $M \vee T \vee W \vee Th \vee F$
  and I know now (Wednesday)
  $\neg M \wedge \neg T$,
  Therefore, my conclusion is
  $W \vee Th \vee F$.
- And the same thing with Tuesday and Monday.

To deduce that the day the exam will take place in Thursday, I should know that on Friday the exam did not take place. I do not know that Friday was not the day if it is still Wednesday afternoon or Thursday morning. I know that Friday could be the day if Thursday is not, but, at this moment (Wednesday afternoon or Thursday morning); I do not know if Thursday was or not, because there are still chances that we can have the exam.

At the same time, it seems quite clear that this paradox, from another point of view, is a case of contingent futures (we are not sure if something is going or not to happen tomorrow, so its truth value is in question). We can find the genesis of this problem in Aristotle, in the chapter nine of the *Peri Hermeneias*. In this chapter, Aristotle thinking about a possible sea battle tomorrow raises the problem of the contingent future and its truth-value. In this century, in the sixties, Arthur Prior tries to formalize this kind of time (following an Ockhamist view), from a philosophical point of view. The system, called OT, is logically and philosophically very interesting and has been developed by several logicians (Burgess, Thomason, and Zanardo). At the same time, computer scientists have developed a similar system called CTL, from very different motivations.

Contingent futures have been analysed also with degrees or multivalues, but contingent futures are solved usually with branching time logics. I can manage different alternatives: in the first one the exam will take place on Monday, in the second on Tuesday, The truth value of the sentences is only relative to the alternative (or branch) that, at the end, comes to place.

We cannot travel through time in the two directions. If it is Thursday, it means that it is not Friday; we do not know what is going to happen. Temporal logic can help us to analyse the paradox, but it suffices classical logic to see gaps in the reasoning. To see that the professor is not going to give the exam on

Friday, we have to be on Friday. Before Friday it does not work.

If Friday the exam does not take place the only problem is with the necessity, “it is necessary that the exam…” and, as Quine would say, the students would have forgotten that there were other possibilities, other alternative futures.

2. And now... What else?

Although I think that the paradox dissolves, it is so suggesting that we can analyze it in different ways. I have not found any analysis using temporal-epistemic logic, but there are formalizations using epistemic logic and dynamic epistemic logic.

2.1. Epistemic Dynamic Logic

Dynamic logic is epistemic logic extended with actions. Jell Gerbrandy, in his dissertation (Gerbrandy, 1999), makes a very interesting dynamic-epistemic analysis of the surprise examination paradox. This is the first analysis of this paradox with dynamic epistemic logic. He studies the announcement "On the basis of what you know, you cannot win the game". He tries to argue that dynamic epistemic semantics offers a natural analysis of how one can learn that such sentences are true while still not coming to believe them. One can even learn such a sentence and come to believe the contrary of the sentence. The act of uttering such a sentence successfully may change the situation in such a way that the sentence becomes false.

Hans van Ditmarsch and Barteld Kooi, in a recent paper (Ditmarsch & Kooi, 2005), study the role of unsuccessful updates in logical puzzles. They see two utterances of the teacher:

1) "There will be an exam next week"
2) "The exact day of the exam will be a surprise"

They say, following Gerbrandy, that the first sentence is an exclusive disjunction over the possible days. They distinguish two readings in the second one:

1) *Given the information the students now have*, the students will not know the day of the exam in advance.
2) The students will not know the day of the exam in advance, even *after they hear this announcement*.

They think that the first sentence can be formalized with the logic of public updates, but the second one not, because it involves self-reference. When "surprise" is announced, the students only learn the exam will take place on another day than Friday. They say the *reductio ad absurdum* cannot go any further, because the announcement is not successful. The problem remains with the reading of the second sentence, as "The students will not know the day of the exam, *even after they hear this announcement*". This announcement is self-referential and relates the paradox with the liar one.

2.2. Hybrid and Epistemic Temporal Logic

It could be interesting to give an account of the way in which the knowledge or the beliefs of the agent change along the time; and it could be also interesting to express the knowledge or the beliefs on an agent, in a concrete instant, with regard to the past and the future. It seems that this paradox is a case of contingent futures, which branching time logics usually solve. The truth-value of the sentences is only relative to the branch that, at the end, comes to place.

In the articles about the topic, I have not found a research using indeterministic temporal logic or temporal-epistemic logic, although there are formalizations using epistemic logic and dynamic epistemic logic[5]. This is surprising when many authors talk about time (or temporal series). It could be interesting to give an account of the way in which the knowledge (or the beliefs) of the agent change along the time and it could be also interesting to express the knowledge or the beliefs on an agent in a concrete instant with regard to the past and the future. In other papers, Rafael Herrera and I have explored this topic.

Among the systems proposed to combine time and knowledge, the most important are:

[5] There have been some interesting formalizations of the paradox using dynamic logic.

the Engelfriet (1996) minimal temporal-epistemic system, the temporal extension of Kraus & Lehmann (1988) system for knowledge and belief and Fagin & Halpern (1988) temporal-epistemic logic.

Engelfriet system combines S5 epistemic system with a linear transitive time one. The problem is that it does not allow the occurrence of temporal operators under the range of epistemic operators. This reduces the system to a temporalitation of epistemic logic. Kraus and Lehmann introduce a system that combines several epistemic and doxastic notions. They enrich this system with temporal operators to express changes of knowledge and belief along the time and different kinds of beliefs that the agents can have about the future. The basic ideas of Fagin and Halpern allow, up to a great extent, to solve the problem of the logical omniscience.

The main difficulties in order to combine temporal and epistemic logic come from the fact of having to combine an absolute temporal perspective with a relative (to each agent) epistemic perspective. That is:

1) Temporal points (instants) are determined from the point of view of an observer placed outside the world, and
2) The epistemic alternatives of each agent (in each instant) are relative to that agent.

Hybrid logics are modal logics that allow referring to the points in the model. In the case of temporal logic, they allow to refer to a particular point of time, an instant. The principal ideas related with hybrid logics were introduced by Prior (1967), after him, it has developed by Bull and reinvented by a group of logicians from the Sofia School. In the 1990s, the research papers about this topic increased, and the principal authors are Blackburn, Areces and other researchers linked to the University of Amsterdam (Areces, Blackburn & Marx, 2001; Blackburn, 2000; Blackburn & Seligman, 1998).

Our hypothesis, in other place, is that hybrid logics simplify the combination of temporal and epistemic logics. Nominals, as used is hybrid logic, allow making reference to points, so we can make reference to concrete instants or to some states in the present, past or future. Doing so, we can avoid building highly complicated models. Instead of that, we build our models upon the notion of "state", where a state is a possible world in a concrete instant. We have two accessibility relations among states, an equivalence relation and an irreflexive partial order. The valuations are similar to those of Engelfriet plus the ones for some new operators we introduce, which include hybrid logic nominals. If an agent knows something (A) about the future in the current state (that is, an agent knows that something will be the case in the future i), we have to check that every state (that the agent considers possible) at the current instant have an ulterior state that is the denotation of i where A is true.

The study of the paradox with the new hybrid-temporal-epistemic logic could offer some new approaches.

References

Alexander, P. (1950). Pragmatic Paradoxes. *Mind, 59*, 536-538.

Areces, C., Blackburn, P. Y Marx, M. (2001). Hybrid Logics: Characterization, Interpolation And Complexity. *The Journal Of Symbolic Logic*, *66(3)*, 977-1009.

Blackburn, P. (2001). Representation, Reasoning, And Relational Structures: A Hybrid Logic Manifesto. *Logic Journal Of The IGPL*, *8(3)*, 339-365.

Cohen, L.J. (1950). Mr. O'connor's Pragmatic Paradoxes. *Mind*, *59*, 85-87.

Engelfriet, J. (1996). Minimal Temporal Epistemic Logic. *Notre Dame Journal Of Formal Logic*, *37*, N. 2.

Fagin, R. Y Halpern, J.H. (1988). Belief, Awareness And Limited Reasoning. *Artificial Intelligence, 34*.

Gabbay, D., Hodkinson, I. Y Reynolds, M. (1994). *Temporal Logic*. Oxford: Oxford University Press.

Gadner, M. (2001). *The Colossal Book Of Mathematics: Classic Puzzles, Paradoxes And Problems*. New York / London: W.W. Norton & Company Ltd.

Gerbrandy, J. (1999). *Bisimulations On Planet Kripke.* Amsterdam: Doctoral Dissertation.

Hall, N. (1999). How To Set A Surprise Exam. *Mind, 108*, 647-703.

Herrera, R. Y Vázquez, M. (2003). Combining Temporal And Epistemic Logic With The Help Of Hybrid Logic. *Twenty First World Congress Of Philosophy*, Estambul.

Herrera, R. Y Vázquez, M. (2005). Towards An Hybrid Epistemic Linear Temporal Logic. *Fifth European Congress For Analytic Philosophy*, Lisbon.

Kraus, S. Y Lehmann, D. (1988). Knowledge, Belief And Time. *Theoretical Computer Science*, 58.

O'Connor, D.J. (1948). Pragmatic Paradoxes. *Mind, 57*, 358-359.

Prior, A. (1967). *Past, Present And Future*. Oxford: Oxford University Press.

Quine, W.O. (1953). On A So-Called Paradox. *Mind*, *67*, 403-407.

Scriven, M. (1951). Paradoxical Announcements. *Mind*, *60*, 403-407.

Van Ditmarsch, H. Y Kooi, B. (2005). The Secret Of My Success. *Synthese, 151(2)*, 201-232.

About the Author

Margarita Vázquez

Assistant Professor of Philosophy (Logic and Philosophy of Science) at the University of La Laguna (Spain). Member of the *Spanish Association for Analytical Philosophy (*She is a member of the Policy Council), of the *Spanish Society of Logic and Philosophy of Science* and of the *System Dynamics Society*. Her research interests are very broad: Philosophy of Technology, Modeling and Simulation, Bounded Rationality, System Dynamics, Paradoxes, Logic, Philosophy of Logic, Temporal, Epistemic and Hybrid Logic, Semantics. She is author of many chapters of books and articles, in spahish and English, about these topics. She is author of several Logic books in Spanish.