# The Surprise Examination Paradox in Dynamic Epistemic Logic

**MSc Thesis** (*Afstudeerscriptie*)

written by

**Alexandru Marcoci**
(born December 19th, 1986 in Bucharest, Romania)

under the supervision of **Prof. Dr. Johan van Benthem** and **Dr. Sonja Smets**, and submitted to the Board of Examiners in partial fulfillment of the requirements for the degree of

**MSc in Logic**

at the *Universiteit van Amsterdam*.

| **Date of the public defense:** | **Members of the Thesis Committee:** |
|---|---|
| *December 14, 2010* | Prof. Dr. Johan van Benthem |
| | Dr. Sonja Smets |
| | Prof. Dr. Peter van Emde Boas |
| | Prof. Dr. Dick De Jongh |
| | Prof. Dr. Frank Veltman |

Institute for Logic, Language and Computation

# Acknowledgements

I wish to express my gratitude to my supervisors, Johan van Benthem and Sonja Smets. They have been excellent guides both through the technical intricacies and the philosophical subtleties that I encountered during my research. I would particularly like to thank Johan for badgering me to always keep in mind the big picture and to Sonja for her day-to-day supervision and careful scrutiny of my every argument. Together, they offered me the impetus and the support I needed in order to successfully complete this project, while leaving me enough space to develop in my own direction. Thank you!

I wish to express my gratitude to the members of the committee for the time they invested in reading this thesis and then in devising questions and in stimulating me to think deeper about the problems that it addresses.

I wish to express my gratitude to the ILLC for offering the perfect environment for studying logic. It is through the mix of classes and conferences organized here that I first got acquainted with dynamic epistemic logic and gained the necessary skills for completing this thesis. I wish to thank my teachers for all they have taught me and, of course, Tanja, who is the ILLC's "invisible hand".

This thesis would have looked very different had it not been for the project on dynamic epistemic logic organized by Johan and Davide Grossi. During that project I had my first go at weaving together dynamic epistemic logic and philosophy and Johan and Davide joyfully guided these first steps. Also, I had a lot to gain from the kind invitation of Johan and Eric Pacuit to present an early version of my thesis during the Workshop on Logic, Rationality and Interaction at ESSLLI 2010. I wish to thank the audience for all their questions and particularly Alexandru Baltag whose challenging comments helped me tune my arguments better.

I wish to express my gratitude to my friends from Amsterdam who, through their kindness and humor made my stay here an unforgettable experience: to Fernando for his contagious optimism; to Kian for the stimulating conversations (philosophical or otherwise) and for the movie nights that we enjoyed together; to Lorenz, for the seminar we organized together and for listening and helping me sharpen many of the ideas that are now in this thesis; to Lucian and Roxana for their kindness to have me over so many times in my first year here and for making Amsterdam feel so close to

home; and to Michael for his humor and his amazing dinner&movie parties, without which my second year in Amsterdam would have been much emptier. Also, I thank my friends who have not been here with me in the past years but who remained as close as always: Andrei, Cristi, Cristiana, Matei, Mihai, and Sebastian.

I wish to express my gratitude to my family who has always been extremely supportive of my every endeavor, academic or otherwise, and for which these last years have not been the easiest: vă mulţumesc că mi-aţi dat atât sfaturile cât şi libertatea necesare pentru a descoperi ceea ce îmi place să fac şi, mai apoi, pentru a reuşi să duc la bun sfârşit acest proiect. Fără voi nu aş fi ajuns aici!

Finally, nothing would have been the same without Diana, who has been at my side from before this thesis started to germinate to the early morning when I finally submitted it and beyond. She has read almost all versions of this thesis and patiently and carefully weighted every argument saving me more than once from unclarity and error. During the last year, she has been immensely understanding and supportive when the writing of this thesis overwhelmed me. She constantly provoked me to better myself and she remains my source of inspiration and comfort.

This thesis was made possible by the generous support of a Huygens scholarship which I hereby acknowledge.

HAARLEM: December, 2010

# Contents

CHAPTER 1

# Introduction

## 1. The surprise examination paradox

> A teacher announces that there will be a surprise examination the following week. A student objects that this is impossible: "The class meets on Monday, Wednesday, and Friday. If the examination is given on Friday, then on Thursday I would be able to predict that the examination is on Friday. It would not be a surprise. Can the examination be given on Wednesday? No, because on Tuesday I would be in a position to predict that the examination will not be on Friday (thanks to the previous reasoning) and know that the examination was not on Monday (thanks to memory). Therefore, on Tuesday I could foresee that the examination will be on Wednesday. An examination on Wednesday would not be a surprise. Could the surprise examination be on Monday? On Sunday, the previous two eliminations would be available to me. Consequently, I would be in a position to predict that the examination must be on Monday. So a Monday examination would also fail to be a surprise. Therefore, it is impossible for there to be a surprise examination." [1]

In most instances, the presentation of this scenario is followed by saying that the teacher gives the students an examination on one of the available days, say Wednesday (although he could give it on any other day as well). Given the above reasoning of the students, they will be in no position to predict that an examination will occur on Wednesday (as they predict that no surprise examination will occur), and hence they will indeed be surprised. The paradoxical flavor of this story comes in when we reflect on what happens when adopting the students' perspective: they begin by assuming that there will be a surprise examination, they deduce that such a surprise examination cannot take place, and because of their deduction they are indeed surprised when the examination actually comes. What is more, even if they have access to the fact that the surprise is precisely the result of their reasoning, they still cannot help being surprised: if they know that by reaching the conclusion that no

[1] Formulation adapted from Sorensen 2009 (48)

surprise examination can occur, they put themselves in a position in which they can indeed be surprised, then they can deduce that a surprise examination could occur in the following week. However, they now restart the backward elimination reaching, again, the conclusion that a surprise examination cannot occur. This makes them once more susceptible to being surprised by an examination and so on. Therefore, thinking that there can be a surprise examination leads to thinking that there cannot be a surprise examination; thinking that there cannot be a surprise examination leads to (thinking that) a surprise examination being (is) possible!

Regarding the origins of the surprise examination paradox[2] it has allegedly been discovered by Lennart Ekbom, a Swedish mathematician. Apparently, during World War II, the Swedish Civil Defense broadcasted the following announcement: "A civil-defense exercise will be held this week. In order to make sure that the civil-defense units are properly prepared, no one will know in advance on what day this exercise will take place." Ekbom, realizing the paradoxical nature of this announcement, began using the announcement as an exercise in his classes, and from there it is supposed to have been picked off by philosophers. O'Connor, the first to write about the paradox in a Philosophy journal, first heard it in 1946 in Chicago during a discussion with Arthur Pap. Quine, one of the early respondents to O'Connor's article, claims to have learned the paradox around 1943 while he was a naval officer in Washington. His impression was that Tarski began circulating it (at least in the United States). This story of how the paradox of the surprise examination came about is told in Bunch 1982 (12), p. 34 and in Sorensen 1988 (47). Sorensen also claims to have independent evidence from the actual correspondence between M. Gardner and L. Ekbom that followed the publication of Gardner's *Scientific American* article on the paradox (Sorensen 1988 (47), p. 253) and from personal correspondence both with O'Connor and Quine (Sorensen 1988, p. 262).

The surprise examination paradox has a long history not only in terms of the time that has passed since it first appeared in print (O'Connor 1948 (35)) but also in terms of the literature that was dedicated to solving it. In the most comprehensive history of this paradox, Sorensen 1988 (47) surveys approximately 30 different pieces that deal with it, while in a recent article, Levy 2009 (29) refers to 52 different articles and book chapters, but without exhausting all that has been written. The reason for such a rich (and still expanding literature) is that there is much debate not only with respect to what the correct solution is, but also with respect to *what a solution*

[2]In this thesis "paradox" should be construed as a puzzle which appears paradoxical (at least) at the level of our intuitions (and possible only there). In other words, when I call the surprise examination paradox a "paradox" I am not committed to the stance (embraced by Kaplan and Montague (28), among others) that it is unsolvable, i.e. that there is a real contradiction in the way in which the scenario is set up. I merely want to suggest that pre-theoretically the scenario presented in the puzzle seems paradoxical.

*has to solve*. There are three fundamental questions that any solution to the surprise examination paradox has to address: (1) is it really a paradox? (2) what exactly is the paradox? and (3) how should the teacher's announcement be construed and what is the exact definition of some of the key components like "prediction" and "surprise"? The way in which one answers these questions shapes his/her solution to the surprise examination paradox. For instance, if one believes that the surprise examination paradox is indeed a paradox, then the solution needs to look differently than in case he/she thinks the paradox is only apparent. I will return to this later this chapter, in section 4 .

In order to ease the understanding of what I will be saying in this thesis, I will need to define what it means for a scenario to be a variation of the surprise examination paradox. First, let an *interpretation* of the surprise examination paradox refer to how we define the terms "prediction" and "surprise" [3], and to how we construe the teacher's announcement. Then, a *variation* of the paradox refers to how we modify the story while keeping an interpretation fixed. For example, there is a debate whether surprise is a knowledge-based phenomenon or a belief-based phenomenon; so, the scenario presented at the beginning of this chapter, where we substitute surprise with a knowledge-based definition (e.g. "the day before the exam will be given you will not know that the exam will be given the folowing day") and the same scenario but in which we substitute surprise with a belief-based definition (e.g. the day before the exam will be given you will not believe that the exam will be given the next day") are two different interpretations of the paradox. However, the scenario presented above and a scenario like the above but in which the students have only one class every week with the professor that made the announcement (see the 1-day surprise examination paradox, *below*) are different variations of the surprise examination paradox (a list of the most important variations of the surprise examination paradox can be found *below*).

I would like to note briefly why the scenario with which I started this chapter is so vague. It is because the scenario is meant to satisfy most interpretations that are being entertained in the literature on the surprise examination paradox. It is meant so as not to commit us to any definition of "predicting", "surprise", or to how teacher's announcement should be construed. But it does commit us to how the students received the information that they will be surprised, to the number of possible examination days and to the order in which the examination days are to be eliminated, among others. So it is only one variation of the paradox among many.

[3]Rigorously, the definitions of "prediction" and "surprise" should be separated. However, it will always be very easy to read off the definition of prediction from the definition of surprise, and hence, I will only be concerned with defining the latter.

The rest of this chapter is organized as follows: (i) I will present the most well-known variations that have been proposed so far in the literature on the surprise examination paradox; (ii) I will briefly survey some of the most well-known solutions to the surprise examination paradox; and (iii) I will explain what I believe solving the surprise examination means, and what the motivation for such an enterprise is.

## 2. Variations of the surprise examination paradox

The purpose of the variations which will be presented in this section is that of showing that the sort of puzzlement we feel when presented with the surprise examination paradox[4] we also feel when presented with these variations. The way in which these variations came about and in which they are used is as counter-arguments to one interpretation or another. The line of reasoning is usually the following: the interpretation to the surprise examination paradox offered by X seems to be well suited for the variation of the surprise examination paradox he analyzes, however, variation Y seems to be manifesting the same type of problem, but X's interpretation does not apply to Y. I will give concrete examples in the next section.

The literature on the surprise examination paradox is extensive and collecting all the variations proposed would be a very difficult task. I will not do this here, but instead I will present just the variations that I believe carry some interesting consequences. Consequently, example 1[5] suggests that the seemingly paradoxical nature of the surprise examination paradox does not depend on the number of classes the students have with the teacher who made the announcement . Example 2 suggests that it does not depend on the assumption that there has to be an examination. Example 3 suggests that it does not depend on the fact that time passes between the teacher's announcement and the examination. Example 4 suggests that it does not depend on the order in which the days are eliminated. Example 5 suggests that it does not depend on the fact that the teacher is making the announcement. Example 6 suggests that it does not depend on the fact that the students can doubt the existence of an exam. Example 7 suggests that it does not depend on the mental attitude the students have.

Reading the different variations presented here, which, again, are just a few, a question can be raised - all these variations seem to rely on the same underlying paradox, but nevertheless they also seem to differ in some respects. Hence the question is: how to separate the general problem that they all instantiate from the differences that obviously exist between them? So far, the answer to this question has been left

[4]From now on, whenever I talk about the surprise examination paradox, I will be referring to the formulation given at the beginning of this chapter. If I refer to another formulation, I will either refer to it by its name, see below, or I will just present the exact scenario I refer to.

[5]All the examples mentioned in this paragraph are presented *below*.

to intuitions, and this is why the literature on the surprise examination paradox is divided into two groups. On the one hand, some (very few) authors who feel that the differences are larger than the similarities, claim that a solution to the surprise examination paradox is sufficiently good if it can solve just one variation. The fact that these authors do not investigate the exact relationship between these variations is not problematic for them, since they think their solutions should only apply to one variation anyway. On the other hand, there are authors who feel that the similarities between all the variations are more relevant than the differences. Nevertheless, even they do not say too much about the relationship between the different variations [6]. However, in their case, the fact that the relationship between the different variations has been left at the level of our intuitions is truly problematic. They should try to explicitate the exact paradox that manifests itself through all these variations. Otherwise, new recalcitrant variations to their proposed solutions are very likely to appear. The literature on the surprise examination paradox is full of such examples. So, one other reason for not collecting in this section all the variations of the surprise examination paradox is that since it is yet unclear what exactly the surprise examination is, collecting all the variations discussed so far cannot give the only significant purpose that could be expected to give, namely exhausting the different guises in which the problem could appear. New and surprising variations could always be generated. So, although each variation has merit in its own right(usually that of showing that some element of the scenario is not necessary for having the feeling of puzzlement prompted by the surprise examination paradox), going through all of them does not give us more.

The purpose of this section is to familiarize the reader with the history of the surprise examination paradox, to ease the understanding of some of the interpretations that I will present in the next section, and to actually give a feeling of the sort of counter arguments that philosophers bring to some of the interpretations to the surprise examination paradox proposed so far, as I will give such a counterargument later in this thesis myself.

This section is structured as follows: first the variation is presented (some of the variations have been dubbed in different ways, so when it applies I will indicate its name); secondly the reasoning of the students is described and lastly some (not exhaustive) bibliographical sources are indicated. All the variations below end with the teacher giving an examination that surprises the students, or the equivalent.

[6] There are two exceptions here, namely McLelland and Chihara 1975 (32) and Sorensen 1988 (47). The former contains a formulation of the paradox that is aimed at being universal, while the latter tries to argue that the surprise examination paradox and all its variations are just instantiations of slippery slopes. However, even Sorensen admits that, at least, the relation between examples 2 and 7 and the rest are truly problematic.

EXAMPLE 1 (1-day surprise examination ). *A teacher announces that there will be a surprise examination the following week. A student objects that this is impossible: "The class only meets on Monday. Could the surprise examination be on Monday? Consequently, on Sunday I would be in a position to predict that the examination must be on Monday. So a Monday examination would fail to be a surprise. Therefore, it is impossible for there to be a surprise examination."*

**Bibliographical Sources:** Quine 1953 (42), p. 66, Binkley 1968 (8), p. 127, Baltag 2003 (3), slide 143.

EXAMPLE 2 (Conditional surprise examination). *A teacher announces that if there is going to be an examination in the following week, it will be a surprise examination. A student objects that this is impossible: "The class meets on Monday, Wednesday, and Friday. If there is an examination and it is given on Friday, then on Thursday I would be able to predict that if there is going to be an examination at all, it is going to be on Friday. It would not be a surprise. Can the examination be given on Wednesday? No, because on Tuesday I would be in a position to predict that if there is going to be an examination, it will not be on Friday (thanks to the previous reasoning) and know that the examination was not on Monday (thanks to memory). Therefore, on Tuesday I could foresee that if there is going to be an examination at all, it will be on Wednesday. An examination on Wednesday would not be a surprise. Could the surprise examination be on Monday? On Sunday, the previous two eliminations would be available to me. Consequently, I would be in a position to predict that if there is going to be an examination at all, is must be on Monday. So a Monday examination would also fail to be a surprise. Therefore, it is impossible for there to be a surprise examination."*

**Bibliographical Sources:** Sorensen 1988 (47), p. 337, Williamson 2000 (58), p. 144.

EXAMPLE 3 (Designated student). *Robinson Crusoe discovers four people on his island in addition to Friday and names the rest of them for each day of the working week. Crusoe decides to teach them English history. Since his resources are limited he can only give one student a test. Since he wants the test to be a surprise he first lines the students up in accordance with the order of the days in the week so that Friday can see the back of Thursday and the backs of all those in front of Thursday, and Thursday can see the backs of Wednesday, Tuesday and Monday (but not Friday's since Friday is behind him), and so on. Robinson Crusoe then shows the students four silver stars and a gold star. He announces that he will put a gold star on the back of the student who has to take the examination and silver stars on all the rest and that the examination will come as a surprise.*

*One of the students objects that such an examination is impossible: "we all know that Friday cannot be the designated student since if he were he would see the four silver stars in front of him and then be in a position to predict that he must have the gold star on his back. But then he would know that he was the designated student. So Friday would not be surprised. We all know that Thursday cannot be the designated student since, if he were, he would see silver stars in front of him and since he knows by the previous deduction that Friday is not the designated student he will be in a position to predict that he is the designated student. So Thursday would not be surprised. In a similar manner, Wednesday, Tuesday, and Monday can be eliminated. Therefore the examination is impossible."*

**Bibliographical Source:** Sorensen 1988 (47), p. 317-318.

EXAMPLE 4 (The random surprise examination paradox). *Consider the game played in the table below.*

| | | |
|---|---|---|
| *1* | *2* | *3* |
| *4* | *5* | *6* |
| *7* | *8* | *9* |

*The object of the game is to discover where you have been initially placed. The seeker may only move Up, Down, Left, or Right, one box at a time. The outer edges are called walls. If the seeker bumps into a wall, say by moving left from* 1*, his move is recorded as* $L^*$ *and his position is unchanged. Bumps help the seeker discover his initial position. For instance, if he is at* 7 *and moves* $U, U, L^*$*, the seeker can deduce that he must have started from* 7*. The seeker has discovered where he started from if he obtains a completely disambiguating sequence of moves, i.e. a sequence which determines the seeker's initial position. Someone with knowledge of the seeker's position announces to the seeker "You have been put in an undiscoverable position." The seeker objects that there is no undiscoverable position: "suppose I am in an undiscoverable position. It follows that I cannot be in any of the corners since each has a completely disambiguating sequence. For instance, if I am in* 3*, I might move* $U^*$*,* $R^*$*, and thereby deduce my position. Having eliminated the corners, I can also eliminate 2, 4, 6, and 8, since any bumps resulting from a first move completely disambiguates. For instance* $U^*$ *is sufficient to show I am in 2. Therefore only 5 remains, but this would mean that I have discovered my position. Therefore, there cannot be any undiscoverable positions."*

**Bibliographical Source:** Sorensen 1988 (47), p. 320-321.

EXAMPLE 5. *A teacher's pupils know that he rings all and only examination dates on the calendar in his office. At the beginning of term, the only knowledge they have of examination dates this term comes from a distant glimpse of the calendar, enough to see that one and only one date is ringed and that it is not very near the end of*

*term, but not enough to narrow it down much more than that. The pupils recognize their situation. They know now that for all numbers $i$, if the examination is $i+1$ days from the end of term then they will not be in a position to predict that it will not be $i$ days from the end $(0 \leq i < n)$. In particular, they know now that if it is on the penultimate day then they will not be in a position to predict now that it will not be on the last day. But they also know now from their glimpse of the calendar that it will not be on the last day. They deduce that it will not be on the penultimate day. They also know now that if it is on the antepenultimate day then they will not be in a position to predict that it will not be on the penultimate day. They deduce that it will not be on the antepenultimate day. And so on. They rule out every day of term as a possible date for the examination.*

**Bibliographical Source:** Williamson 2000(58), p. 135.

EXAMPLE 6 (Guessing the card). *Someone announces that he has placed the ace of spades in a position in the deck such that as the cards are turned over one by one, the audience will not be able to predict when the ace of spades will be revealed. The audience is permitted to verify the normality of the deck and knows that the dealer has no talent as a magician. Somebody in the audience objects that there is no such position: if the ace of spades is the last card in the deck, then after all but one card have been turned over, he will be in a position to predict that the last card is the ace of spades. Hence, the ace of spades cannot be the last card. Also if the ace of spades is the penultimate card, then after all card but two have been turned over, he will be in a position to predict that the ace of spades can be either the last or the penultimate card. However, if it is the last, the previous reasoning applies, and hence it can only be the penultimate card. But then he will be in a position to predict that it is the penultimate card before it is turned over, so it cannot be the penultimate card either. Similarly, that person in the audience concludes that the ace of the spades is not in the deck, which is an unacceptable conclusion*

**Bibliographical Sources:** Ayer 1973 (1), p. 125-126, Sorensen 1988 (47), p. 311.

I believe that there are reasons for arguing that the next scenario is not part of the family of variations to the surprise examination paradox. However, I include it here to give you a feeling of how wide-ranging the problem with the surprise examination is sometimes taken to be (I will say more about this later in a different context).

EXAMPLE 7 (Intention-based surprise examination paradox). *A wealthy teacher makes the following offer to his students. Beginning next week, the students will be paid $ 1,000 if at midnight they intend to take an unpleasant test the following Monday afternoon. What is more, if the students carry out their intentions and do*

*take the test they earn an opportunity to have the offer renewed for the following day. However, there is a maximum of four renewals. It is commonly understood that students hate tests but love money. At a first glance, the students stand to make $ 5,000. However, they then realize that they have no reason to carry out their intentions of taking the exam on Friday afternoon, since by doing so they win nothing. But then, they cannot form the corresponding intention on Thursday evening. So, the offer to win $ 1,000 by taking an unpleasant exam on Friday is worthless, and hence there is no reason for them to take an unpleasant exam on Thursday afternoon. This reasoning is applied again and again until they reach the conclusion that they cannot make any money from the teacher's offer.*

**Bibliographical Source:** Sorensen 1988 (47), p. 337.

## 3. Solutions to the surprise examination paradox

The surprise examination paradox has been the topic of many philosophical papers, but despite its long history, no long-lasting solution has yet emerged. However, there are two solutions that have received more attention than others, namely Quine 1953 (42) and Wright and Sudbury 1977 (59). Both entered the philosophical folklore as definite solutions and there are still philosophers who would easily state that Quine solved the paradox or that Wright and Sudbury solved it. Quine himself was one of the philosophers who believed that his solution solved the paradox: "[i]t is clear to me that I solved the puzzle, but is still perhaps not clear to all concerned." [7] Nevertheless, despite their empirically proven appeal, both solutions have received serious criticism.

The landscape of solutions is very varied; the surprise examination paradox has been seen as (A) a paradox about: epistemic/doxastic logic (Kaplan and Montague (28), Wright and Sudbury 1977 (59), among many others), belief revision and trust (Quine 1953 (42), Baltag 2009,2010(5; 6) and Baltag and Smets 2010 (7)), intentionality (Sorensen 1988 (47)), game theory (Cargile 1967 (13), Olin 1983, 1988 (36; 37), Sorensen 1988 (47)); (B) another manifestation of: Moore's paradox (Binkley 1968 (8), Wright and Sudbury 1977 (59), Gerbrandy 1999, 2007 (20; 21), Baltag and Smets 2010 (7)), Kavka's toxin puzzle (Sorensen 1988 (47)), prisoner's dilemma (Olin 1983, 1988 (36; 37)), self-referentiality (Shaw 1958 (44) Kaplan and Montague 1960 (28)), KK principle (McLelland and Chihara 1975 (32)), Godel's sentence (Nerlich 1961 (34), Fitch 1964 (18)), vagueness (Dietl 1973 (16), Smith 1984 (46), Williamson 2000 (58)). In addition to the variety of solutions proposed, first-line philosophers

[7] *The Time of My Life*, Cambridge, MA: MIT Press, 1985, p. 234.

have contributed to these debates, among them: Alfred Ayer, Charles Chihara, Frederic Fitch, David Kaplan, Saul Kripke[8], Richard Montague, Willard Quine, Timothy Williamson, Crispin Wright. Despite this plurality of views none managed to survive careful scrutiny. Below are a few examples:

Quine 1953 (42) claims that since after the backward argument the students conclude that there can be no examination, then this should also be a possibility on Thursday evening. Hence the students will no longer be able to eliminate Friday as a possible examination day. This obviously relies on the idea that they can doubt the fact that the exam will take place that week, that is, that they can doubt the teacher. However, in the "Guessing the Card" variation (initially formulated by Ayer 1973 (1)) "doubts akin to the one about (...) the fidelity of the teacher do not create uncertainties. Only a broad skepticism seems to suffice to prevent the audience from knowing there is an ace of spades if fifty-one of the remaining cards have been removed from the deck."(Sorensen 1988 (47), p. 311).

Wright and Sudbury 1977 (59) claim that the announcement of the teacher can be known by the students. However, if all possible examination days, but the last, pass without an exam being given, then at that point the students are faced with a Moore sentence (the teacher's announcement together with their information that the exam has not been given until the last day imply a Moore sentence): "The examination is on the last day but the students do not believe it". Therefore the students loose the knowledge they had. Sorensen 1988 (47) argues that Wright and Sudbury's solution relies on the temporal, sequential aspect of the variation they are using: some days pass without knowledge being lost and there is a day whose passing generated a lost in knowledge. Sorensen thus construct the "Designated student" variation in which, after the teacher has placed the stars on the student's necks, has them break formation and by that act the students collectively and instantly learn which of them had silver stars and who had the gold star. Therefore, in this variation both the elimination of silver star students and the discovery of the gold star student happen in one go, as it were. Of course, one could argue that even in Sorensen's variation, the students do not learn "in one go" everything, since every student observes the stars of the other students one at a time, so he performs successive updates with the type of the stars of the other students, much like the students in the surprise examination paradox performs successive updates with examination-free days. However, a second, more powerful criticism comes from Chihara 1985 (14) who argues that the Wright and Sudbury solution fails to work if we take a variation of

[8]In a lecture given to the Moral Sciences Club in Cambridge in 1971 titled "Two paradoxes of knowledge", forthcoming in Kripke, S., *Collected Papers, Volume 1*, New York: Oxford University Press.

the surprise examination paradox in which the teacher guarantees to the students they will not reach a position in which they will loose their knowledge. [9]

Kaplan and Montague 1960 (28) devote their attention not to proving that there is no inherent paradox in the way in which the scenario of the surprise examination paradox is set up, as most other authors do, but on how to make sure and prove that the scenario is indeed paradoxical. They identify the source of the paradox in holding all of the following three principles at the same time:

**S1:** $K_S(\bar{\phi}) \supset \phi$
**S2:** $K_S(K_S(\bar{\phi}) \supset \phi)$
**S3:** $[I(\bar{\phi}, \bar{\psi}) \wedge K_S(\bar{\phi})] \supset K_S(\bar{\psi})$

The $K$ in the above formulas refer to a knowledge predicate, and Kaplan and Montague's conclusion is that "[I]f any of $S1-S3$ is removed, it can be shown that the remaining schemata are compatible with the principles of elementary syntax."(88-89) I will not criticize it here, although there are lines of criticism directed against their analysis in Wright and Sudbury 1977, section IV, for instance. The reason is that I believe their approach is significantly different than the other approaches considered here, instantiating a special type of attitude one could have towards the surprise examination paradox.

The list of solutions offered and of criticisms being addressed to them is extremely long. As a final example, even one of the most recent attempt to solve the paradox, i.e. Williamson 2000 (58) has been met with criticism, especially due to its heavy reliance on the Margin for Error principle. But I will not insist on this any further. Let us now turn our attention to what it means to solve the surprise examination paradox.

## 4. Solving the surprise examination paradox

As mentioned before, there are several issues that surround the surprise examination paradox. Therefore, talking about solving the surprise examination paradox is too vague. So, this section is dedicated to the clarification of what exactly I mean by solving the surprise examination paradox.

The most fundamental issue with the surprise examination paradox is whether or not it is indeed a paradox. Let us explore both alternatives. If the surprise examination paradox is indeed a paradox then the scenario (at some point, but most likely in the teacher's announcement) contains a genuine contradiction. Obviously,

[9] As a side remark, I would re-interpret the solution of Wright and Sudbury in terms of a belief-based definition of surprise and of a revision of their beliefs. But even so, Chihara's criticism still stands, although Sorensen criticism based on "Designated student" paradox does not, since the revision with the type of stars the other students have would have to be sequential, and not "in one go".

the scenario does not contain any explicit contradictions, therefore the contradiction must be phrased in an unfamiliar way. If the surprise examination paradox is indeed a paradox, a solution to it must reveal, beyond the shadow of a doubt, what the contradiction really is and nothing except that. Together with such a solution, an explanation of the feeling of puzzlement that one has when introduced to the puzzle comes for free: paradoxes are, by their very nature, puzzling. The only additional problem that such a solution would need to address is the problem of its scope, i.e. should it extend to all variations or not? However, as I said in the previous section, this depends (at least for now) on the intuitions of the author holding that solution, and it does not have any intimate connections to the attitude one has with respect to the surprise examination paradox.

If one were to take this first attitude towards the surprise examination paradox and treat it as a real contradiction, the implications of a correct solution (according to the criterion spelled out in the previous paragraph) could be very far-reaching. At least, this is the lesson Kaplan and Montague 1960 (28) advance. They argue, as stated in the previous section, that the surprise examination paradox is indeed paradoxical and that, since the paradox stems from holding $S1-S3$, one of them has to be dropped. But the underlying assumption if this conclusion is that the scenario whose analysis leads to paradox is not really paradoxical. In other words, revision at the level of our theory of knowledge is needed when a paradox is found in the formulation of the surprise examination paradox only if our formulation cannot be itself contradictory. This is Kaplan and Montague's assumption, and I endorse it. It should be obvious to anyone, I believe, that situations as those described by the surprise examination paradox occur every day: teachers around the world announce surprise examinations, and their students end up being surprised by them. Therefore, in the real world, the scenario is not contradictory. Ergo, our way of understanding the real world is deficient if it contains a contradiction: it generated a contradiction where there is none. So, for those who hold that the surprise examination paradox is truly paradoxical, solving it could have a big philosophical relevance: it may lead to significant revision in our theories about the real world.

On the other hand, if the surprise examination paradox is not a real paradox, then, of course, there is no hidden contradiction in its formulation. In this case, a solution to the surprise examination paradox has to do more than before. Firstly, it has to offer the right kind of framework for understanding it and the right interpretation in that framework of all the relevant aspects of the paradox. This may be game theory, or epistemic logic, but no matter which framework it is, the conclusion will be the same: when we understand all the facts of the scenario by means of the right tools, the paradoxical flavor withers away. However, if indeed there is no paradox, a solution has to offer a bit more, namely the reason why the paradox remained

unsolved for so long[10] and why the puzzling reasoning of the students seems so intuitive[11]. Nevertheless, if the surprise examination paradox proves to be solvable, there is no need for revising our theories about the real world. Moreover, the reason why the surprise examination paradox has been around for so long is explained by this view as being just the fact that no one has managed to cast it in the right light. But this does not mean that there is nothing to be learned from solving the surprise examination paradox when construing it as a non-paradoxical puzzle.

These two ways of seeing the surprise examination paradox are not necessarily complementary. I see no contradiction in someone holding that if the paradox is interpreted in, say, game theory it generates a real paradox in that field which, in turn, leads to a revision of game theory, but that if we interpret the paradox in epistemic logic, the surprise examination paradox can be shown to be only apparently paradoxical, and a corresponding solution can be found. However, I prefer the second view, for reasons that I will present below.

Let us try to make more precise the conditions that a solution to the surprise examination paradox (according to my preferred view, i.e. the second) has to meet. The first to explicitly discuss conditions for a solution were Wright and Sudbury (59). Below I compile a list of criteria that I believe most philosophers (who take the same attitude as I do towards a solution to the surprise examination paradox, that is almost everyone besides Kaplan and Montague) would agree the solution to the surprise examination paradox should meet. This list is mostly based on Wright and Sudbury's.

(1) The solution should make the teacher's announcement satisfiable [12].
(2) The solution should make it clear that the teacher can carry out the announcement even after he has announced it [13].
(3) The solution should do justice to the intuitive meaning of the announcement.
(4) The solution should do justice to the intuitive plausibility of the pupils' reasoning.
(5) The solution should make it possible for the students to be informed by the announcement[14].

[10] One possible explanation could be that the right tools for analyzing the surprise examination paradox have just been discovered.

[11] One possible explanation could be that there is a very subtle difference, that in most common situations we ignore.

[12] "since a surprise examination is, palpably, a logical possibility", Wright and Sudbury 1977(59).

[13] "since, palpably, he can." Wright and Sudbury 1977.

[14] Sorensen 1988, interprets this condition as saying that the students should be able to come to know what the teacher announced: "if you cannot trust your teachers, who can you trust?" (p. 312).

(6) The solution should explain the role, in the generation of the puzzle, of the announcement's being made to the students [15].
(7) The solution should account for the fact that before the teacher's announcement the students are indifferent both with regards to there being an exam and, conditional on there being an exam, with regards to its actual date.
(8) The solution should do justice to the intuition that there is something truly interactive in the scenario: after the teacher's announcement, both the teacher and the students learn something about one another.
(9) The solution should be applicable to *all variations* of the surprise examination paradox.

I will not defend these intuitions here, since I believe they are straightforward for someone who agrees that a surprise examination can be given in real life and that students are indeed surprised by it. Moreover, most of them are already common place in the philosophical literature on the surprise examination paradox. It can even be said that although the history of the paradox is full of new beginnings (almost everyone starts an analysis by saying that all the analyses before him were wrong), the criteria that a solution has to meet have had a cumulative evolution. For instance, (1)-(6) are Wright and Sudbury's conditions, and almost everybody after them accepts them as being correct. Moreover, (8) is just a strengthening of condition (5), and it should be obvious that also the teacher has to adapt his strategy according to what he announces: the teacher cannot act in the same way if he announces that the examination will come as a surprise and if he does not announce that. Finally, condition (9) is almost never explicitly stated, but it is assumed by most authors (indeed not all, Baltag and Smets 2010 explicitly say that different variations mean different paradoxes, no matter how big the similarities between them are). However, this principle is assumed by Ayer 1973, Sorensen 1988, and Williamson 2000 among many others.

I divide these conditions into two categories: strong conditions (1,2,5,6,7,8,9) and lax conditions (3,4). When evaluating a solution according to this grid, the questions regarding the strong conditions are objective questions: the teacher's announcement is either satisfiable or not, the teacher can either carry it out (after making the announcement), or not; whether there is a difference between announcing the students and announcing another teacher, or not; the students are either informed by the announcement or not; the initial scenario accounts for the student's ignorance or not; the attitudes of both teacher and student are influenced by announcing that a surprise examination will come, or not; and finally, the solution is applicable to all

[15] "there is, intuitively, no difficulty if, e.g. the teacher tells only the second teacher or keeps his intentions to himself."

variations or not [16]. On the other hand, the lax conditions are more subjective since they engage our intuitions; however, it can be a matter of interpretation whether a certain solution captures the intuitive notion of surprise or whether according to that solution it is intuitive why the students erred in their initial reasoning. Although I will be interested in how all of these conditions are satisfied by the solutions I will discuss in this thesis (including my own solution), I just want to direct your attention mainly on the strong ones, as those are, in my opinion the most relevant and the most decisive (any rational agent should accept that different agents consider other things as being intuitive).

## 5. Motivations

The title of this thesis is "The Surprise Examination Paradox in Dynamic Epistemic Logic". The purpose of this section is to motivate why the surprise examination paradox is an interesting topic and why we should investigate it by means of dynamic epistemic logic.

I will begin with the first issue. In the previous section I have presented the conditions that a solution to the surprise examination paradox should meet, given my preferred way of viewing it, namely as a non-paradoxical puzzle. However, now, that we know how a solution should look like, assuming that such a solution is indeed possible, the philosophically significant question becomes: "what can be accomplished by solving the surprise examination paradox?". In other words, if the surprise examination paradox is not really a paradox, so that solving it we revise our theories, what can a solution (in the above sense) can teach us? There are two ways of answering this question. The first possible answer is a very non-philosophically one: "nothing; it's just a puzzle! What do you accomplish when you solve a Sudoku game other than polishing your puzzle-solving skills!" I agree that this is a possible attitude, however, if this is the most we can say about the meaning of a solution to the paradox, then all the literature so far has been nothing more than a display of cleverness.

The second possible answer, which I prefer, is a more philosophically interesting one: besides the fact that we could give an answer to a puzzle and approach the end of a debate that has lasted for more than 60 years, a solution to the surprise examination paradox could also prompt us to change our preferred way of understanding certain types of scenarios. Sorensen represents the source of this second possible answer:

> a problem's recalcitrance is evidence that we have underestimated the relevance of the problem to our cherished beliefs. Often we learn that the apparent irrelevance was an illusion generated by

[16] Intuitively, if these answers are not straightforward, there is something wrong with the solution, or at least its presentation.

> misplaced loyalties or a gap in our understanding. Such was the case with the prediction[17] paradox. (p. 256)

The surprise examination paradox is a paradox that deals with problems related to the exchange of information between agents. One of the central aspects of the scenario of the surprise examination paradox is precisely how to interpret the announcement the teacher makes to the students and how this announcement affects the students. Moreover, the different ideas that have been proposed for clarifying the teacher's announcement and its effect on the students have been ideas that are typical for analyzing the exchange of information between different agents, i.e. do the agents trust each other? are the agents ideal? what is the background information the agents have? etc. Even if we think of a variation of the surprise examination paradox with only one agent, like in Williamson's variation, the same problems can still emerge: can the students trust the fact that the teacher only circles examination days in her calendar? are the students ideal? etc.

The fact that no solutions so far have been able to account for all the intuitive conditions presented in the previous section suggests either that our intuitive understanding of the paradox is mistaken and that it is indeed an impossible scenario, or that the theories we use to reason about the paradox are not well-suited for that purpose. The former alternative would be very strange: all the above intuitive conditions are more or less supported by actual similar scenarios as the one of the surprise examination paradox. Therefore, finding a solution to the surprise examination paradox that meets the above conditions would have great significance: it would represent a very strong argument to the effect that the framework employed by that solution is very appropriate for investigating scenarios that deal with the exchange of information between agents. In additions to that, the way in which the notion of surprise would be interpreted in that framework might be relevant to a theory about surprise. In consequence, the surprise examination paradox could help us discriminate between different frameworks that are suited to analyze the sort of topics that the paradox involves, and which are all related to the process of information exchange. However, there need not be *one* right framework for analyzing the surprise examination paradox. There is no contradiction in saying that more frameworks can give equally good solutions to the paradox (given the above conditions for a solution). This is perfectly plausible. It means that the surprise examination paradox cannot distinguish between those frameworks. But, since the focus of this thesis is the surprise examination paradox, and since no solution that satisfied the conditions from the previous section has yet been found, finding one such solution will be sufficient for the present purposes.

[17] This is the general name that Sorensen gives to the surprise examination paradox so that it covers all of its variations

So, the way in which the investigation of the surprise examination paradox will unfold is similar to an existential proof. The idea is to interpret the puzzle in a right framework, trying to interpret the relevant aspects of the puzzle (i.e. surprise, the teacher's announcement) in different ways given so that all conditions are met. Of course this is not a blind search, but nevertheless, once a right framework and a right interpretation of the key fact in that framework is met so that all conditions are met, the process stops: the paradox has been solved and there would be no real motivation in going further.

Therefore, if the frameworks employed so far for solving the surprise examination paradox have failed to vindicate our intuitions about how a solution should look like (given my prefered view of the paradox), but a new framework manages to do that, it means that, although our theories do not need revision, they do need to be changed: our cherished ways of investigating the world do not stand the tribunal of the surprise examination paradox and our loyalty to them has been misplaced. This is the meta-philosophical lesson that a solution (according to my preferred view) to the surprise examination paradox can teach us.

To sum up: we need to continue studying the surprise examination paradox since no solution satisfying the conditions specified in the last section has yet been given, and because such a solution might have some significant lessons to teach us and can lead to a change in our preferred way of looking at the world.

In this thesis I attempt to give a solution to the surprise examination paradox by means of dynamic epistemic logic (DEL). Such a project could raise the following question: why use DEL in order to solve the surprise examination paradox? Here are a few reasons: DEL has proven to be a suitable framework for formalizing different epistemic puzzles (e.g. Muddy Children, see van Ditmarsch et al. 2007 (55), ch. 4); DEL has proven to be a nice framework for making some connections between cognitive science and belief revision theory which concluded with a (dynamic) logic of surprise (see Lorini and Castelfranchi 2007 (31)); DEL has proven to be well suited for dealing with different types of communication, e.g. between deceitful agents, over time, etc. (see van Benthem forthcoming (52)). In addition to this DEL has proven well suited for capturing intentions and game theoretical reasoning (e.g. van Benthem et al. forthcoming (53), respectively). However, not all these connections will be of interest for the present thesis. The reason for this enumerations was just to show that DEL is well equipped for dealing with the issue that surround the surprise examination paradox. Moreover, DEL is one of most respectable frameworks for dealing with the way in which information is transferred between agents and with the way in which information affects agents. All of this would make DEL one of the most plausible framework to give a satisfactory solution to the surprise examination paradox (i.e. a solution that respected the conditions presented in the previous section).

However, someone familiar with the DEL literature might know that there already are two analyses of the surprise examination paradox in this framework: (i) Gerbrandy 1999 (20), restated in Gerbrandy 2007 (21) and Kooi and van Ditmarsch 2008 (54); and (ii) Baltag and Smets 2010 (7), restated in Baltag 2009 (5) and Baltag 2010(6). These authors believe that a solution to the surprise examination paradox could be given if we analyze the facts with the right tools (public announcement logic according to Gerbrandy and dynamic belief revision for Baltag and Smets). However, their solutions do not really meet the criteria presented above: they claim that there is no solution (or better said their solutions do not work) if the teacher's announcement is meant to be fulfilled (Gerbrandy), or if the students are to trust the teacher (Baltag and Smets). Therefore, despite the fact that such approaches have both philosophical and logical merit and all authors manage to put their conclusions to work towards far-reaching philosophical conclusions regarding knowledge and the way in which agents revise their beliefs in face of new information, they fail to meet widely accepted criteria that philosophers expect a solution to the surprise examination paradox to meet. For example, the students should be surprised even after the teacher's announcement, which Gerbrandy cannot accommodate, and a surprise examination should indeed be possible, which Baltag and Smets cannot accept. Therefore these solutions to the surprise examination paradox in DEL are not the right kind of solutions I am looking for. I will show that the DEL "solutions" are not good enough solutions and I will show that a solution that meets the conditions from the previous section is possible in DEL.

In conclusion, the question that drives the investigation in this thesis is: "Can DEL vindicate our pre-theoretical intuitions regarding how a solution to the surprise examination paradox should look like?". So far a lot of solutions have been defended by recourse to a logical analysis of the scenario, e.g. Wright and Sudbury 1977 (59), McLelland and Chihara 1975 (32), Binkley 1968 (8), and many others. All of them have received great consideration in the literature and all of them have been proven to miss at least one of the intuitions presented above, e.g. all three have been shown to be unable to solve a variation to the surprise examination paradox. Moreover, in recent years the logical analysis of the paradox has gained new thrust by the interest dynamic logicians have shown for this paradox. Their solutions have not yet been evaluated. In my thesis, I will fill this gap. I will present and evaluate the DEL solutions to the surprise examination paradox. The outcome of my evaluation will be that not even they manage to account for the pre-theoretical intuitions. In consequence, one might be inclined to believe that these intuitions, although apparently coherent and consistent, are in fact impossible to satisfy. This would be a wrong answer, in my opinion. After all, they are guided by actual scenarios in which students are announced surprise examinations and in which they are actually surprised when the examinations actually come. Therefore, the starting point of my

investigation is the feeling that there has to be a way of modelling the scenario of the surprise examination paradox in such a way that reality (the reality of being able to be announced and to receive surprise examinations) is saved. The conclusion will be that such a solution exists.

CHAPTER 2

# Logical Preliminaries

In this section I briefly present the different logics that I later refer to or use in this thesis. They consist in: public announcement logic, dynamic epistemic logic, epistemic temporal logic. In addition to these I will present a way in which DEL and ETL can be put together into one single framework and a way in which beliefs can be revised in the framework of DEL. The big absent in this list is obviously epistemic logic and the reason is that I consider it to be too well entrenched in both the philosophical and in the logical traditions for it to require any further exposition. Nevertheless, some basic facts about it will be presented along side public announcement logic. One note about the meaning of DEL. In recent years, DEL has become the general name for a large family of different logics that are all inspired by the idea of "a dynamic turn in logic". This is the same way in which I use this term in the other chapters of my thesis. So, from the point of view of the other chapters, all the logics presented in this chapter are DEL logics. However, in this current chapter, DEL is intended to refer to the system developed in Baltag et al. 1998 (2), and presented in section 3.

As it should have already been obvious from the introduction, this thesis is meant to be more a thesis in (formal) epistemology than in logic, and for this reason I will omit any discussions regarding the axiomatic systems and the completeness results of the logics presented. Also, in evaluating different frameworks the emphasis will always be on modelling virtues rather than on computational ones. Therefore, in this chapter, as well as in the rest of my thesis I will only focus on semantic issues.

The structure of the following sections will be as follows: I will first give a few bibliographical references focusing mainly on the origin of the logic presented in that section and on general overviews of that logic. Secondly, I will give some motivations for the logic presented. Thirdly, I will present the basic features of that logic, focusing on models, language and semantics. In the end I apply the logic to solve the scenario that served as its motivation.

## 1. Public Announcement Logic

**Bibliographical Note:** Developed in Plaza 1989 (41), and in Gerbrandy and Groeneveld 1997 (19) and Gerbrandy 1999 (20). Overviewed in van Ditmarsch et al. 2007, chapter 4 and van Benthem forthcoming (52), chapter 3.

EXAMPLE 1 (3 friends in a bar not to mention the car). *Alan, Bob and Carl are sitting at table in their favourite bar. Alan, who is well known among his friends as always being sincere says: "I bought a car!".*

This is a very trivial example, and epistemic logic can very well explain what goes on: you have two different time points 1 and 2. At 1 Carl knows that he bought a car, while Bob and Carl do not know it. However, at 2, all three friends know that Alan bought a car, and what is more they know that they all know it (and that they all know it that they know it ...). This is completely correct. However, what public announcement logic (PAL) can add on top of this is an explanation of how the three friends get from not knowing that Alan bought a car to them knowing that Alan bought a car and that they all know it (and that they all know it that they know it ...). In other words, PAL can give an account of the transition from one state (1) to another (2). That is what the dynamic nature of PAL is all about.

DEFINITION 1 (Epistemic Model). *Given a (finite) set of agents, Agt, with $i, j, k, \ldots$ as its elements, and a countable set of atoms Prop,with $p, q, r, \ldots$ as its elements, a PAL model is a structure $M = \langle S, R_i, V \rangle$, where:*

*(1) $S$ is a (finite) non-empty set of states;*
*(2) $R_i$ is a set of equivalence relations on $S$ for every agent $i \in Agt$;*
*(3) $V$ assigns a set of states to each propositional variable from the set Prop, that is $V : Prop \to 2^S$.*[1]

DEFINITION 2 ($\mathfrak{L}_{PAL}$). *Given a (finite) set of agents, Agt, with $i, j, k, \ldots$ as its elements, and a countable set of atoms Prop, with $p, q, r, \ldots$ as its elements, the language of PAL, $\mathfrak{L}_{PAL}$ is given by the following Backus-Naur form:*

$$\varphi ::= p | \neg\varphi | \varphi \wedge \varphi | K_i\varphi | [!\varphi]\varphi$$[2],

The above definition might seem circular due to the occurrences of $\varphi$ on both sides of the symbol for definition. However, this disturbing observation can be overcome by noticing that $\varphi$ does not represent a formula but rather the *type* of a formula. Therefore, a formula is well formed if it is an atom, or if it is formed by negating a well formed formula, or if it is formed through the conjunction of two well-formed formulas, and so on. van Benthem forthcoming explains this in an elegant and

[1] This is also the same type of model on which epistemic logic is based.

[2] The fragment of this language without the dynamic formula is the language of epistemic logic.

perspicuous way: "formulas are all and only the syntactic strings arising from this recursive process in a finite number of steps."(p.) This might be confusing, but definitely not circular.

The language of PAL extends the language of epistemic logic with formulas of the type $[!\varphi]\varphi$. Their intended interpretation is: "a sentence of the type $\varphi$ holds after a public and truthful announcement of a sentence of the type $\varphi$ has been made."[3] Remark that this language allows for (finite) iterations of both the knowledge operator and the announcement operator and mixes in between, that is: $K_iK_j\varphi$ and $[!\varphi][!\psi]\chi$ as well as $[!\varphi]K_i\psi$ and $K_i[!\varphi]\psi$.

DEFINITION 3 (PAL Semantics). *Given a PAL model $M = \langle S, R_i, V\rangle$, with $s \in S$:*

$$
\begin{aligned}
&M, s \vDash p \text{ iff } s \in V(p)\\
&M, s \vDash \neg\varphi \text{ iff } M, s \nvDash \varphi\\
&M, s \vDash \varphi \wedge \psi \text{ iff } M, s \vDash \varphi \text{ and } M, s \vDash \psi\\
&M, s \vDash K_i\varphi \text{ iff } \forall t : sR_it \Rightarrow M, t \vDash \varphi^{4}\\
&M, s \vDash [!\varphi]\psi \text{ iff } M, s \vDash \varphi \Rightarrow M^{!\varphi}, s \vDash \psi
\end{aligned}
$$

DEFINITION 4 (Update Model). *Given an epistemic model $M = \langle S, R_i, V\rangle$ and a sentence $\varphi \in \mathfrak{L}_{EL}$, the updated model, $M^{!\varphi} = \langle S^{!\varphi}, R_i^{!\varphi}, V^{!\varphi}\rangle$, is defined in the following manner:*

$$
\begin{aligned}
S^{!\varphi} &= \{u \in S : M, u \vDash \varphi\}\\
R_i^{!\varphi} &= R \cap (S^{!\varphi} \times S^{!\varphi})\\
V^{!\varphi} &= V
\end{aligned}
$$

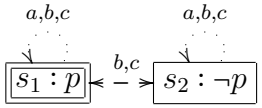
The meaning of all but one of the above semantic clauses should be straightforward for someone familiar with epistemic logic, and I will not explain them further here. The only interesting clause is the one for $[!\varphi]\psi$ (as usual, $\langle\rangle ::= \neg[]\neg$). Notice first that the dynamic nature of PAL is clear from its semantics: the statement that $[!\varphi]\psi$ is interpreted at model $M$, by reference to a different model $M^{!\varphi}$.

The effect of the public announcement of sentence $\varphi$ on an epistemic model M is the elimination of all states in M at which $\varphi$ did not hold. However, this does not imply that whenever $\varphi$ is announced $\varphi$ will hold, or will become known, that is:

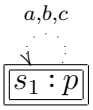
[3] It is not necessary to interpret the public announcement as being always truthful, and indeed Gerbrandy 1999 and 2007 defines a public announcement as being always executable. In other words, Gerbrandy's public announcement can be performed even if false. This obviously changes the semantics for a public announcement.

$\not\models_{PAL} [!\varphi]\varphi$ and $\not\models_{PAL} [!\varphi]K_i\varphi$. In other words, it is not the case that all formulas are successful. A trivial example is $\varphi ::= p \wedge \neg K_i p$, although this might very well be the case in some epistemic model (it is a truism that there are true things that we do not know) it is a well-known fact that it can never be the case that $K_i(p \wedge \neg K_i p)$. Nevertheless, if we restrict the language of PAL to atoms closed under boolean operations (that is if we do not allow announcements of higher-order information), then both principles hold. A final interesting fact about the semantics of PAL presented here is that since the announcements have to be truthful ($\boldsymbol{M, s \models \varphi \Rightarrow} M^{!\varphi}, s \models \psi$), it is not the case that anything can be announced, that is $\not\models_{PAL} \langle\varphi\rangle\top$. The jargon is: "not all announcements are feasible in a given epistemic model".

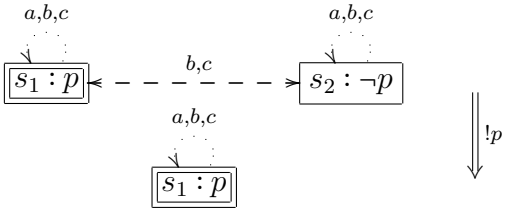
Returning to the initial scenario, let $PROP = \{p\}$, where $p$ means "Alan bought a car", $Agt = \{a, b, c\}$, where $a$ stands for Alan, $b$ for Bob and $c$ for Carl. The initial epistemic model is $M$:




Alan's announcement can be formalized as $!p$, and the effect is the following updated model, $M^{!p}$




In one picture, the scenario can be modelled in the following way:




This formalization not only expresses that $b$ and $c$ do not know initially that Alan bought a new car ($p$), although Alan knows it ($M \models K_a p \wedge \neg K_b p \wedge \neg K_c p$) and that after Alan's announcement they all know it and know that they know it, etc. ($M^{!p} \models K_a p \wedge K_b p \wedge K_c p \wedge K_a K_b p \wedge K_c K_a K_b K_c p \ldots$[5]), but also it give an account of

[5]Note however, that the language of PAL presented here is not powerful enough to express the true fact that they have common knowledge, since that would imply having an infinite conjunction. The problem is easily avoidable by extending $\mathfrak{L}_{PAL}$ with a common knowledge operator. But I will not do that here. See van Ditmarsch et al. 2007, chapter 4 for details.

how the information changes, i.e. through the public announcement of $p$ (in other words the transition is explained in the model). That is:

$$M \vDash [!p](K_b p \wedge K_c p)$$

Remark here that there is an important piece of information not represented in this formal setting, namely the fact that Alan was the one announcing that he bought the car, and it was he who was well known (among his friends) as being truthful. This omission might become very relevant, but I will come back to this problem later.

## 2. Reasoning with soft information

**Bibliographical Note:** Developed in van Benthem 2007 (50) and Baltag and Smets 2008 (4). Overviewed in van Benthem forthcoming (52).

Notice that in the previous section the focus was mainly on knowledge. However, from a philosophical point of view, the more important attitude is belief. The following definition presents a framework in which we can reason about both belief and knowledge.

Definition 5 (Epistemic-doxastic models). *Given a (finite) set of agents, Agt, with $i, j, k, \ldots$ as its elements, and a countable set of atoms Prop,with $p, q, r, \ldots$ as its elements, an epistemic-doxastic model is a structure $M = \langle S, R_i, \leq_i, V \rangle$, where:*

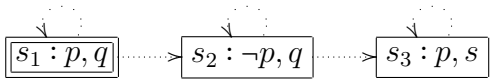
*(1) $S$ is a (finite) non-empty set of states;*
*(2) $R_i$ is a set of equivalence relations on $S$ for every agent $i \in Agt$;*
*(3) $\leq_i$ is a set of reflexive, transitive, binary relations on $S$, such that every non-empty subset has maximal elements;*
*(4) $V$ assigns a set of states to each propositional variable from the set Prop, that is $V : Prop \to 2^S$.*

In this framework, the $\leq_i$ relation is interpreted as a plausibility relation and $s \leq_i t$ reads "agent $i$ considers $t$ more plausible than $s$". It makes sense to define the belief that an agent has not as something true in all the states that he considers possible, but only the states that he considers *most* probable. After all, only those seem relevant. The next definition distinguishes between more types of belief an agent can have.

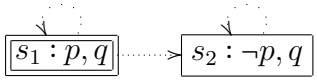
Definition 6. *Given an epistemic-doxastic model $M = \langle S, R_i, \leq_i, V \rangle$, with $s \in S$, let $Max_i$ stand for the set of maximal states for agent $i$, and let $MAX_i = \{t : sR_it\}$:*

$$M, s \vDash B_i\varphi \text{ iff } \varphi \in MAX_i$$
$$M, s \vDash B_i^{\psi}\varphi \text{ iff } \varphi \in (MAX_i \cap \{s \in S : M, s \vDash \psi\})$$
$$M, s \vDash B_i^{+}\varphi \text{ iff } \forall t \text{ such that } s \leq_i t : M, t \vDash \varphi$$

The intended interpretations of the above formulas are: "agent $i$ believes $\varphi$" (I will refer to this from now on as the maximal definition of belief, in order to distinguish it from the KD45 operator which is standardly used to define belief), "conditional on $\psi$, agent $i$ believes $\varphi$", and "agent $i$ safely believes that $\varphi$". The meaning of "safely believes" is that this attitude is not affected by public announcements in the way in which mere belief is. Take the following example, in which only the doxastic relations are represented (all states are epistemically indistinguishable):



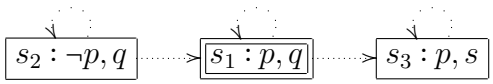

In this model the agent believes that $p$ is the case, since $p$ is true in his maximal state ($s_3$). It so happens that $p$ is really true in the actual state ($s_1$). So, it may be said that he believes the right thing, but for the wrong reasons (this is the essence of Gettier examples). Now, consider the effect of a (truthful) public announcement, $!\neg s$:




The agent now believes $\neg p$, which is false. So, the agent lost his truthful belief by being exposed to a truthful announcement. This seems to suggests that belief is a very instable attitude. This motivates the introduction of safe belief, whose behaviour is more stable.

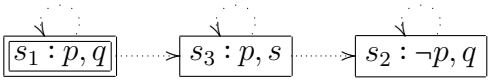
Nevertheless, when dealing with beliefs one important thing one needs to account for is the way in which beliefs are modified by incoming information. In this context it makes sense to explore ways in which the states that an agent entertains as possible get re-arranged, not only eliminated, like in the case of public announcement. The type of information that leads to the elimination of states has been dubbed "hard information", while the information that merely induces a change in the order of plausibility that an agent assigns to the states he entertains as possible has been dubbed "soft information". Here are two ways in which incoming information can be taken "softly":

(1) Lexicographic upgrade ($\Uparrow$). After a lexicographic upgrade with sentence $\varphi$, all the $\varphi$-states become more plausible than the $\neg\varphi$- states, while the order within the $(\neg)\varphi$-states remains unchanged. For instance, the effect of a lexicographic upgrade with $p$ on the previous epistemic-doxastic model would be:




(2) Conservative upgrade ($\uparrow$). After a conservative upgrade with sentence $\varphi$, the most plausible of the $\varphi$-states becomes the most plausible state. The rest of the states remain in the same relation to one another. For instance, the

effect of a conservative upgrade with $q$ on the previous epistemic-doxastic model would be:

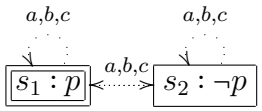



## 3. Dynamic Epistemic Logic

**Bibliographical Note:** Developed in Baltag et al. 1998 (2) and overviewed in van Ditmarsch et al. 2007, chapter 6 and van Benthem forthcoming (52), chapter ?.

Example 2 (3 friends in a bar and a horse race). *Alan, Bob and Carl are sitting at a table in their favourite bar. They all have bets placed on the same horse, but unfortunately none of them was able to attend the race and the race is not on TV. Therefore, they are all waiting for Bob to receive a call from a friend of his who is at the race track and who promised to announce Bob the outcome of the race as soon as it ends. Bob's friend finally calls, and Bob talks to him, and from what Alan and Carl hear they deduce that Bob now knows the outcome of the race (since Bob is known among his friends as being very calm in all situations, they cannot read the outcome of the race from his behaviour).*

In this example an informational event occurs: Bob learns the outcome of the horse race, but his friends do not know that outcome, although they know Bob knows and Bob knows that they know he knows. Let $PROP = \{p\}$, where $p$ means that Alan, Bob and Carl won some money and $Agt = \{a, b, c\}$, as before. Then the initial epistemic model $M$ is (supposing that they actually won):




Now, Bob's conversation with his friend changes this initial model in the following way: Bob now knows the outcome, and Alan and Carl know that Bob knows it, although they do not know what the outcome of the race is. However, a public announcement logic could not capture this change. The plausible public announcement $!(K_b p \vee K_b \neg p)$, meaning that Bob knows whether they won or not cannot be truthfully announced at $M$, since the sentence $M \not\models K_b p \vee K_b \neg p$. Dynamic epistemic logic gives a way to compute the outcome of witnessing a non-public act given an initial epistemic model. The private act (Bob's conversation with his friend) is being interpreted in this framework as an event model.

Definition 7 (Event Model). *Given the language $\mathfrak{L}_{EL}$ and the sets $Agt$ and $PROP$, an event model is a structure $E = \langle \mathfrak{E}, \rightarrow_i, pre \rangle$, such that*

*(1) $\mathfrak{E}$ is a non-empty set of events;*

*(2)* $\rightarrow_i$*:* $Agt \rightarrow 2^{\mathfrak{E}\times\mathfrak{E}}$ *is an accessibility relation between events;*
*(3)* $pre : S \rightarrow \mathfrak{L}$ *is a function that assigns to each action in* $S$ *a precondition (a proposition in* $\mathfrak{L}$*)*

$\mathfrak{E}$ is meant to represent a set of possible events that could take place. Thus, the intended meaning of $e \rightarrow_i f$ is "when event $e$ happens agent $i$ thinks that $f$ has in fact happened." Of course, if $e \rightarrow_i f$ and $f \rightarrow_i e$, then agent $i$ cannot distinguish between the situation in which e has happened and the situation in which $f$ has happened. Of course the events that can happen at a certain state have to be in some way related to the state at which they happen. For example, if the state is such that I haven't played the lottery, the event $e$: "I win the lottery" cannot happen. The precondition functions takes care of this connection between events and states. It assigns to each event a formula from the language of epistemic logic that is interpreted as the minimum fact that need to hold in order for the event to be able to occur. In our example, this fact was "having a lottery ticket".

Definition 8 (DEL language). *Given a (finite) set of agents, Agt, with* $i, j, k, \ldots$ *as its elements, and a countable set of atoms Prop, with* $p, q, r, \ldots$ *as its elements, the language of DEL,* $\mathfrak{L}_{DEL}$ *is given by the following double recursion:*

$$\varphi ::= p|\neg\varphi|\varphi \wedge \varphi|K_i\varphi|[\mathfrak{E}, e]\varphi$$

The language of DEL is the same as the language of EL but with an extra dynamic modality (different than the PAL dynamic modality). The intended interpretation of the formula This way of writing the language of DEL may seem puzzling to some. The dynamic modality described here is a model (an event model), and this would seem to suggest a certain circularity between the syntax and the semantics of DEL. A good exploration of this problem can be found in van Ditmarsch et al. 2007, chapter 6. In any case, the intended meaning of the dynamic modality is "if the event $(\mathfrak{E}, e)$ happens, then $\varphi$ will be the case".

Since, the language of DEL is mostly the same as the language of epistemic logic, and since the semantics of epistemic logic has been review in the previous section, I only focus on the semantics for the new dynamic modality.

Definition 9 (DEL semantics). *Given an epistemic model* $M = \langle S, R_i, V\rangle$*, with* $s \in S$*, and an event model* $E = \langle \mathfrak{E}, \rightarrow_i, pre\rangle$*:*

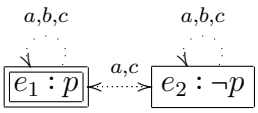
$$M, s \vDash [\mathfrak{E}, e]\varphi \text{ iff } M, s \vDash pre(e) \Rightarrow M \otimes \mathfrak{E}, (s, e) \vDash \varphi$$

The dynamic nature of this logic is again conspicuous by observing the fact that the dynamic modality is interpreted on different models: the precondition of $e$ has to hold on the epistemic model M, while $\varphi$ has to hold on the product update of M and the event model $\mathfrak{E}$, $M \otimes \mathfrak{E}$, where:

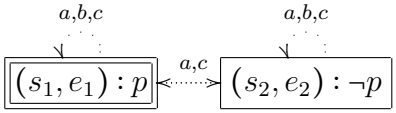
Definition 10 (Product update). *Given an epistemic model $M$ and an event model $\mathfrak{E}$, the product update of $M$ and $\mathfrak{E}$ is $M \otimes \mathfrak{E} = \langle S', R_i', V' \rangle$, such that:*

$$S' = \{(s, e) \in S \times \mathfrak{E} : M, s \vDash pre(e)\}$$
$$R_i' = \{((s, e), (s', e')) : sR_is' \text{ and } e \to_i e'\}$$
$$V'(p) = \{(s, e) \in S \times \mathfrak{E} : s \in V(p)\}$$

Going back to our example with the three friends in a bar, we can now model the effect of Bob receiving the phone call from his friend in front of Alan and Carl by means of an event model $\mathfrak{E}$




Therefore, after Bob receives the telephone call, the informational states of the three agents are denoted by the product update of their initial model and the event model described above:




This means that after the telephone call Bob received from his friend, Alan and Carl cannot tell if they witnessed event 1 (They won), or event 2 (They lost), however Bob can tell them apart and both Alan and Carl know that.

## 4. Epistemic Temporal Logic

**Bibliographical Note:** Developed in Halpern and Fagin 1989 and Parikh and Ramanujam 1985/2003 (39), overviewed in Fagin et al. 1995 (17).

Definition 11 (Protocol). *Given a non-empty set of events $\Sigma$. A history, $h$, on $\Sigma$ is a finite sequence of events from $\Sigma$. Let $\Sigma^*$ be a set of histories on $\Sigma$. Then, a protocol on $\Sigma$ is a set $H \subseteq \Sigma^*$ closed under non-empty finite prefixes.*

Let $he$ denote the fact that history $h$ is followed by event $e$. Given $h, h' \in \Sigma^*$ and $e \in \Sigma$, (i) $h \preceq h'$ denotes that $h$ is a prefix of $h'$, that is, $\exists h'' \in \Sigma^*$ such that $hh'' = h'$ and (ii) $h \prec_e h'$ denotes that that $h'$ represents the moment in time after $e$ occurred in $h$. We say that $H \subseteq \Sigma^*$ is closed under non-empty finite prefixes if $\forall h \in H : h' \preceq h \Rightarrow h' \in H$.

Definition 12. *Given a set of agents, Agt, with $i, j, k, \ldots$ as its elements, an ETL model is a structure $\mathfrak{H} = \langle \Sigma, H, R_i, V \rangle$ such that:*

*(1) $\Sigma$ is a non-empty set of events;*

(2) *$H$ is a protocol on $\Sigma$;*
(3) *$R_i$ assigns an accessibility relation to each agent, $R_i : Agt \to 2^{H \times H}$;*
(4) *$V : PROP \to 2^{H \times H}$ assigns a set of histories to each propositional variable from the set $PROP$.*

DEFINITION 13 (ETL Language). *The language of ETL is generated by the following Backus-Naur form:*

$$\varphi ::= p | \neg\varphi | \varphi \wedge \varphi | K_i\varphi | \langle e \rangle \varphi,$$

*where $p \in PROP$, a set of propositional variables, $i \in Agt$, and $e \in \Sigma$.*

The meanings of the first four formulas are as usual, while the intended meaning of $\langle e \rangle \varphi$ is "event $e$ *can* occur and after its occurrence $\varphi$ is the case". The dual of the $\langle\rangle$-operator, the $[]$-operator, is also defined as usual, that is $[e]\varphi = \neg\langle e \rangle \neg\varphi$. The intended meaning of $[e]\varphi$ is "after event $e$ occurs, $\varphi$ is the case"[6]. This is a very basic language for ETL. A discussion of different languages for describing ETL models can be found in Halpern and Vardi 1989 (24), Fagin et al. 1995 (17) or Halpern et al. 2004 (25). Examples of different ETL languages include languages with common/distributed knowledge operators, languages with branching/linear temporal operators or languages with arbitrary future/past operators (As Halpern and Vardi show these variations may lead to a high complexity of the validity problem).

DEFINITION 14 (ETL Semantics). *Given an ETL model $\mathfrak{H} = \langle \Sigma, H, R_i, V \rangle$ with $h \in H$:*

$$\mathfrak{H}, h \vDash p \text{ iff } h \in V(p)$$
$$\mathfrak{H}, h \vDash \neg\varphi \text{ iff } \mathfrak{H}, h \nvDash \varphi$$
$$\mathfrak{H}, h \vDash \varphi \wedge \psi \text{ iff } \mathfrak{H}, h \vDash \varphi \text{ and } \mathfrak{H}, h \vDash \psi$$
$$\mathfrak{H}, h \vDash K_i\varphi \text{ iff } \forall h' \in H : hR_ih' \Rightarrow \mathfrak{H}, h' \vDash \varphi$$
$$\mathfrak{H}, h \vDash \langle e \rangle \varphi \text{ iff } \exists h' \in H : h \prec_e h' \text{ and } \mathfrak{H}, h' \vDash \varphi$$

The semantic clause for the dual of the $\langle\rangle$-operator is defined as usual:

$$\mathfrak{H}, h \vDash [e]\varphi \text{ iff } \forall e \in \Sigma : he \in H \Rightarrow \mathfrak{H}, he \vDash \varphi$$

## 5. DEL-generated ETL models

**Bibliographical Note:** Developed in van Benthem et al. 2009 (51) and overviewed in Hoshi 2010 (27).

The motivation behind merging DEL and ETL is primarily philosophic. DEL and ETL are two different frameworks for dealing with intelligent interaction. On the one hand, DEL describes agents' epistemic states, and how these change as a result

[6]Note that this does not exclude the fact that $\varphi$ might be the case even before event $e$ occurs.

of informational events. However, it is a common assumption of the framework that, given some consistency constraints, the agents' informational states can change in any way. On the other hand, ETL describes the processes that govern the interactions between agents, but without giving a good account of how agents' informational states are affected by informational events. The results that such a combination between the two different frameworks produces are elegantly presented in Hoshi 2010, (27), pp. 414-415: (i) "[u]sing DEL product update to produce ETL models provides a "bridge" between the two logical paradigms allowing us to investigate their precise relationship"; (ii) "[t]he merged framework generalizes models in DEL. We can then reinterpret the formal language of DEL over this class of generalized models and investigate new logical systems"; and (iii) "[m]odels in the merged framework (...) are powerful tools for studying concrete scenarios of intelligent interactions." A good philosophical introduction to these issues can be found in Hoshi 2010 (27), section 1. Some reasons to the effect that ETL does not have the last word on intelligent interaction are presented in Halpern and Fagin 1989 (23), see especially Theorem 5.1. Wang 2010 (57), section 9.3.2, has a brief but informed presentation of the advantages of reasoning with protocols in DEL vs. reasoning with protocols in ETL.

Let $E$ be the class of all pointed models, $E = \{(\mathfrak{E}, e) : \mathfrak{E}$ an event model and $e \in D(\mathfrak{E})\}$. A history is a finite sequence of event models from $E$. Let $E^*$ be the set of all histories built from elements of $E$. Let $\sigma$ denote an element of $E^*$ (so $\sigma$ is a sequence of pointed event models). I write $\sigma_n$ for the initial segment of $\sigma$ of length $n$ ($n \leq len(\sigma)$) and write $\sigma_{(n)}$ for the $n$th component of $\sigma$. For example, if $\sigma = (\mathfrak{E}_1, e_1)(\mathfrak{E}_2, e_2)(\mathfrak{E}_3, e_3)$, then $len(\sigma) = 3$, $\sigma_{(2)} = (\mathfrak{E}_2, e_2)$ and $\sigma_2 = (\mathfrak{E}_1, e_1)(\mathfrak{E}_2, e_2)$.

Definition 15 (DEL protocol). *A DEL protocol is a set $P \subseteq E^*$ closed under finite initial segments.*

Let $Ptcl(E)$ be the class of all DEL protocols, then

Definition 16 (State-dependent DEL protocol). *Let $M$ be an epistemic model. A state-dependent DEL protocol on $M$ is any function $p : D(M) \to Ptcl(E)$.*

If all states in a model get assigned the same protocol, then that model has a uniform DEL protocol. The reason for distinguishing between state-dependent and uniform protocols is because we want to be able to tell apart models in which the protocol is common knowledge (in which it is uniform) from model in which the agents may have uncertainties regarding which protocol is running (state-dependent DEL protocols). For the rest of this section we will be focusing on state-dependent DEL protocols.

Definition 17 ($p$-generated model). *Given an epistemic model $M = \langle S, R_i, V\rangle$ and $p$, a state-dependent DEL protocol on $M$, the $p$-generated model at level $n$ is $M^{n,p} = \langle S^{n,p}, R_i^{n,p}, V^{n,p}\rangle$ such that:*

(1) $S^{0,p} = S, R_i^{0,p} = R_i, V^{0,p} = V$;
(2) $s\sigma \in S^{n+1,p}$ *iff* $(1) s \in S; (2) len(\sigma) = n{+}1; (3) s\sigma_n \in S^{n,p}; (4) \sigma \in p(s); (5) M^{n,p}, s\sigma \vDash pre(\sigma_{(n+1)})$;
(3) *For each* $s\sigma, t\sigma' \in S^{n+1,p}$ : $s\sigma R_i^{n+1,p} t\sigma'$ *iff* $s\sigma_n R^{n,p} t\sigma'_n$ *and* $\sigma_{(n+1)} \rightarrow_i \sigma'_{(n+1)}$;
(4) *For all* $q \in PROP$ : $V^{n+1,p}(q) = \{s\sigma \in S^{n+1,p} : s \in V(q)\}$.

Now, the way to get from this construction to an ETL model is described in the following definition.

Definition 18 (DEL-generated ETL model). *Given an epistemic model* $M = \langle S, R_i, V\rangle$ *and a state-dependent DEL protocol on* $M$, $p$, *and ETL model,* $Forest(M, p) = \langle H, R'_i, V'\rangle$ *is defined as follows:*

(1) $H = \{h :$ *there is a* $s \in S, \sigma \in \bigcup_{s_i nS} p(s)$ *with* $h = s\sigma \in S^{len(\sigma),p}\}$;
(2) *for all* $h, h' \in H$ *with* $h = s\sigma$ *and* $h' = t\sigma'$, $hR'_i h'$ *iff* $len(\sigma) = len(\sigma')$ *and* $s\sigma R_i^{len(\sigma),p} t\sigma'$;
(3) *for every* $q \in PROP$ *and* $h = s\sigma \in H$, $h \in V'(q)$ *iff* $h \in V^{len(\sigma),p}(q)$.

van Benthem et al. 2009 (51) prove that for an arbitrary epistemic model and state-dependent DEL protocol, $Forest(M, p)$ is indeed an ETL model.

There are a few logics that could be interpreted on these models and the two most obvious represent extensions of PAL and DEL, i.e. TPAL and TDEL. Interpreting public announcement logic and dynamic epistemic logic on these models is not trivial. One significant difference when adding a protocol is that reduction axioms from the dynamic logic to its static counterpart are no longer available. However, since I said that I do not want to talk about syntax in this thesis, I will not explore this issue further. Instead I will define a few operators on these models that will be helpful later in the thesis. These definitions are taken from Hoshi 2009 (26):

Let $X$ be a set of pointed event models, such that if $(\mathfrak{E}, e) \in X$ then $(\mathfrak{E}, e') \in X$, for all $e' \in D(\mathfrak{E})$. The following operators could be defined in TDEL(X):

$$H, h \vDash NEXT\varphi \text{ iff } \exists e \in X : he \in H \text{ and } H, he \vDash \varphi$$
$$H, h \vDash FUTURE\varphi \text{ iff } \exists \sigma \in X^* : h\sigma \in H and H, h\sigma \vDash \varphi$$
$$H, h \vDash BEFORE\varphi \text{ iff } \exists e \in X, \exists h' : h = h'e \text{ and } H, h' \vDash \varphi$$

The intended interpretation of these operators is as follows: "some event can happen such that afterwards $\varphi$ holds", "some sequence of events can happen such that afterwards $\varphi$ holds", and respectively "event $e$ has happened and before it happened $\varphi$ was holding". Of course, duals could be defined in the usual way.

CHAPTER 3

# The surprise examination in dynamic epistemic logic

As briefly mentioned in the introduction, the surprise examination paradox has recently made its way in the dynamic epistemic logic literature. In this section I briefly present the solutions proposed so far, and I expand on the criticism already presented in the introduction.

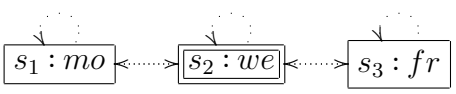
There are two different approaches to the surprise examination paradox in DEL. The first one (also historically) was developed by Gerbrandy 1999. It has recently been restated in Gerbrandy 2007 and van Ditmarsch and Kooi 2008. However all three references discuss the same solution. The second one was put forward by Baltag and Smets in a series of presentations. Despite some differences that exist between the different expositions, the main idea (which I present below) is always the same. It is worth remarking that both Gerbrandy and Baltag and Smets frame the surprise examination paradox in the same way: the school in question is a school who has among its rules that there is an examination every week. This is taken as being an unbreakable rule. Therefore the teacher's announcement is simply that "next week the examination will come as a surprise".

The structure of this chapter is the following: I present briefly Gerbrandy and Baltag and Smets' solutions. At the end of each presentation I first show why their solutions do not respect the criteria for a solution presented in the introduction and then I give a counter-example to their solution, showing that there is at least one variation of the surprise examination paradox that they cannot account for.

## 1. Gerbrandy's solution

Gerbrandy's solution follows the tradition started by Binkley and continued by Wright and Sudbury of seeing the surprise examination paradox as stemming from Moore's paradox. The conclusion of Gerbrandy's analysis is that the teacher's announcement, just like a Moorean sentence, can be announced but it cannot successful.

Gerbrandy models the initial situation of the students, before the teacher's announcement as:



This model however, is not an S5 model, but a K45 one. This means that, for Gerbrandy, knowledge is not factive ($\not\vdash K\varphi \to \varphi$). This interpretation of knowledge is provoked by the paradox he analyzes in which it seems that the students can reach a contradiction and still carry on with their lives: after the students perform their elimination they end up with a contradiction but they still go to school every day until the get the examination and they are surprised by it[1]. From this Gerbrandy derives the idea that knowledge does not need to be factive and that an agent can be surprised when he knows a contradiction.

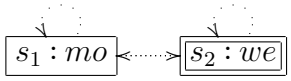
Therefore, the definition of surprise, according to Gerbrandy is

$$S = (mo \wedge \neg Kmo) \vee (we \wedge [!\neg mo]\neg Kwe) \vee (fr \wedge [!\neg mo][!\neg we]\neg Kfr) \vee K\bot$$

This reads: "the students are surprised if (i) the examination comes on Monday and the students do not (K45) know that it comes on Monday; or (ii) the examination comes on Wednesday and after the students learn that the examination does not come on Monday they do not (K45) know that it comes on Wednesday; or (iii) the examination comes on Friday and after the students learn that the examination does not come on Monday and that the examination does not come on Wednesday, they do not (K45) know that it comes on Friday; or (iv) the students know a contradiction". As it is obvious, Gerbrandy identifies predicting with (K45) knowing that it is the case.

The teacher's announcement is construed as the public announcement of S, that is "$!S$". This allows Gerbrandy to model the scenario as: $[!S]KS \vdash [!S]K\bot \wedge [!S]K\bot \vdash [!S]S$. That is: "After the teacher's announcement the students come to know that it is the case that $S$, from which they deduce a contradiction and because they deduce the contradiction, they indeed are surprised." However, formalizing the paradox like this, Gerbrandy is able to show where the problem with this reasoning stands. Although it seems intuitive, it is not the case that $[!S]KS$. In other words, it is not the case that the teacher's announcement is successful. Although it is intuitive that not all sentences can become known after being uttered/written (think of "I have not written this sentence!"), people rarely take this into account. They always assume that once uttered, a sentence can become known. But just as the Moore sentence, the surprise sentence as Gerbrandy formalizes it cannot become known after being uttered.

The effect of the teacher's announcement on the student's initial epistemic state is given by the updated model $M|S$:




[1]From this perspective Gerbrandy may also be seen as continuing the tradition of Kaplan and Montague.

It can be easily checked that the surprise sentence, $S$ does not hold in this model. Therefore, Gerbrandy concludes that the surprise examination paradox is not really paradoxical and that when the scenario is interpreted by means of public announcement logic, it becomes obvious that the error of the students lie in assuming that the teacher's announcement is successful, which under scrutiny, is proven to be wrong. Of course, Gerbrandy does not ignore the fact that some may want a bit more, namely that the teacher's announcement to be interpreted as saying that even the announcement is made, the students will still be surprised. However, Gerbrandy quickly dismisses such an interpretation as being a real (self-reference) paradox[2]

## 2. Evaluating Gerbrandy's solution

I will restate every condition that a solution (in the sense I advocate in this thesis) has to meet, and see whether Gerbrandy's solution indeed meets it.

(1) The solution should make the teacher's announcement satisfiable.

The teacher's announcement according to Gerbrandy is only satisfiable if the teacher does not plan to give the examination on Friday, or on Wednesday, since after the announcement, $S$ is only satisfied on $mo$.

(2) The solution should make it clear that the teacher can carry out the announcement even after he has announced it.

Again, he can do this only if he plans to do it on Monday, for the same reasons as before.

(3) The solution should do justice to the intuitive meaning of the announcement.

This I believe is one of the weakest points of Gerbrandy. According to him, if the teacher's announcement gets the highly intuitive reading that the students should be surprised by the examination even after the teacher's announcement, then the surprise examination paradox is indeed a paradox.

(4) The solution should do justice to the intuitive plausibility of the pupils' reasoning.

This is one of the strongest points of Gerbrandy's analysis: the students reasoning and their mistake get a very clear statement and explanation in his solution.

(5) The solution should make it possible for the students to be informed by the announcement.

[2]In this sense, Gerbrandy seems to follow ... . What these footnotes at Gerbrandy are meant to suggest is that Gerbrandy's solution are by no means new, and in fact, my criticism to his solution, are just a restatement of the criticism brought to ... .

Again, a strong point. It is obvious that the teacher's announcement modifies the student's information (they go from not distinguishing between three states to not distinguishing between two).

(6) The solution should explain the role, in the generation of the puzzle, of the announcement's being made to the students.

Again, a good point for Gerbrandy, since in the surprise sentence, the knowledge operators are indexed with the students, therefore the surprise sentence has the Moorean effect only for the students and not for anybody else (e.g. $K_{parents}S$ is perfectly fine, whereas $K_{students}S$ is not)

(7) The solution should account for the fact that before the teacher's announcement the students are indifferent both with regards to there being an exam and, conditional on there being an exam, with regards to its actual date.

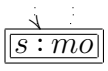
This is not exactly so in Gerbrandy's analysis, since Gerbrandy starts directly with a model in which the students know that surprise will be given. However, I believe that it does not change Gerbrandy's analysis in any way if we were to explicitly introduce the school's announcement that every week has an examination.

(8) The solution should do justice to the intuition that there is something truly interactive in the scenario: after the teacher's announcement, both the teacher and the students learn something about one another.

The teacher, on Gerbrandy's analysis does learn something after he makes the announcement, namely that the students now know that Friday cannot be the day of the examination, and that they will expect an examination on Wednesday. However, in Gerbrandy's analysis, it does not make a significant difference between making the announcement and not making it: after the announcement the only difference is that the model looses its last state. Otherwise, the same problems exist as with the first, initial model.

(9) The solution should be applicable to *all variations* of the surprise examination paradox.

This is again a problematic point for Gerbrandy. Take the 1-day variation of the surprise examination paradox. In this variation, the initial state of the students is:




It is easy to check that $S$ does not hold in this model. Therefore the effect of the teacher's announcement is the elimination of $s$, concluding with the empty domain. Thus: $[!S]K\bot \wedge [!S]K\bot \vdash [!S]S$. In conclusion, Gerbrandy analysis does not amount to anything more than recognizing the paradoxical nature of the 1-day surprise examination paradox.

## 3. Baltag and Smets' solution

Baltag and Smets interpret the surprise examination paradox as a paradox about trust. Their conclusion is that the paradox withers away once one realizes that there is no reason for the students to continue trusting the teacher after the perform the backward elimination argument.[3] Therefore, after they reach contradiction (which Baltag and Smets argue that they do reach as long as the trust the teacher), the students revise their attitude towards the teacher until they can integrate what the teacher announces with the rest of their beliefs. Baltag and Smets prove that such a consistent attitude is indeed possible, and what is more, unique.

The main focus of Baltag and Smets' analysis is on the trust that the students have towards the teacher and that they revise after hearing his announcement. They define different levels of trust that correspond to an agent's willingness to perform a certain upgrade with the information coming from the trusted source. We will refer to these levels of trust by $!, \Uparrow, \uparrow$, where, e.g. $!\varphi$ means that the agent assigns the source (we can also index the symbols, $!_i$, in order to take into account more than one source) of the announcement infallibility. As it is intuitive, an agent will always have some attitude towards the source of an announcement and he will always trust it as much as it can, that is, an agent always applies a principle of charity when interpreting an announcement.

In order to capture this formally Baltag and Smets introduce atomic sentences $!, \Uparrow, \uparrow, MAGM$. [4] that correspond to the agent's attitudes towards the source of the announcements (variable $\tau$ will range over them). The dynamic modality $[*\varphi]$ is interpreted as follows: the transformation $*\varphi$ is an upgrade that reorders each partition cell $[s]$ by applying the corresponding type of upgrade $\tau$, where $s \vDash \tau$. When this reordering is inconsistent for some $\tau$, the worlds that satisfy $\tau$ are eliminated, hence $*\varphi$ cannot be executed in them. On top of this new structure a hierarchy (of obligations) can be added, in the form of a total pre-order, $\precsim$, on the states of $S$, so that:

$$s \vDash O\psi \Leftrightarrow Max_{\precsim} S \subseteq \|\psi\|$$

$$s \vDash O(\psi|\varphi) \Leftrightarrow Max_{\precsim} \|\varphi\| \subseteq \|\psi\|$$

A possible way to define the $\precsim$ relation is by making all $\precsim$-maximal states satisfy $!$, the next best ones $\Uparrow$, and so on. Then, all states will satisfy (the charity principle):

$$O(!) \wedge O(\Uparrow |\neg!) \wedge O(\uparrow |\neg! \wedge \neg \Uparrow) \ldots$$

[3] From this perspective, Baltag and Smets seem to continue the tradition of Quine.

[4] $MAGM$ represent the minimal trust attitude that an agent might consistently have towards a source when in face of revision with higher-order belief sentences. It corresponds to $[*\varphi]\neg K \neg (BEFORE\varphi) \Rightarrow [*\varphi] B(BEFORE\varphi)$

and

$$O(MAGM)$$

Then we can have special actions that lead to a revision of the norms, which semantically corresponds to a specific change (depending on the revision) in the $\precsim$ relation. But the more interesting fact is that regular revisions can lead to revisions of the $\precsim$ relation. This is the case with the infallibility norm ($O(!)$) and the Moore sentence.

Therefore the initial situation from the point of view of the students is

$$\cdots \leftrightarrows we, \Uparrow \leftrightarrows th, \Uparrow \leftrightarrows fr, \Uparrow \rightarrow we, ! \leftrightarrows th, ! \leftrightarrows fr, !$$

where *day* means "the exam is going to take place on *day*", and the arrows represent the plausibility that the students assign to the exam being on *day*.

Now, for the meaning of the teacher's announcement. To begin with, Baltag and Smets construe the fact that the exam will be a surprise as "the evening before the exam day, the students will not believe that the exam is tomorrow"[5]:

$$Surprise = \bigwedge_{we \leq i \leq fr} (i \rightarrow [!(\bigwedge_{we \leq j < fr} \neg j)]\neg Bi)$$

However, Baltag and Smets believe that the teacher means more by his announcement than just "the evening before the exam day, the students will not believe that the exam is tomorrow". He intends that "even after the announcement the exam's date will still be a surprise"[6]. Formally, the teacher's announcement is $*(NEXTsurprise)$.

As the student has to give as much credit to the teacher as it is consistent to give, Baltag assumes the two hierarchies presented above, namely:

$$O(!_{Teacher}) \wedge O(\Uparrow_{Teacher} |\neg!_{Teacher}) \wedge O(\uparrow_{Teacher} |\neg!_{Teacher} \wedge \neg \Uparrow_{Teacher}) \ldots$$

$$O(MAGM_{Teacher})$$

Now, Baltag and Smets prove that $!, \Uparrow$, and $\uparrow$ are not possible attitudes the students might have towards the teacher's announcement in this situation, by showing that $!(NEXTsurprise), \Uparrow (NEXTsurprise)$, and $\uparrow (NEXTsurprise)$ are not compatible with the students knowing that the exam will take place on one of the five days. (I will omit the proof)

[5]Remark that here believe is not a regular Kripke modality, but a modality defined in terms of the maximal worlds in a plausibility ordering. Compare to Gerbrandy who takes 'surprise' to be a K45 (Kripke) modality.

[6]This sentence is self-referential. Compare to Gerbrandy who rejects any form of self-referentiality.

Further, Baltag (2010) proves that there is however a unique revision policy that respects the $MAGM$ norm, for which no contradiction arises. From, $MAGM$ and

$$BEFORE(NEXTsurprise) \Leftrightarrow surprise$$

we can derive that

$$[T]\neg K\neg surprise \Rightarrow [T]Bsurprise$$

However,

$$K(\bigvee_{1\leq i\leq 5} i) \Rightarrow \neg Bsurprise$$

Therefore,

$$[T]\neg K\neg surprise \Rightarrow [T]FALSE.$$

So an upgrade $T$ is executable if and only if $\neg[T]\neg K\neg surprise$ holds. That is if $K\neg surprise$ holds after the upgrade. But this is only if

$$we \longleftarrow th \longleftarrow fr\ ,$$

So the student will know the teacher lied, and the exam will not be a surprise, whenever it comes (even on Monday!).

## 4. Evaluating Baltag and Smets' solution

As in the case of Gerbrandy's solution, my evaluation will be in terms of the 7 criteria presented in the introduction. Therefore,

(1) The solution should make the teacher's announcement satisfiable.

Baltag and Smets' solution imply that after the announcement not only that the students do not believe what the teacher announced, but the actually believe the opposite. Therefore, their solution does not make the announcement satisfiable (they even refer to it as a Moore sentence).

(2) The solution should make it clear that the teacher can carry out the announcement even after he has announced it.

Their solution establishes exactly the opposite: after the teacher's announcement the students will be in such a position that no matter when the examination comes, they will expect it.

(3) The solution should do justice to the intuitive meaning of the announcement.

This is a good point for Baltag and Smets. Their interpretation of the announcement seems very plausible as they take into account its self-referential component.

(4) The solution should do justice to the intuitive plausibility of the pupils' reasoning.

I fail to see how their solution explains the students' reasoning. Their answer would be that the students did not realize that they can revise their trust towards the teacher. This is an unsatisfactory explanation.

(5) The solution should make it possible for the students to be informed by the announcement.

As all of the solutions in DEL, the students are indeed informed by the announcement. But against Sorensen, the way in which they are informed is not that they come to know what the teacher announced but they come to know the opposite. So if Sorensen is right in thinking of this condition as stating that the students need to trust the teacher, then this condition is also failed by Baltag and Smets' solution.

Sorensen's comment regarding this condition stems from his belief that an epistemological ideal source in epistemological ideal circumstances is always reliable. Therefore no correct chain of arguments will end by contradicting something an ideal source has announced. Baltag and Smets' conclusion that the only way in which the students can make sense of the teacher's announcement is by knowing he has lied, implies more or less Quine's conclusion. That is, it implies that *the teacher's announcement is insufficient evidence for the students to know that the exam is going to be a surprise*. Hence, a similar criticism to that of Sorensen's, can be raised against Baltag's solution: nothing in the scenario seems to indicate that the teacher might be deceitful (lying) or that he may be wrong.

(6) The solution should explain the role, in the generation of the puzzle, of the announcement's being made to the students.

As in the case of Gerbrandy (and of all who think that the teacher's announcement is a form of Moore sentence), the teacher's announcement is only problematic to the students since the doxastic operators it contains only refer to the students.

(7) The solution should account for the fact that before the teacher's announcement the students are indifferent both with regards to there being an exam and, conditional on there being an exam, with regards to its actual date.

This is not exactly so in Baltag and Smets' analysis, since they start, just like Gerbrandy, directly with a model in which the students know that surprise will be given. However, I believe that it does not change Baltag and Smets' analysis in any way if we were to explicitly introduce the school's announcement that every week has an examination.

(8) The solution should do justice to the intuition that there is something truly interactive in the scenario: after the teacher's announcement, both the teacher and the students learn something about one another.

Baltag and Smets' analysis has a counterintuitive result. To begin with, the teacher does learn something about the students once he makes the announcement,

namely that the students can no longer be surprised. This is emphasized by Baltag and Smets by drawing the game tree of a game played between the teacher and the students at which at every state, the teacher can either give the examination or not and the student can either learn for the examination or not. They show that the student always has a winning strategy in such a game. However, if this is so, why does the teacher make the announcement in the first place? If surprise was already possible and the announcement makes it impossible, what would be the motivation of the teacher for announcing his intentions? This is a question to which Baltag and Smets do not give an answer.

(9) The solution should be applicable to *all variations* of the surprise examination paradox.

As in the case of Gerbrandy there is something to be said here. There is a very simple counterexample to Baltag and Smets, namely to formulate a variation to the surprise examination paradox in which students always trust the teacher. For example, the teacher could add to his announcement that he is telling the truth. In Baltag and Smets' analysis such an extra condition will generate a real paradox: if the students believe that the teacher is telling the truth, they deduce, via the announcement that there will be a surprise examination, that the teacher is not lying; if they assume that the teacher is lying, i.e. that there will be no surprise examination, a surprise examination will come (because they will not expect it since thinking that it is not possible), and hence the teacher was telling the truth.

CHAPTER 4

# A new look at the surprise examination paradox

## 1. Introduction

In this chapter I will present the constructive part of my thesis, namely my proof that there is a solution to the surprise examination paradox that meets all the intuitive conditions that a solution to the paradox should meet. I begin by specifying what methodology I will be using, what my main question is and the motivations for answering it.

First of all, let me recapitulate where we stand right now: (1) there are some intuitive conditions that a solution to the surprise examination paradox should accommodate; (2) it seems intuitive to think of the surprise examination paradox as involving information exchange (that a surprise examination will be given) between two ideal agents (the teacher and the prototypical student) and, naturally, (3) logicians working on logics of information change (in which DEL occupies, by now, a respectable place) rightfully manifested an interest in the surprise examination paradox; however, (4) the only two attempts of solving the puzzle by means of DEL have some consequences that do not meet the intuitive conditions that a solution to the surprise examination paradox is expected to meet: both attempts imply that the students cannot be surprised after the announcement; and both attempts fail to solve another puzzle which intuitively is just a variation to the surprise examination paradox.

At this point one might be tempted to take one of the following two attitudes towards the fact that the most promising framework for solving the surprise examination paradox has so far failed to vindicate the pre-theoretical intuitions regarding what a solution to the paradox should look like. The first possible attitude is to hold that one of the DEL analyses of the surprise examination paradox is the "correct" one (which Baltag and Smets claim) and to argue that the pre-theoretical intuitions about the right solution to the paradox are misguided.

However, I believe that to dismiss our pre-theoretical intuitions because they do not fit the logic is to all intended purposes the same as not preparing your favourite dish because you do not have the right ingredients, although these ingredients are available at the grocery store near you. Of course, in this example, you could just go to the store and buy the appropriate ingredients. Similarly, if the logic does not fit

your intuitions you should try to modify your framework so as to meet them. The only problematic part of this analogy is whether or not it is possible to come up with a modification of the DEL frameworks which meets all the intuitive conditions (i.e. that there are ingredients at the grocery store). This chapter is dedicated to proving that it is indeed possible.

Therefore, the second attitude, which I will adopt in this chapter is that of trying to find within DEL the right interpretation for surprise and for the teacher's announcement so that the pre-theoretical intuitions (behind the conditions spelled out in the introduction) are vindicated.

The rest of the chapter is organized as follows: first I explore the variety of logical formulas that might be suitable candidates for the notion of surprise. I then argue that in fact, out of the plethora of such logical formulas only some seem to meet our intuitions and I defend a certain formula. I turn my attention to the meaning of the teacher's announcement and show that DEL can nicely capture this meaning. I end up by proving that the definition of surprise coupled with the interpretation of the teacher's announcement give us an interpretation of the surprise examination puzzle on which the intuitive conditions of a solution to the surprise examination paradox are met.

## 2. The notion of "surprise"

As already mentioned in the introduction, the surprise examination paradox has a long history and many solutions have been proposed. However, despite the (sometimes) huge differences between the various attempts to solve this paradox, most solutions have one thing in common, namely the way in which they define surprise:

$$Surprise ::= \varphi \wedge [\,]\neg \Box \varphi$$

The way the formula above is understood varies between authors. Nevertheless it is always the case that there is a (finite) sequence of elements (the fact that it is a sequence and not just a set will become relevant shortly) and $\varphi$ says that an element $i$ from that sequence has a certain feature. For instance, the elements have been taken to be days (e.g. Quine 1958, Wright and Sudbury 1977), Robinson's helpers (e.g. Sorensen 1988), random numbers (e.g. Sorensen 1988), hours in a day (e.g. O'Connor 1948), etc. The feature that $\varphi$ attributes to the $i^{th}$ element of the sequence has been taken to be an examination (e.g. Wright and Sudbury 1977), a hanging (e.g. Quine 1953), a blackout (e.g. O'Connor 1948), etc. That is, $\varphi$ can say that day $i$ is an examination day, or a day on which a hanging will take place, or a time at which a blackout will occur. $\Box$ has been taken to stand for knowledge (e.g. Williamson 2000), belief (e.g. Baltag and Smets 2010), safe belief (e.g. Baltag 2003), intention (e.g. Sorensen 1988), a knowledge predicate (e.g. Kaplan and Montague

1970), etc. In any case, the structure of the sentence has almost always been the same; even when surprise was not formalized, the informal explanation followed the same structure. The dynamic operator, "[]", means that all the elements in the sequence up to $i$ cannot have the property expressed by $\varphi$. This dynamic operator should be interpreted as a public announcement which eliminates all elements of the sequence up to $i$. For example, if the elements of the sequence are days of the week or hours in a day, then their order in the sequence is the chronological one. Therefore $[]\neg\Box\varphi$ means that after all the days of the week or all the hours in a day up to the $i^{th}$ have lost the possibility of having the feature that $\varphi$ now attributes to $i$, the agent will not believe that $i$ has that feature. This way of expressing the passing of days, hours, etc. through a dynamic operator represents one of the contributions DEL has made to the literature on the surprise examination paradox. Other authors simply index the regular belief or knowledge operators. For instance, Chow 1998 (15) writes $Ka$: "on the eve of the first day the students know ...", see p. 47.

In this section I will not be concerned with what $\varphi$ stands for, or with what exactly is the sequence that I mentioned in the last paragraph, but only with ways of altering the structure of the sentence expressing surprise and of varying the definition of the $\Box$. By "the structure of the sentence expressing surprise" I mean the string of logical connectives and operators that appear in that sentence. In the case of the standard way of defining surprise, the structure of the sentence is: proposition, and, dynamic operator, negation, a modal operator, proposition (same as before). It seems obvious, I would say, that the conjunction has to be where it is, after all surprise cannot emerge simply from our belief in something or from something being the case. It needs some sort of clash to come about, and the conjunction intuitively connects the two things that clash, thus generating the surprise. Furthermore, it also seems intuitive that we have the same proposition on both sides of the conjunction, with a negation in front of it on one side; after all, for the clash to occur there has to be a tension of some sort and this is between $\varphi$ and its negation. This restricts the ways in which the sentence can vary (for example, we cannot eliminate the negation or replace the conjunction with a disjunction, say). Therefore, I will explore some of the remaining different ways in which the structure of the surprise sentence can vary. I will argue against some types of variations and finally conclude that there is *at least* one way of construing surprise which is not only intuitive but also makes it possible to come up with a solution to the surprise examination paradox that respects the constraints on a solution to this paradox discussed in the introduction.

It is important to remember that the purpose of this analysis of the different ways of defining surprise is to show how surprise could be formalized so that we get a solution that meets all of our pre-theoretical intuitions. Therefore, the purpose in this section is not to exhaust all the possibilities and defend only one, but rather the purpose is to familiarize the reader with the space of possibilities. I will, then, only

investigate a few ways of construing surprise. However, if one is not pleased with the definition I will finally argue for, and wishes to argue for another one, as long as that new definition also meets the pre-theoretical intuitions, I have nothing against it. The rationale is that, as I already mentioned in the introduction, the most important conditions to be met by a solution to the surprise examination paradox are the strong conditions and not the lax ones; formalizing surprise is a lax condition: it is dependent on our intuitions, and these can differ. So, as long as another interpretation of surprise (which has some intuitive support) leads to a solution that meets all of the strong conditions, it represents, in my view, an equally good interpretation of surprise. To wit, we are not after *the* solution to the surprise examination paradox, but after *one* solution.

The common structure of surprise is taken to be: $\varphi \wedge []\neg \Box \varphi$. Here are a few ways in which we can change this definition. I divide them in three categories. The first type of variation focuses on the second conjunct. The second type of variation focuses on the first conjunct, while the third type focuses on the exact interpretation of the $\Box$ operator.

The second conjunct could look in any of the following 6 ways: $[]\neg \Box \varphi$, $\neg[] \Box \varphi$, $\neg \Box []\varphi$, $[] \Box \neg\varphi$, $\Box[]\neg\varphi$, $\Box\neg[]\varphi$.

The first conjunct could look in any of the following 5 ways: $\varphi$, $\Box\varphi$, $[]\varphi$, $[] \Box \varphi$, $\Box[]\varphi$.

The meaning of the $\Box$ operator can be any of the following 5: (i) if defined on a Kripke frame: S5, KD45, K45[1]; (ii) if defined on a plausibility frame: maximal, safe belief. Of course, this is again just part of the story, as one can very well define the operator syntactically; this is especially interesting if the point would be to analyze real-life situations in which the agents are not ideal (see Lorini and Castelfranchi 2007: the *Test* operator *below*).

In between the options I have presented above (which again are just a small part of a larger number of available options, and we'll see below one more type of variation) we have 510 structurally different ways of construing the surprise sentence. Each form of the second conjunct is combined with every form of the first conjunct and every definition of the $\Box$ operator. However, when both conjuncts contain a $\Box$, it is by no means necessary that both operators are of the same type. That is: six variations of the first conjunct times three variations of the first conjunct that contain a $\Box$ times 25 (the number of distinct pairs formed from two types of $\Box$ operators, out of the five we consider here). This means 450 variations. To this we add six variations of the second conjunct times two variations of the first conjunct (those not containing a $\Box$ operator) times five different $\Box$ operators. This means 60 variations. In total, there are 510 different syntactic variations. Of course, some of these ways are not

[1]Although of course, one could explore weaker operators as well.

logically correct. Nevertheless, most of them represent logical correct combinations and eliminating them depends solely on their capacity of expressing the intuitive notion of "surprise". Let us explore a few variations:

**1.** $\varphi \wedge [] \neg \Box \varphi$ is the standard way of defining surprise. However, this conception of surprise is too wide. In other words it over-generates examples of surprising situations. Take the following situation: I believe (I will be ambiguous with respect to the exact definition of belief I am using) neither that Harry Potter 7.1 is playing at the Pathe cinema in Haarlem, nor that it is not playing (obviously, in this case, I do not believe that it is playing, either). Visiting Haarlem, I notice that it is actually playing. Will I be surprised? I hardly think so, as I was entertaining the possibility that it was playing. The problem with this way of construing surprise is more far-reaching than just finding out out what movies are playing in Haarlem. On this account of surprise it would mean that we cannot learn anything new without being surprised, since learning something new presupposes that before learning it one did not believe it. That is, if we learn for the first time a true fact $\varphi$, after $\chi$ happens (whatever $\chi$ means), obviously $\varphi \wedge [\chi] \neg \Box \varphi$ holds. It should be intuitive that this extends to all the five interpretations of the $\Box$ operator mentioned.

This suggests that if the first conjunct is $\varphi$, then out of the six different variations of the second conjunct, only $[]\Box\neg\varphi$, $\Box[]\neg\varphi$ and $\Box\neg[]\varphi$ represent suitable components to a definition of surprise. Having the negation outside the scope of the $\Box$ operator would make surprise trivial. But such a conclusion puts constraints on the type of operator. Consider the following interpretation of surprise.

**2.** $\varphi \wedge [] \Box_{S5} \neg\varphi$ denotes an impossible situation. If, indeed $\varphi$ is the case, then no agent can (S5) know that $\varphi$ is not the case. (public announcements cannot change this fact!)

Hence, if the first conjunct is $\varphi$, the $\Box$ in the second conjunct cannot be a S5 operator. But then, if the first conjunct is $\varphi$, as most philosophers interpret it, surprise cannot be a (S5) knowledge-based phenomenon, but it has to be a belief based one. But this is not a very strong conclusion since interpreting the first conjunct as $\varphi$ is strange. The reason is that in similar situations, learning a lie should be as surprising as learning the truth. Take the following example: you believe that receiving an exam on Monday is impossible, but, during the class on Monday the teacher tells you that in fact Monday is the exam day and he gives you a test. However, at the end of the class, he tells you that it was just a joke, and that you were initially right, Monday is not the "real" examination day. So the test you just took will not be graded, but you will receive another "real" test in one of the other days of the week. It is obvious that $\varphi$ (with the meaning that the examination is on Monday) is not the case, but that you still were surprised. Therefore, $\varphi$ being the case, as well as you knowing that $\varphi$ are both two strong formulation of the first conjunct. What is required is that you believe that the examination in on a certain date. It is not

necessary for the examination to actually be on that date for you to be surprised! All of the definitions of belief mentioned above work equally well.

In consequence, the first conjunct has to be one of the $B\varphi$, $[]B\varphi$, $B[]\varphi$. However, if we interpret the first conjunct as $[]B\varphi$ then after the elimination of the days prior to $\varphi$ one of the following situations could obtain (here it is relevant to distinguish between different types of belief):

**3.** $[](\Box_{max} \wedge \neg \Box_{S5} \varphi)$ This represents a consistent formula. However, it fails to capture any intuitive notion of surprise. It merely says that the situation that you consider most likely to be the case is one in which the exam is on day $i$ ($\varphi$ holds). In addition to this, you do not know that the exam is on $i$, so you also consider possible days in which the exam is on another day, other than $i$. This is a perfectly plausible situation. But if this were to mean that we are surprised, then, again, surprise would become a trivial attitude. Take the following example: after some announcement, call it $\chi$(and think of it as the elimination of the days prior to $i$), I come to believe that most likely the world is in such a way that the examination is on Wednesday. However, I also come to believe that it is possible for the examination not to be on Wednesday. Am I surprised after learning $\chi$? I would say not. Most will argue that whenever you learn something, it is rational to also entertain the possibility that it is false. So this scenario is by no means surprising. But if the first conjunct cannot be $\varphi$, what does this tell us about the second conjunct? The previous conclusion is no longer relevant. Consider, then the following formulation of surprise.

**4.** $\varphi \wedge []\Box_{K45} \neg\varphi$ is obviously possible from a logical point of view, but by no means a good definition of surprise in our case. We are after a definition of surprise for ideal agents, who, among other things, are supposed to be perfect logicians. However, if their beliefs can be rendered by a K45 operator, then they should have no problems entertaining contradictions. But this is something that perfect logicians would like to avoid. Perfect logicians (to use the vernacular of Baltag and Smets) "go crazy" when they end up believing a contradiction. So, the operator in the second conjunct should be either a KD45 operator, or a maximal operator, or a safe belief operator.

**5.** $[]\,\Box_{S5}\,\varphi \wedge []\,\Box\,\neg\varphi$ denotes an impossible situation, whatever the interpretation of the second $\Box$ operator is. By the rules of public announcements this sentence translates to $[](\Box_{S5}\varphi \wedge B\neg\varphi)$, which can never happen (because of intuitive relations that have to hold between knowledge and belief: if one knows that $\varphi$, he cannot believe $\neg\varphi$).

Of course, there many more variations that could be attempted. But the purpose here was not to go through all of them, but rather to show a sample of how a logician could play with the traditional way of defining surprise. However, as promised in the first part of this chapter, I will now show another way of varying the definition of surprise, which is a bit different that all the previous variations (even from those

not spelled out) and which, in addition, will lead us towards a new solution to the surprise examination paradox.

***

I will begin by remarking that surprise appears to be an attitude that always accompanies an act of learning. If you sit on your couch and you see nothing new, or you make no inferences in your head, etc. you cannot get surprised. Therefore, surprise is always an act of learning a certain type of new fact. Remember the temporal operators defined in chapter 2, i.e. $BEFORE(\varphi)$. They were defined on DEL-generated ETL frames and they were interpreted at one level of the model by reference to the previous level. The idea they can formalize is that you learn something new if you now believe or know something that you did not believe/know before. As already obvious from the previous analysis, I favour some particulars about the definition of surprise: namely that it has to be a completely belief-based phenomenon (you can be surprised with false information, and while holding false beliefs); (thus) the first conjunct should be a doxastic formula as well; and the negation in the second conjunct should be inside the scope of the belief operator in order for surprise not to become trivial. Again, there may be other ways of intuitively defining surprise, but the purpose of this chapter is to show that there exists at least a way of vindicating our pre-theoretical intuitions about the surprise examination paradox. From now on, in order to be more concrete, I will define the $\Box$ operator as a maximal belief operator on an epistemic-doxastic frame, and I will write $B$ for that. So, think about the following definition of surprise:

$$B\varphi \wedge BEFORE(B\neg\varphi)$$

This reads: "you now believe that $\varphi$ but at the last step you believed that $\varphi$ cannot be the case". The state at which this sentence is true is the state at which you learned that $\varphi$. The $BEFORE$ operator allows us to identify the state at which the act of learning occurred. Moreover, this is a special type of learning: you learned that something is the case, while before you believed it to be impossible. This seems to be a good way of thinking about surprise. Take the following scenario: I believe that Harry Potter 7.1 is not playing at the Pathe in Haarlem (say because I believe that the cinemas in Haarlem only show Dutch movies and Harry Potter 7.1 is not Dutch), I go to the Pathe in Haarlem and I learn that it is actually playing there. I believe that this would really represent a surprising fact since something that I was expecting not to happen, actually happened, contradicting my expectations. In consequence, learning new information is surprising only if that information seemed impossible before learning it. I believe that this is indeed, a very plausible definition of surprise and in the next section I will defend it by comparing it to some recent work

done at the intersection of logic and cognitive science on the notion of "mismatch-based surprise".

## 3. A caveat: the cognitive structure of surprise

Lorini and Castelfranchi (2007) identify several types of surprise, out of which I will only present the one that seems to apply to the surprise the students in the scenario of the surprise exam experience:

> **Mismatch-based surprise** (given the conflict between a perceived fact and a scrutinized representation). The cognitive configuration of *mismatch-based surprise* relative to the mismatch between a perceptual datum $\psi$ and a scrutinized representation $\varphi$ is defined by the following facts:
> (1) $\psi$ is the agents perceptual datum;
> (2) $\varphi$ is the representation scrutinized by the agent; and
> (3) the agent believes that $\varphi$ and $\psi$ are incompatible facts.
> Formally,
>
> $$MismatchSurprise(\psi, \varphi) =_{def} Datum(\psi) \wedge Test(\varphi) \wedge Bel(\psi \rightarrow \neg\varphi)$$

In order to really understand this definition it is necessary to introduce some of the technical apparatus of Lorini and Castelfranchi's[2].

$Datum(\Phi)$ means that $\Phi$ is a datum perceived by an agent. $\Phi$ is just "some piece of information gathered by the agent's sensors which is a candidate for becoming a belief of the agent."[3] $Datum(\Phi)$ is true at a world $w$ iff $DATA(w) = \Phi$, where $DATA$ is a function from worlds to the set of propositional formulas, which assigns to each world $w$ the datum obtained by the agent's sensors at world $w$.

$Bel(\Phi)$ means that the agent believes that $\Phi$, and $Bel$ is a K modal operator, that is:

(1) All instances of propositional tautologies, and Modus Ponens
(2) $\vdash Bel(\Phi \rightarrow \Psi) \wedge Bel(\Phi) \rightarrow Bel(\Psi)$
(3) $if \vdash \Phi, then \vdash Bel(\Phi)$

$Bel(\Phi)$ is true at a world $w$ iff $\Phi$ is true at all the worlds in $B(w)$, where $B$ is a function from the worlds to their powerset, which assigns to each world $w$ the alternative worlds that the agent considers possible at $w$.

$Test(\Phi)$ means that $\Phi$ is the representation that the agent is scrutinizing. $\Phi$ is "the representation on which the agent is focusing its attention and that the agent

[2]To this end I will present selections from Lorini and Castelfranchi (2007), especially section 2: pp. 136-139.

[3]Lorini and Castelfranchi (2007), p. 136.

matches with the perceptual data"[4]. $Test(\Phi)$ is true at a world $w$ iff $TEST(w) = \Phi$, where $TEST$ is a function from worlds to propositional formulas which assigns to each world the represenatation that the agent is scrutinizing at that world.

If $Test(\Phi)$, that is if an agent has a scrutinized representation that $\Phi$, or in simpler terms, if an agent expects $\Phi$ then obviously $\Phi$ is believed to be true by that agent. However,

$$\nvdash Test(\Phi) \rightarrow Bel(\Phi),$$

since agents are rarely aware of the consequences of the beliefs they focus on at a certain moment, whereas $Bel$ is closed under known entailment (the K axiom). The formal counterpart of this is that $Test$ is defined syntactically by the $TEST$ function, whereas $Bel$ is defined semantically. In consequence, $\nvdash (Test(\Phi) \wedge (\Phi \leftrightarrow \Psi)) \rightarrow Test(\Psi)$, but $\vdash (Bel(\Phi) \wedge (\Phi \leftrightarrow \Psi)) \rightarrow Bel(\Psi)$. So, Lorini and Castelfranchi use $Test$ to designate explicit beliefs[5], defined as "the set of beliefs that the agent can use to make inferences and which is not closed under classical inference."[6]

Therefore (mismatch-based) surprise is "a belief-based phenomenon (...) based on an actual or potential prediction formulated on the basis of the other beliefs"[7] So the definition of mismatch-based surprise would read: "surprise emerges if one explicitly believes $\varphi$, but observes that $\psi$, which he believes to imply $\neg\varphi$".

First of all, remark the two significant similarities between Lorini and Castelfranchi's definition of mismatch-based surprise and my definition of surprise. The first similarity concerns the first conjunct: $Datum(\varphi)$ is the datum observed *by an agent*. This strengthens my argument that the first conjunct cannot simply be a fact; the agent has to be brought in the picture! Of course, the way Lorini and Castelfranchi are bringing the agent in the picture is different from the way I do it. They have a syntactic operator, whereas I use a maximal belief operator. Moreover, they are taking a somewhat objective perspective, since the focus is on some fact being observed by the agent, while my focus is on the fact that the agent came to believe some fact (subjective). For this reason, I believe that my definition has some advantages over theirs. They can only express situations in which the agents observe the fact that generates the surprise. My way of defining surprise does not have this constraint, since I do not focus on the act of gathering the information that triggers the surprise, but on the fact that the agent comes to believe that information. But an agent can come to believe something also through testimony, not only through observation. So, in this sense, my framework is more general: it accommodates both learning through testimony and through observation, as long as the effect is that

[4]Lorini and Castelfranchi (2007), p. 137.
[5]See note 13, p. 141.
[6]Lorini and Castelfranchi (2007), p. 141.
[7]Lorini and Castelfranchi (2007), p. 135.

of coming to believe some fact. This makes my framework more applicable with respect to the way in which the surprise-generating information is gathered by the agents. However, in another sense, their definition is more general: their syntactic operator can accommodate both ideal and resource-bounded agents, which makes their framework more widely applicable with respect to the types of agents.

The second similarity concerns the second conjunct: the negation that helps in generating the tension between the first and second conjunct is inside the belief operator. The similarities are greater than that, I believe. The meaning of the $Test$ operator is that of having the expectation that $\psi$ will happen. However, expecting $\psi$ to happen after observing $\varphi$ (and while believing that the two are incompatible) does not make sense. So, although not clear from their analysis, I believe that they would agree with my observation that the expectation expressed by the second conjunct was in the past, i.e. before learning that $\varphi$. However, there is a difference between my understanding of the second conjunct and theirs. They define the belief expressed by the second conjunct as a syntactic operator, which extends the range of application for their definition to non-ideal agents. The fact that they define the expectation not as a belief, but as a syntactic operator is the reason why they need a third conjunct, i.e. to express the contradiction between the perceived fact and their expectation.

Lorini and Castelfranchi's motivation for investigating a definition of surprise for resource-bounded agents is to formulate a theory of belief revision for resource-bounded agents. Since this type of agents cannot spend too much time revising their beliefs, they have to choose when a revision is essential and when not. Lorini and Castelfranchi propose the surprising situations as triggering the need for performing a revision of an agent's beliefs. Their idea could represent a further motivation for studying the surprise examination paradox: the surprise examination paradox is obviously a paradox that involves the notion of surprise. Clarifying the paradox and especially if this is done by changing our understanding of surprise (as I suggest in this thesis), would have great consequence on the philosophy behind belief revision.

In conclusion, Lorini and Castelfranchi's analysis, which has the advantage of being empirically informed as they come up with it by studying the Cognitive Science literature on surprise, supports the way in which I define surprise. Moreover, my definition can help their framework be extended to testimony-generated surprising situations. These connections would be worth pursuing further, but I postpone them for another time. For now, I will return to the teacher's announcement.

## 4. The teacher's announcement

Let us return to the variation of the surprise examination paradox in its most common form. First, the students are indifferent between there being an examination in the following week and not. That is,

$$M: \qquad \boxed{s_1 : mo} \prec\!\!-\!\!\succ \boxed{s_2 : we} \prec\!\!-\!\!\succ \boxed{s_3 : fr} \prec\!\!-\!\!\succ \cdots$$

The meaning of $mo$, $we$ and $fr$ is "the examination is scheduled for Monday", "the examination is scheduled for Wednesday", and respectively "the examination is scheduled for Friday". The teacher makes his announcement: "next week you will receive a surprise examination!" Although superficially very simple, the exact meaning of this announcement is actually more complicated. I believe there are three distinct parts to this announcement. The first part of the announcement (A) is an existential claim: one of $mo$, $we$ and $fr$ is the case. Since I am after trying to have a solution that meets the condition spelled out in the introduction, I will interpret the teacher as a source highly trusted by the agents. This naturally implies that after the announcement of part (A), the students have to believe that (A), that is the students have to believe that there will be an exam the following week. There are more ways of ensuring that, but I will opt for interpreting (A) as a hard information (although one could interpret it as a soft information as well, as long as after the upgrade prompted by (A), the students end up believing that there is going to be an announcement in the following week):

$$M|(mo \vee we \vee fr): \qquad \boxed{s_1 : mo} \prec\!\!-\!\!\succ \boxed{s_2 : we} \prec\!\!-\!\!\succ \boxed{s_3 : fr}$$

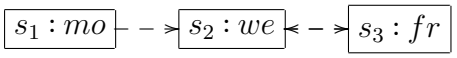
The second part of the announcement, (B), is a possibility claim about surprise: "a surprise should be possible". Remember the last definition of surprise defined in the previous section, and call it *surprise*

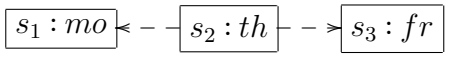
$$surprise ::= B\varphi \wedge BEFORE(B\neg\varphi)$$

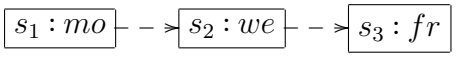
In other words, the meaning of (B) is: it should be possible that when the students learn (come to believe) the exact date of the exam it is the case that previously they had believed that the exam cannot be on that date. First of all, remark that $M|(mo \vee we \vee fr) \not\models surprise$. Therefore, in order for the teacher's announcement to work in such a way that the students come to believe that they could receive a surprise examination (which would meant that they trust the teacher with respect to the second part of his announcement), the teacher's announcement has to convey the information that the students are in such a situation such that when they learn when the exam comes it is possible that they will not have expected it. In other words, the second part of the teacher's announcement has to be such that after it, the students come to know that it is possible they will be surprised: $KFUTUREsurprise$ ("the students know that there exists a sequence of events after which surprise holds"). In other words, given that we model the second part of the teacher's announcement as an event model $(E, e)$: $M|(mo \vee we \vee fr) \models [E, e]KFUTUREsurprise$ ("after the teacher's announcement the students know that they will be, at some point in the

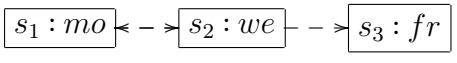
future, surprised"). Remember that, although with the introduction of the temporal operators we are already working in a DEL-generated ETL model, the protocol running is the trivial protocol "anything goes!".

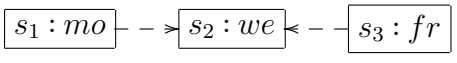
Therefore, we can think of the second part of the teacher's announcement as an event model $(E, e)$ which satisfies the condition that after the product update between $M|(mo \vee we \vee fr)$ and $(E, e)$ the students know that even if they are not surprised then, they could be surprised at some point in the future (there exists a sequence of events that leads to a state in which *surprise* holds). If the initial model is $M|(mo \vee we \vee fr)$, then there are 7 ways in which the students information could be changed by the act performed by the teacher so that the intended effect to obtain (let $exam ::= mo \vee we \vee fr$):

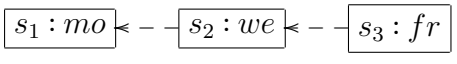
$(M|exam) \otimes (E_1, e_1)$ : 

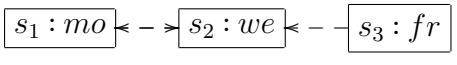

$(M|exam) \otimes (E_2, e_2)$ : 


$(M|exam) \otimes (E_3, e_3)$ : 

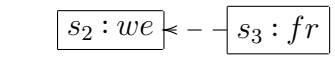

$(M|exam) \otimes (E_4, e_4)$ : 


$(M|exam) \otimes (E_5, e_5)$ : 


$(M|exam) \otimes (E_6, e_6)$ : 


$(M|exam) \otimes (E_7, e_7)$ : 


It is easy to check that in all of these 7 models, surprise could at some point emerge, that is, the students will at some point come to believe that the examination is going to be on a certain day, while before learning that they had believed that it was impossible for the examination to be on that day. For instance take $(M|exam) \otimes (E_7, e_7)$: after Monday passes with no exam taking place, the students update with $!\neg mo$, which leaves them with

$((M|exam) \otimes (E_7, e_7))|\neg mo$ : 


In this new model, the students will believe that the exam is on Wednesday, and they will believe that an exam cannot be on Friday. However, when Wednesday passes without an exam being given, the students will come to believe that $fr$, and before coming to believe that $fr$ they had believed that $fr$ cannot be the case. Therefore, surprise is indeed possible if the teacher's action is construed as the event model $(E_7, e_7)$, and if $!\neg mo; !\neg we$ occurs.

However, *surprise* is by no means necessary yet. There is nothing that prevents the teacher (so far) from giving an examination on a date predictable to the students. For instance, given that the trivial protocol is running, $(M|exam) \otimes (E_1, e_1) \vDash KFUTURE\neg surprise$, that is "there is a sequence of events such that if the students interpret the teacher's announcement as $(E_1, e_1)$, they will not be surprised by the examination". That sequence of events is $!\neg mo$. Of course, this is very far away from the meaning of the teacher's announcement and this is why the teacher's announcement has a third part, (C).

The third part of the teacher's announcement is a procedural claim: it tells the students how the teacher will act. The effect of this third part will be not on the number of states the students entertain as possible (as (A) was) or on the plausibility relations of the students (as (B) was), but on the protocol that the students think is running. The third part of the announcement informs the students that the teacher will act in such a way so that the students receive a surprise examination, so the worry discussed in the last paragraph dissolves. In other words, after the teacher's announcement (mark the effect of this third part on the students initial model, $M$, by $*$), the students will know that $\neg FUTURE\neg surprise$:

$$((M|exam) \otimes (E_i, e_i))^* \vDash \neg FUTURE\neg surprise$$

That is, "after the teacher's announcement all sequences of events will lead to the students being surprised". If the necessity of (A) should be obvious to everyone, the necessity of both (B) and (C) might be more obscure. The motivation behind them is the following: if only (C) were the case, then, given that the initial model of the students was $M$, then the teacher would not have been able to give a surprise examination, since it would not have been possible to surprise the students (according to *surprise*). The outcome of only holding (B) was already explored. Thus (A), (B) and (C) represent the three different parts of the teacher's announcement.

However, there is something unpleasant with saying that the teacher's action can be interpreted as any of these 7 models. First remark that it is obvious that the students know what surprise means. So, when the teacher announces that the students will receive a surprise examination, he does not only convey that the students have to be in a certain state so that they can indeed be surprised (B), but also that he will act in such a way so that the students are surprised (C). Therefore, after the teacher's announcement the students also know that the examination will come only in those days that they consider impossible.

With respect to this idea, the 7 models above divide into two categories. On the one hand there are $(M|exam) \otimes (E_1, e_1)$, $(M|exam) \otimes (E_2, e_2)$, $(M|exam) \otimes (E_4, e_4)$, $(M|exam) \otimes (E_6, e_6)$, $(M|exam) \otimes (E_7, e_7)$, and on the other $(M|exam) \otimes (E_3, e_3)$ and $(M|exam) \otimes (E_5, e_5)$. What distinguishes these two classes is that in the former

there is only one world in each model on which if the exam were to come, the student would be surprised, i.e. $mo$, $we$, $we$, $fr$ and $fr$, respectively. On the other hand, in the latter class there are two worlds in each model, i.e. $mo$ and $we$, and $mo$ and $fr$, respectively. The problem with the first set of models is that they have such a structure that the students are able to predict from their beliefs and from the fact that the teacher knows their beliefs that the exam will come on the day they consider impossible. Since there is only one such day, then the students are able to predict that at the next round they will learn the exact day of the examination. But this does not really seem really surprising.

Therefore, I believe that a small proviso needs to be added to the definition of surprise, namely the proviso that before coming to believe that the exam is on a certain day, the students should not be able to predict that they will learn that. That is, the new and final definition of surprise is:

$$SURPRISE ::= B\varphi \wedge BEFORE(B\neg\varphi) \wedge BEFORE(\neg BNEXT(B\varphi))$$

It reads "an agent is surprised by $\varphi$ at a state if he believes $\varphi$, but before that state he believed that $\neg\varphi$ and moreover, before that state, he was not expecting to come to believe $\varphi$". It is now easy to check that according to this definition of surprise, the first set of models do not satisfy the intuitive condition that after the announcement the students still have to be surprised, whereas the second set of models do.

$$((M|exam) \otimes (E_1, e_1))^* \not\models K\neg FUTURE(SURPRISE)$$
$$((M|exam) \otimes (E_2, e_2))^* \not\models K\neg FUTURE(SURPRISE)$$
$$((M|exam) \otimes (E_4, e_4))^* \not\models K\neg FUTURE(SURPRISE)$$
$$((M|exam) \otimes (E_6, e_6))^* \not\models K\neg FUTURE(SURPRISE)$$
$$((M|exam) \otimes (E_7, e_7))^* \not\models K\neg FUTURE(SURPRISE)$$

Whereas,

$$((M|exam) \otimes (E_3, e_3))^* \models KFUTURE(SURPRISE)$$
$$((M|exam) \otimes (E_5, e_5))^* \models KFUTURE(SURPRISE)$$

So, let us return to the main question of this section: what is the teacher's announcement? Let us first think of what the teacher's announcements does: it lets the students know that there is an examination in the following week; it places the students in a state in which they can be surprised and it informs the students that the teacher will surprise them. Therefore, the teacher's announcement contains

three components: a public announcement of *exam*; an event model that places the students in a state in which they can be surprised; and the protocol that the teacher will follow from that moment onwards (i.e. (A)+(B)+(C)). I will not formalize the teacher's announcement here since the teacher's announcement would have to contain a way of conveying a protocol. This is yet a very new area in DEL and only little research has been done, see Wang 2010 (57) and Gosh 2010 (22).

Before moving forward to assessing this analysis of the surprise examination paradox against the intuitive conditions presented in the introduction, I would like to make three comments about this way of interpreting the teacher's announcement. The first comment is that, in principle, the teacher can now give the exam on all days of the week. However, he cannot give the exam whenever he wants; he has to take into the account whether the students have chosen $(E_3, e_3)$ or $(E_5, e_5)$. So the teacher has to be aware of how the students understood his announcement. This can easily be taken care of by stipulating that there exists a certain convention between the students and the teacher that together with the teacher's announcement makes it clear to the teacher how the students will revise their beliefs. Or it might be the case that the teacher can rely on past experience or simply that he has some sort of privileged access to the students' reasoning.

Secondly, this solution to the surprise examination paradox makes this paradox a true interactive paradox. The fact that the teacher makes the announcement to the students makes the the students possible to be surprised. That is, before the teacher's announcement it was not the case that the students would be surprised. The reason is that, before the teacher's announcement, the protocol governing the scenario was the trivial protocol: everything was possible. This means that an unfolding of the scenario with the teacher giving the examination in one of the days the students were foreseeing as a examination day could have happened. Moreover, the fact that the students trust the teacher makes it possible for them to be surprised. If the person making the announcement was not trusted by the students, then they would not revise their beliefs so that they get in a position to be surprised.

Thirdly, and very importantly, it is almost unanimously accepted that the paradoxical feeling that the surprise examination paradox has relies on the teacher's announcement. My analysis of the teacher's announcement as being three distinct actions sheds more light to this idea. I believe that the problematic and seemingly contradictory parts of the teacher's announcement are (B) and (C), that is the possibility claim that the students should be possible to be surprised and the procedural claim saying that the teacher will act so that he will surprise the students. The problematic part is easy to see: if the students are told by what procedure the teacher will try to surprise them, how can they be surprised when the teacher fulfills his own announcement? I agree that this seems contradictory, but I hope the earlier analysis established how this can happen. Namely, there are models in which the

students can be surprised on more than one day, and thus although the students will know that they will learn that the examination is on either of the possible days, they cannot know what exactly they will come to learn.

Let us now return to the criteria that a solution needed to respect:

(1) The solution should make the teacher's announcement satisfiable.

I have shown that it is satisfiable, by given two distinct models in which it is true, namely $(M|exam) \otimes (E_3, e_3)$ and $(M|exam) \otimes (E_5, e_5)$

(2) The solution should make it clear that the teacher can carry out the announcement even after he has announced it.

I have shown that not only can the teacher carry out the announcement (the students indeed are surprised), but that he can only carry it because he makes the announcement that he makes. Before the announcement, the students were not in a position in which they could have been surprised (according to the definition defended in the previous section, at least).

(3) The solution should do justice to the intuitive meaning of the announcement.

As already mentioned in the introduction the lax condition are more difficult to argue for, since they essentially rely on our intuitions. Of course, I believe that the way in which I interpret the teacher's announcement is intuitive:

(4) The solution should do justice to the intuitive plausibility of the pupils' reasoning.

This is one of the most difficult conditions to satisfy if one interprets the surprise examination paradox as not really being a paradox. The teacher's announcement is interpreted as containing three different parts. One of them contains the protocol the teacher will follow. The students' reasoning comes from the fact that interpret the second and the third part of the teacher's announcement as being truly contradictory, whereas in fact one talks about what is happening in the present, as it were, while the other talks about what can happen in the future, so the contradiction is only apparent. The nice thing of explaining a puzzle by recourse to a protocol is succinctly put by Bovens and Ferreira 2010 (10): "[w]hat makes an appeal to protocols so inviting is that it provides us not only with a correct treatment ..., but also with an error theory of all the confusion in this area." (p. 480) The confusion stems from not taking into consideration the underlying protocol.

(5) The solution should make it possible for the students to be informed by the announcement.

This is obviously so. The students' beliefs are changed by the teacher's announcement.

(6) The solution should explain the role, in the generation of the puzzle, of the announcement's being made to the students.

As already mentioned, if the teacher does not make the announcement then the students cannot be surprised.

(7) The solution should account for the fact that before the teacher's announcement the students are indifferent both with regards to there being an exam and, conditional on there being an exam, with regards to its actual date.

This is obviously so. Before updating with the first part of the teacher's announcement, in the initial model, $M$, the students manifest this sort of indifference.

(8) The solution should do justice to the intuition that there is something truly interactive in the scenario: after the teacher's announcement, both the teacher and the students learn something about one another.

In the previous analysis, the students obviously learn a lot from the teacher's announcement. However, also the teacher learns something about the students by announcing them that they will receive a surprise examination, namely he learns that they can be surprised, which, before his announcement he knew it was false. So, the teacher's announcement affects both the students and the teacher.

(9) The solution should be applicable to *all variations* of the surprise examination paradox.

This is a very important point, and I believe that my solution can show such robustness. However, simply going through the existing variations will not prove much (new variations can always be generated). What would be needed is a general way of thinking about all the variations of the surprise examination paradox. Proving what the underlying structure of the surprise examination paradox is is a very complicated task. However, in order to show that my solution meets this last condition, I will try to identify some of the key components that exist in all the variations of the surprise examination paradox:

(1) there is always a trusted source of information and an agent that ends up being surprised (the source can be either another agent or Nature, as in Williamson's variation);
(2) there is always a sequence of elements that are eliminated one by one in the order given by the sequence;
(3) there is always a feature that all the elements of the sequence are equally probable to have, but *at most* one can have it! (this takes care of Williamson's conditional surprise examination paradox)
(4) there is always an information coming from the source saying that: (i) there exists an element in the sequence so that, if that element has the mentioned feature then the agent will be surprised when he comes to believe that it has it; and (ii) that element can have that feature only if the agent would be surprised if he were to learn that it [that element] has it. (This generates

the apparent paradoxical nature of the surprise examination paradox: this fourth condition seems to be circular)

Any scenario that meets these four criteria is a variation of the surprise examination paradox. In order to support this statement, let us see how it applies to some of the variations presented:

**Conditional Surprise Examination Paradox:** (1) the teacher and the student;
(2) days of the week;
(3) being examination days; there is at most one examination, if an examination at all!
(4) (i) the teacher announces that if there is an examination the following week, the students will be surprised by it; (ii) the teacher's announcement conveys also the protocol the teacher will follow: only give examinations on days in which the students do not believe that examinations are possible.

**Random Elimination Paradox:** (1) the game master in the game and the seeker;
(2) random numbers in a table;
(3) being undiscoverable positions;
(4) (i) the game announces that there is one cell in the table that is undiscoverable; (ii) the game master's announcement also conveys that he has assigned the seeker a certain cell only if the seeker would be surprised when he were to come to belief in which cell he had been placed.

**Designated Student:** (1) Robinson and the prototypical helper;
(2) several helpers in a row;
(3) having gold start on their necks;
(4) (i) Robinson announces that there is at least one helper that has a gold star on his neck; (ii) Robinson's announcement conveys also the protocol he will follow: only give put stars on he necks of the helper that does not consider it possible for him to have a golden star on his neck.

The paradoxical flavour of the surprise examination paradox is given by the tension between what the agent is announced that can happen and the fact that he is announced by what sequence of events it can happen.

Nevertheless, from this attempt to generalize some of the variations to coming up with a characterization of the surprise examination paradox is still a long way. A discussion of what exactly the surprise examination paradox is together with an assessment of the actual relations that hold between its variations is something I will leave for another time. I believe that such an analysis is possible and that it will rely significantly on the protocol that the teacher follows in giving the examination.

However, for now, I will only offer my solution that respects all but this last of the intuitive conditions that a solution to the surprise examination paradox has to meet.

CHAPTER 5

# Conclusions and further work

This thesis brings a few new contributions to the literature on the surprise examination paradox. (1) The problem of what exactly can be gained by solving the surprise examination paradox is more clearly stated than in previous works (such as Sorensen 1988 and Chow 1998); (2) the fact that we should be after *one* solution and not *the* solution is for the first time made clear; (3) the set of criteria that a solution to the surprise examination paradox has to meet is for for the first time explicitly stated in this extensive form (Wright and Sudbury only mention the first six conditions); (4) it is the first solution that questions the way in which surprise has to be construed, authors generally just inherit the definition of surprise and change everything else; and (5) it is the first solution reached by trying to satisfy the criteria that a solution is expected to meet (a few authors even check whether their solutions meet these, or similar criteria).

This thesis studied the surprise examination paradox in dynamic epistemic logic. It presented and criticized the already existing solutions given in this framework, and suggested a new one. The advantages of this new solution are philosophical. The solution presented in the previous chapter manages to fulfill all the criteria which a solution to the surprise examination paradox is expected to fulfill. However, there are another two interesting philosophical consequences that my solution has which were not yet explored, but which are extremely interesting.

In a recent series of articles, Bovens 2009, Bovens and Ferreira 2010 and Bovens forthcoming, bayesian epistemologists have began studying various probabilistic puzzles by uncovering the probabilistic protocol that underlie them. Their conclusion is that if two scenarios share the same protocol, then they share a lot of structure. This allowed them to offer a very informative comparison between Monty Hall, Sleeping Beauty and Judy Benjamin. Now, it might seem that their notion of protocol is very different from the notion of protocol employed in this thesis. However, this is not so. The notion of protocol they are using comes from Shafer 1985 and it refers to a tree, called a protocol and a probability function defined on that tree, called an evaluation. The DEL notion of protocol is copying the ETL notion of protocol, and is exactly a tree. So it seems that from the ETL/DEL notion of protocol to the probabilistic one there is one step, namely a probability distribution on the branches of that tree.

This observation was also made by Halpern and Tuttle 1993 who acknowledge that the two notions of protocols are mostly the same and they show that extending the ETL notion of protocol to Shafer's probabilistic notion is very simple indeed: add a probability distribution on the branches of the tree.

The reason why I mention this here is that in my solution to the surprise examination paradox, protocols played a crucial role: the teacher's announcement conveyed not only information about the current state of the system (there is an exam scheduled for one of the days of the week), but also procedural information about how the teacher will behave over time. This could only be accomplished by explicitly referring to a protocol. Furthermore, when discussing a way of uniting all the different variations, it seemed intuitive that what always remains the same in all variation is the protocol the teacher, or Robinson, or ... follow. Hence, I argued there that although there indeed are differences between all the variations, what makes them manifest the same type of puzzlement is precisely the fact that the agents in all variations seem to follow the same protocol. This suggests that Bovens and Ferreira's intuition with regards to the fact that paradoxes seem to become clearer once their underlying protocols are explicitated extends beyond the probabilistic setting. This might be an interesting development for their framework since they only consider probabilistic puzzles and do not state a way in which the probabilities can be dropped so that we are left only with protocols. Moreover, there would be an interest in developing a probabilistic protocol in DEL since, DEL has a lot of machinery for dealing with complicated scenarios of interaction between agents, especially for deceiving agents. Bovens and Ferreira's framework seems ill-equipped for such scenarios since they only condition propositional information. I will not explore these issues further, but a possible way of defining a probabilistic protocol on a PAL-generated ETL frame would be:

Definition 19 (Evaluation for a PAL Protocol). *Given a PAL protocol, $P$, a function $\mu_i : Agt \to (P \to [0,1])$ is an evaluation for $P$ iff*

*(1) if $\mu_i(\sigma) > 0$ and $len(\sigma) = n$ then $\mu_i(\sigma_{n-1}) > 0$;*

*(2) given that $\|\sigma_n\| = \{\sigma' : len(\sigma') \geq n \text{ and } \sigma'_{n-1} = \sigma_{n-1}\}$, for all $\sigma \in P$ ($len(\sigma) \geq n$) such that $\mu(\sigma_n) > 0$, $\sum_{\sigma' \in \|\sigma_n\|} \mu_i(\sigma') = 1$.*

Condition 1 would take care of the fact that in order for an element in a sequence to have positive probability, all its antecedents have to be possible. Condition 2 stipulates that the sum of all the nodes who have the same immediate parent has to be 1.

Definition 20 (Probabilistic PAL protocol). *Given a PAL protocol $P$ and an evaluation for $P$, $\mu_i$, a pPAL protocol is a structure $\pi = \langle P, \mu_i \rangle$.*

I will not explore this further here.

A second interesting philosophical consequence that my solution would have is also related to protocols, namely to the philosophy of protocols. Protocol constructions suggest a new type of knowledge: procedural knowledge, that is knowledge about the way in which an agent will act in the future, or about what can be learned in the future. However philosophers have so far neglected this type of knowledge. For instance Schwitzgebel 2010 (45) mentions procedural knowledge, but what he is referring to is the term used in cognitive science to refer to skills gained unconsciously. That is, autistic patients who are exposed to the same task every day improve their performances with the passing of time, however they have no memories about doing the task. Hence, it cannot be said that they know, in the declarative sense, how to solve the task, but they do know, in the practical sense, how to solve it. But this is not what protocol knowledge refers to. Protocol knowledge gives an agent information about future actions that himself or others will do. One nice application of this is the philosophy of trust. You can try define what it means to trust another agent by the following a protocol: if the trusted agent believes $\varphi$ you will revise your beliefs so that you come to believe $\varphi$. This might be too strong as the other agent could believe that your wife is cheating you and that you do not know that. Of course, you will never be able to come to believe this sentence (since it is a Moorean sentence). However, weaker definitions could be explored. This would vindicate trust as a notion that has to do with repeated interactions rather than with the present moment. Furthermore, in view of Bovens and Ferreira's arguments protocols could have great significance for identifying some intermediate level of common structure that more scenarios share: they are not identical, but nor do they completely differ. But I will not address these issues here, either.

There remain, of course, a few open questions:

**1.** The most significant open question which is also of great relevance to my solution is what exactly is the surprise examination paradox. In this thesis I mention that it feels intuitive that all the variations presented so far (maybe with the exception of the intention-based one) share the same source of puzzlement. However, this was never spelled out. In addition, when evaluation my solution, I go over the question of whether my solution is indeed robust since in order to really check this one would need the characterization of the surprise examination paradox; going through all known variations will not do any good. My intuition was that a good way of approaching the essence of the surprise examination paradox is through the protocol followed by the teacher and which the students come to know from his announcement. This would be coherent with a recent attempt of Bovens and Ferreira 2010 to compare probabilistic paradoxes by uncovering their underlying probabilistic protocols. It might be the case that the surprise examination paradox is a good example of showing that the same mechanism can work in a non-probabilistic environment.

However, these intuitions need to be spelled out in a more systematic way and a characterization of the (essence of) surprise examination paradox is still missing.

**2.** The second question that requires further investigation is what exactly is the connection between the surprise examination paradox and other puzzles such as the prisoner's dilemma, the toxin puzzle, etc.? It has been proposed in the literature that in fact the surprise examination paradox hides in itself a series of other puzzles (see Sorensen 1988 for an overview of the puzzles that seem to be related to the surprise examination paradox). Can a solution to the surprise examination paradox shed any light on these other puzzles? Even if there is no direct way of applying the solution to the surprise examination paradox presented here to those puzzles, it might be relevant to apply the same framework of the DEL-generated ETL frames to solve them. Therefore, interesting new application open for trying to understand the prisoner's dilemma or the toxin puzzle, or vagueness by means of DEL-generated ETL frames.

**3.** Finally, there are a few logical problems surrounding my solution that need to be explored. For example, I suggested that the teacher's announcement contains an announcement about the protocol that the teacher will follow when he will give the examination. This was not at all clear in the formalism. In order to be able to express that we need a logic in which protocols can be announced. An example would be Wang 2010. However, I do not investigate the issue further in the thesis. Also, the protocol that seems to be in place seems to have future oriented preconditions. This is not the way in which it was set up in Chapter 2. There a protocol only accepted preconditions coming from the language of epistemic logic. If the preconditions talk about the future, a few complications arise. See Hoshi 2009 for a discussion.

In conclusion, DEL-generated ETL frame can help us clarify what surprise means and what the teacher's announcement implies in such a way that the surprise examination paradox is solved in an intuitively expected way. Furthermore, such a solution would have interesting philosophical consequences both on the way in which we reason about scenarios of the same sort as the surprise examination paradox, and on the way in which we analyze paradoxes in general.

# Bibliography


[1] Ayer, A., 1973, "On a Supposed Antinomy", *Mind* **82**, 125-126

[2] Baltag, A. , Moss, L. , Solecki, 1998, "The Logic of Public Announcements, Common Knowledge, and Private Suspicions", in Gilboa, I., ed., *Proceedings of the 7th Conference on Theoretical Aspects of Rationality and Knowledge (TARK 98)*, 43-56.

[3] Baltag, A., 2003, "Logics for Communication: Reasoning About Information Flow in Dialogue Games", presented at *NASSLLI 2003*, http://www.indiana.edu/ nasslli/2003/program.html, slides.

[4] Baltag, A., Smets, S., 2008, "A Qualitative Theory of Dynamic Interactive Belief Revision", in Bonnano, G., van der Hoek, W., Wooldridge, M. (eds), 2008, *Logic and the Foundation of Game and Decision Theory (LOFT 7)*, Amsterdam: Amsterdam University Press, 11-58.

[5] Baltag, A., 2009, "SURPRISE!? An Answer to the Hangman, or How to Avoid Unexpected Exams!", presented at *Logic and Interactive Rationality Seminar (LIRA)*, slides.

[6] Baltag, A., 2010, "Dynamic-Doxastic Norms Versus Doxastic-Norm Dynamics", presented at *Formal Models of Norm Change 2*, http://www.cs.uu.nl/events/normchange2/program.html, slides.

[7] Baltag, A., Smets, S., 2010, "Multi- Agent Belief Dynamics", course given at *NASSLLI 2010*, Indiana University, http://www.vub.ac.be/CLWF/nasslli2010lecture3.pdf.

[8] Binkley, R., 1968, "The Surprise Examination in Modal Logic", *The Journal of Philosophy* **65**, 127-136.

[9] Bovens, L. , 2010, "Judy Benjamin is a Sleeping Beauty", *Analysis* **70**, 23-26.

[10] Bovens, L. , Ferreira, J.L. , 2010, Monty Hall Drives a Wedge between Judy Benjamin and the Sleeping Beauty: a Reply to Bovens, *Analysis* **70**, 473-481.

[11] Bovens, L. , forthcoming, "Does it Matter whether a Miracle-Like Event Happens to Oneself rather than to Someone Else?", manuscript.

[12] Bunch, B., 1982, *Mathematical Fallacies and Paradoxes*, New York: van Nostrand.

[13] Cargile, J., 1967, "The Surprise Test Paradox", *The Journal of Philosophy* **64**, 550-563.
[14] Chihara, C., 1985, "Olin, Quine and the Surprise Examination", *Philosophical Studies* **47**, 191-199.
[15] Chow, T., 1998, "The Surprise Examination or Unexpected Hanging Paradox", *The American Mathematical Monthly* **105**, 41-51.
[16] Dietl, P., 1973, "The Surprise Examination", *Educational Theory* **23**, 153-158.
[17] Fagin, R., Halpern, J., Moses, Y., Vardi, M., 1995, *Reasoning about Knowledge*, Cambridge, MA: The MIT Press.
[18] Fitch, F., 1964, "A Goedelized Conception of the Prediction Paradox", *American Philosophical Quarterly* **1**, 161-164.
[19] Gerbrandy, J., Groeneveld, W., 1997, "Reasoning about Information Change", *Journal of Logic, Language and Information* **6**, 147169.
[20] Gerbrandy, J., 1999, *Bisimulations on Planet Kripke*, PhD Thesis, University of Amsterdam.
[21] Gerbrandy, J., 2007, "The Surprise Examination in Dynamic Epistemic Logic", *Synthese* **155**, 21-33.
[22] Gosh, S., 2010, "Changing Protocols and More ...", *The Many Faces of Protocols and Knowledge Workshop*, University of Amsterdam, 13 September 2010.
[23] Halpern, J., Fagin, R., 1989, "Modelling Knowledge and Action in Distributed Systems", *Distributed Computing* **3**, 159-177.
[24] Halpern, J., Vardi, M., 1989, "The Complexity of Reasoning about Knowledge and Time. I. Lower Bounds", *Journal of Computer and System Sciences* **38**, 195-237.
[25] Halpern, J., van der Meyden, R., Vardi, M., 2004, "Complete Axiomatizations for Reasoning about Knowledge and Time", *SIAM Journal of Computing* **33**, 674-703.
[26] Hoshi, T., 2009, *Epistemic Dynamics and Protcol Information*, PhD Thesis, Stanford University.
[27] Hoshi, T., 2010, "Merging DEL and ETL", *Journal of Logic, Language and Information* **19**, 413-430.
[28] Kaplan, D., Montague, R., 1960, "A Paradox Regained", *Notre Dame Journal of Formal Logic* **1**, 79-90.
[29] Levy, K., 2009, "The Solution to the Surprise Exam Paradox", *The Southern Journal of Philosophy* **47**, 131-158.
[30] Lorini, E., Castelfranchi, C., 2006, "The Unexpected Aspects of Surprise", *International Journal of Pattern Recognition and Artificial Intelligence* **20**, 817-833.
[31] Lorini, E., Castelfranchi, C., 2007, "The Cognitive Structure of Surprise: Looking for Basic Principles, *Topoi* **26**, 133-149.

[32] McLelland, J., Chihara, C., 1975, "The Surprise Examination Paradox", *Journal of Philosophical Logic* **4**, 71-89.
[33] Mosteller, F., 1965, *Fifty Challenging Problems in Probability with Solutions*, Reading, Mass.: Addison-Wesley.
[34] Nerlich, G., 1961, "Unexpected Examinations and Unprovable Statements", *Mind* **70**, 503-513.
[35] O'Connor, D., 1948, Pragmatic Paradoxes, *Mind* **57**, 358-359.
[36] Olin, D., 1983, "The Prediction Paradox Resolved", *Philosophical Studies* **44**, 225-233.
[37] Olin, D., 1988, "Predictions, Intentions and the Prisoner's Dilemma", *Philosophical Quarterly* **38**, 111-116.
[38] Pacuit, E., Simon, S., "Reasoning with Protocols Under Imperfect Information", manuscript.
[39] Parikh, R. , Ramanujam, R. , 2003, "A Knowledge Based Semantics of Messages", *Journal of Logic, Langauge and Information* **12**, 453-467.
[40] Parikh, R. , Ramanujam, R. , "A Knowledge Based Semantics of Messages", online.
[41] Plaza, J., 1989, "Logics of Public Communications", *Proceedings of the 4th International Symposium on Methodologies for Intelligent Systems*, 201216.
[42] Quine, W.v.Q., 1953, "On a So-Called Paradox", *Mind* **62**, 65-67.
[43] Shafer, G., 1985, "Conditional probability", *International Statistical Review* **53**, 261-277.
[44] Shaw, R., 1958, "The Paradox of the Unexpected Examination", *Mind* **67**, 382-384.
[45] Schwitzgebel, E., 2010, "Belief", *The Stanford Encyclopedia of Philosophy (Winter 2010 Edition)*, Zalta, E. (ed.), http://plato.stanford.edu/archives/win2010/entries/belief/.
[46] Smith, J., 1984, "The Surprise Examination on the Paradox of the Heap", *Philosophical Papers* **13**, 43-56.
[47] Sorensen, R., 1988, *Blindspots*, Oxford: Clarendon Press.
[48] Sorensen, R., 2009, "Epistemic Paradoxes", in Zalta, E., 2009, *The Stanford Encyclopedia of Philosophy (Spring 2009 Edition)*, URL = ¡http://plato.stanford.edu/archives/spr2009/entries/epistemic-paradoxes/¿.
[49] van Benthem, J., 2004, "What One May Come to Know", *Analysis* **64**, 95105.
[50] van Benthem, J., 2007, "Dynamic Logic for Belief Revision", *Journal of Applied and Non-Classical Logics* **17**, 129-155.
[51] van Benthem, J. , Gerbrandy, J. , Hoshi, T. , Pacuit, E. , 2009, "Merging Frameworks for Interaction", *Journal of Philosophical Logic* **38**, 491-526.
[52] van Benthem, J. , forthcoming, *Logical Dynamics of Information and Interaction*, Cambridge, MA: Cambridge University Press

[53] van Benthem, J., Pacuit, E., Roy, O., "Towards a Theory of Play: a Logical Perspective on Games and Interactions", manuscript.
[54] van Ditmarsch, Kooi, K., 2006, "The secret of my success", *Synthese* **153**, 201-232.
[55] van Ditmarsch, H, van der Hoek, W., Kooi, B., 2007, *Dynamic Epistemic Logic*, Dordrecht: Springer.
[56] vos Savant, M., 1990, "Ask Marily", *Parade Magazine*, 15.
[57] Wang, Y., 2010, *Epistemic Modeling and Protocol Dynamics*, PhD Thesis, University of Amsterdam
[58] Williamson, T., 2000, *Knowledge and Its Limits*, Oxford: Oxford University Press.
[59] Wright, C., Sudbury, A., 1977, The Paradox of the Unexpected Examination, *Australasian Journal of Philosophy* **55**, 41-58.