The Surprise Examination Paradox
A review of two so-called solutions in dynamic epistemic logic

Alexandru Marcoci

1 The paradoxical scenario

In a school where exams can never be less than seven days apart, a teacher announces his students that they will get an exam the following week and that it will be a surprise. For simplicity let us assume that due to the seven day condition the only possible exam days in the following week are Wednesday, Thursday and Friday[1].

Given the teacher's announcement a student will reason in the following manner:

- Assume that by Friday I will not have received an exam. Since there has to be an exam on one of the three days, it will have to be on Friday. However, I will then be able to predict it before it occurs. Therefore Friday cannot be the day of the exam.
- Assume then that by Thursday I will not have received an exam. Since there has to be an exam on one of the three days, and cannot be on Friday (by the previous argument), it has to be on Thursday. However, I will then be able to predict it before it occurs. Therefore Thursday cannot be the day of the exam.
- But then, since there has to be an exam on one of the three days, and it cannot be on Thursday or Friday (by the previous arguments), it has to be on Wednesday. However, I will then be able to predict it before it occurs. Therefore Wednesday cannot be the day of the exam.
- So, it is false that I will be surprised when the exam will come, since there can either be an exam which I will predict, or there can be no exam.

Nevertheless an exam is given on one of the three days and the student is surprised: by assuming that there is going to be a surprise exam the student is led to believe that there can be no surprise exam, only to receive a surprise exam.

[1] This simplification was inspired by Gerbrandy (2007)

2 Aim of the paper

The surprise examination paradox has been the topic of numerous philosophical papers. Ken Levy (2009) lists fifty-four articles dedicated to it in various philosophical journals. In recent years it has made its way in the dynamic epistemic logic literature as well. First of all, Gerbrandy (2007) has argued that the students' reasoning can be best encapsulated in the framework of public announcement logic, and that the error lies in taking ”success” for granted. Also, in a series of conferences, Baltag (2009a, 2009b, 2010) claims to have found the ”correct” solution. His idea is to have the students revise their attitude towards the teacher once they reach paradox. The intuition is that their backward argument only works as long as they trust the teacher and assume both of his claims (about the existence of an exam and about its unpredictability) to be true. Baltag shows that once you are willing to manifest some distrust towards the teacher there is a (unparadoxical) solution that will imply that the teacher has lied and that will make the students expect the exam.

In this paper I intend to review the dynamic logic literature on the surprise examination paradox, and I will criticize the solutions advanced so far on the ground that they fail to solve some equivalent reformulations of the paradox. In the end, different further developments are hinted at.

3 Philosophical assumption

The following principle is never invoked in the dynamic logic literature on the surprise examination, however, it is commonplace in the philosophical literature[2]:

ALL Any real solution to the surprise examination paradox has to hold for all its possible reformulations.

ALL is very intuitive: as long as a variation of the initial scenario depicts the same paradox, indeed the solution of the original paradox should extend to the modified version. But naturally, this raises a problem of degree - how much can we modify the original scenario and still call the result an instance of the same paradox? I will not offer an answer to this question in this paper, but at the end I will hint at a possible strategy for finding such an answer.

4 Gerbrandy's solution

Gerbrandy (2007) believes that public announcement logic (PAL) can not only offer a formal proof for the informal reasoning of the student, but it can also unveil the error in the students' reasoning. I will assume familiarity with PAL. However, there is an essential

[2] E.g. See Ayer, Sorensen (1988), p. 311, and Williamson (2000), p. 137.

difference between Gerbrandy's formulation of PAL and the usual one[3], namely that its axiomatization contains neither the T, nor the D axioms for knowledge[4]. That is, Gerbrandy assumes neither that knowledge is factive (the frame is not reflexive), nor that it is consistent (the frame is not serial):

$$\nvdash K\varphi \rightarrow \varphi, \text{ and}$$

$$\nvdash \neg K\bot.$$

Although I agree with Gerbrandy that the scenario requires a K45 modal operator I am sceptical that (i) this can be interpreted as knowledge - how can knowledge be inconsistent? - or (ii) that it should be interpreted as knowledge - see the section on the cognitive structure of surprise below. I think a better way of construing it is as belief, so from now on in presenting Gerbrandy's analysis I will consistently write B (Belief) wherever he writes K (Knowledge). This will not affect Gerbrandy's proofs, since the principles governing the agents' doxastic attitudes will not change (B will still remain a K45 operator!).

Gerbrandy then formalizes the teacher's anouncement in this logic as follows:

$$S = (we \wedge \neg Bwe) \vee (th \wedge [!\neg we]\neg Bth) \vee (fr \wedge [!\neg we][!\neg th]\neg Bfr) \vee B\bot$$

The meaning of all disjuncts but the last is very straightforward. The last disjunct comes from the fact that the scenario suggests that after eliminating all days of the week and thus $B\bot$, the students do not "go crazy" and are still surprised when the exam comes, surprise being a consistent doxastic attitude. This motivates why a $K45$ operator is needed.

Now the reasoning of the students can be represented in PAL as the following proof:

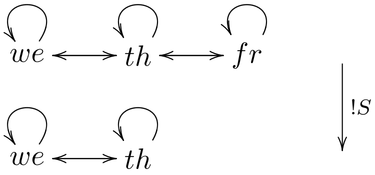
$$[!S]BS \vdash [!S]B\bot \wedge [!S]B\bot \vdash [!S]S$$

That is, after the announcement that the exam is going to be a surprise, the students will know that. But from knowing that the exam is going to be a surprise they infer $B\bot$ (inconsistent beliefs), which in turn implies that the exam is going to be a surprise. The fact that $[!S]B\bot \vdash [!S]S$ is obvious from the definition of S. However, the fact that $[!S]BS \vdash [!S]B\bot$ requires a proof that due to lack of space I will omit

Gerbrandy believes that the faulty reasoning is evident in this formalism: the students assume S to be successful. However, this does not hold for all sentences, and a non-problematic solution emerges (Gerbrandy believes) if we take surprise to be unsuccessful. What Gerbrandy aims at cutting short by this analysis is $[!S]B\bot$. And the dynamic semantics guarantees this:

[3] See for example Ditmarsch, van der Hoek, and Kooi (2007), chapter 4.

[4] Gerbrandy also defines a Kripke frame in an unusual way, in terms of information states. However, this is not essential since there is a way of translating it in the regular format, see his n2, p. 32.



Gerbrandy's conclusion is that although the students are right in thinking that the exam cannot be on Friday (as indeed it will no longer be a surprise), they are not warranted in extending this argument to the other days of the week. In this case, after the announcement of S, fr is eliminated as a possible exam day, and wednesday and thursday are conserved as indeed they are the only two days in the original model when the exam can come as a surprise. However, in the updated model this is no longer the case, since if the exam does not come on Wednesday,the students will anticipate it on Thursday. But Gerbrandy thiks that this is unproblematic. So, ”before the Teacher's announcement, it **was** the case that (**if** the Teacher **didn't make the announcement, then**) the exam's date **would have been** a surprise”[5] and this is becasue the announcement is truthful (as it is a public announcement). That is,

$$M, we \vDash S \cdots !S \cdots M|S, we \vDash S$$

$$M, th \vDash S \cdots !S \cdots M|S, th \nvDash S$$

Gerbrandy realizes, however, that this might be an oversimplification. There seems to be something self-referential in the teacher's announcement that would require that even after the announcement surprise should continue to hold. He investigates two possible such self-referential reformulations of the announcement:

1. $\delta \leftrightarrow S \wedge [\delta]S$, where δ is the meaning of the teacher's announcement.

2. $\delta \leftrightarrow S \wedge [\delta]\delta$, where δ is the meaning of the teacher's announcement.

He rejects (proof omitted) all self-referential alternative formalizations, since they are not true irrespective of when the teacher gives the examination. That is, there are situations in which assuming that the announcement is self-referential leads to contradiction[6].

In conclusion, for Gerbrandy the meaning of the teacher's announcement is fully grasped by S, there is no paradox involved as long as the teacher decides to give the examination on any day of the week except for Friday, and the reason why the informal argument is wrong is that it assumes that the announcement is successful, when in fact it is not.

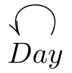
5 Critique of Gerbrandy's solution

The surprise examination paradox can be formulated even when we take only one possible exam day. That is, if a teacher announces that there will be a surprise examination the

[5] Baltag 2009 gro, slide 4. Original emphasis.

[6] Gerbrandy (2007), p. 29.

following day, the students will reason that if there is an exam, then there will be no surprise, and hence they eliminate the following day as a possible (surprise) exam day, being surprised when the exam actually occurs. By **All**, if a solution to the surprise examination paradox is found, then this solution should also solve the one-day version of the paradox. However, it is not the case that Gerbrandy's solution solves this version. The initial situation, M_0, is:



The teacher's announcement, in Gerbrandy's formalization, would be:

$$S'' = (Day \wedge \neg K Day) \vee K\bot$$

However, since $M_0, Day \nvDash \neg K Day \vee K\bot$, after the announcement $!S''$, Day will be deleted, thus recreating the original paradox, without invoking success. That is,

$$[!S'']\bot \vdash [!S'']K\bot \wedge [!S'']K\bot \vdash [!S'']S''.$$

But this is exactly what Gerbrandy was hoping to avoid when he claimed that assuming the success of S is what leads the students to paradox.

In conclusion, Gerbrandy's solution only works if the initial model contains more than one state, that is, if the initial situation contains more than one possible (surprise) exam day. This would be arbitrary and hence cannot be accepted as a/the correct solution.

6 Baltag's solution

I assume the definitions of plausibility models, knowledge, and (conditional) belief (for example, slides 27-30 of Baltag (2010)). In this section I will closely follow Baltag (2010).

Belief upgrades Assume the initial model to be $M = \langle W, \sim, \leq, V\rangle$, then after the belief upgrade with φ, we get the new model $*\varphi(M) = \langle W', \sim', \leq', V'\rangle$, such that

1. $W' = W$
2. Knowledge increases or stays the same: $s \sim' t \Rightarrow s \sim t$
3. Given that $[s] = \{t : t \sim s\}$, $[s]' = \{t' : t \sim' s\}$ and $||\varphi||_M = \{w \in W : M, w \vDash \varphi\}$,

   $$||\varphi||_M \cap [s]' \neq \varnothing \Rightarrow Max_{\leq'}[s]' \subseteq ||\varphi||_M$$

   That is, the most plausible worlds of the new information cell are among the ones that satisfied φ in the old model (if there are such worlds).
4. $V' = V$

Dynamic operators $[*\varphi]\psi \Leftrightarrow$ "ψ will surely be true after the upgrade with $*\varphi$".

Temporal operators We can identify a sequence of plausibility models, $M_0, M_1, M_2 \dots$ obtained by successive upgrades $*\varphi_0, *\varphi_1, *\varphi_2 \dots$, $M_{n+1} = *\varphi_n(M_n)$, by means of a temporal operator "$NEXT$":

$$M_n, w \vDash NEXT\psi \Leftrightarrow M_n, w \vDash [*\varphi_n]\psi$$

Dually,

$$M_n, w \vDash BEFORE\psi \Leftrightarrow M_{n-1}, w \vDash \psi$$

It is obvious that updates, lexicographic upgrades and conservative upgrades are all examples of such belief upgrades, and that they correspond to different levels of trust that the agent that is revising is assigning to the source making the announcements. 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 introduces atomic sentences $!, \Uparrow, \uparrow, FAGM, MAGM$[7] that correspond to the agent's attitudes towards the source of the announcements (variable τ will range over them). Also he introduces the dynamic modality $[*\varphi]$, which he interprets as follows: the transformation $*\varphi$ is an upgrade that reorders each partition cell $[s]$ by applying the corresponding type of upgrade τ, where $s \vDash \tau$. When this reordering is inconsistent for some τ, the worlds that satisfy τ 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, $\lesssim$, on the states of S, so that:

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

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

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

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

and

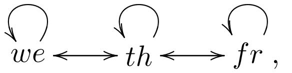
$$O(FAGM) \wedge O(MAGM|\neg FAGM)$$

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 $\lesssim$ relation. But the more

[7] $FAGM$ and $MAGM$ represent the attitudes that an agent might consistently have towards a source when in face of revision with higher-order belief sentences. They correspond to $K\neg\varphi \vee [*\varphi]B(BEFORE\varphi)$, and $[*\varphi]K(BEFORE\neg\varphi) \vee [*\varphi]B(BEFORE\varphi)$, respectively

interesting fact is that regular revisions can lead to revisions of the $\lesssim$ relation. This is the case with the infallibility norm ($O(!)$) and the Moore sentence.

Baltag (2010) depicts the initial situation from the point of view of the students as



where i means "the exam is going to take place on the i-th day of the week", and the arrows represent the plausibility that the students assign to the exam being on the i-th day.

The crux is identified by Baltag to be the meaning (and implicitly the formal counterpart) of the teacher's announcement. First of all, the fact that the exam will be a surprise has to be construed as "the evening before the exam day, the students will not believe that the exam is tomorrow"[8]:

$$surprise = \bigwedge_{we \le i \le fr} (i \to [!(\bigwedge_{we \le j < fr} \neg j)]\neg Bi)$$

However, Baltag believes 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"[9]. Formally,

$$*(NEXTsurprise)$$

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:

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

$$O(FAGM_{Teacher}) \wedge O(MAGM_{Teacher}|\neg FAGM_{Teacher})$$

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

Further, Baltag (2010) proves that the remaining revision norm, $MAGM$ works, as there is an unique upgrade satisfying $MAGM$ for which no contradiction arises. From, $MAGM$ and

$$BEFORE(NEXTsurprise) \Leftrightarrow surprise$$

we can derive that

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

[8] Cf. Gerbrandy who takes 'surprise' to be about ($K45$) knowledge.

[9] This sentence is self-referential. Cf. Gerbrandy.

However,

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

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

$$\begin{tikzcd}
we \arrow[loop above] & th \arrow[l] \arrow[loop above] & fr \arrow[l] \arrow[loop above]
\end{tikzcd},$$

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

7 Critique of Baltag's solution

7.1 The teacher isn't a liar

First of all, it is questionable if Baltag's solution is indeed a solution to the Surprise examination paradox. As explained above, his analysis concludes that if the students trust the teacher, then they really reach a contradiction. In more formal terms, if $M \models !_{Teacher} \wedge (\Uparrow_{Teacher} |\neg !_{Teacher})$, which means that the students cannot distrust the teacher, then $M \models [*NEXTsurprise]\bot$, where $*NEXTsurprise$ stands for the teacher's announcement. In other words, the students do not err in their informal reasoning, provided they trust the teacher. So, the only thing left to establish is if there is a way in which the paradox can be avoided, and according to Baltag, such a way out would be for the students to doubt the teacher and believe that he has lied to them. For Baltag's case such a way out is present in the hierarchy of norms that he assigns to the students. However, Baltag is not very clear what is the meaning of his hierarchy of norms: is it inherent in the meaning of the upgrades, or is it given relative to the agents? Since the atoms that depict the norms the students adopt, and hence the trust they have in face of the teacher's announcements, are only indexed with respect to the source (i.e. the teacher) it might be the case that he assumes the former. But this would mean that all upgrades fit in a strict ordering with respect to the level of trust they correspond to. This might be the case for $!$, $\Uparrow$, and $\uparrow$, but I do not believe that more unusual upgrades, e.g. # (suggestion), have such a clear counterpart on the trust scale. So, I believe that what Baltag has in mind here is that the hierarchy of norms is given in the model, and is not in any way derived from the meanings of the upgrade operations. But in this case, his solution lies on a very arbitrary fact of the students having MAGM somewhere in their hierarchy of norms. This might simply not be the case. As at the beginning of this paragraph, M might *only* satisfy $!_{Teacher} \wedge (\Uparrow_{Teacher} |\neg !_{Teacher})$. In this case, the only thing that Baltag's solution can offer is the conclusion that the paradox is really paradoxical and that there is no way out. However, this hardly is a solution.

It might be argued that the students *should* adopt the hierarchy that Baltag assigns to them. I believe, however, that this is not the case. The reason is encapsulated in Sorensen's critique to Quine's solution to the Surprise Examination Paradox:

> Quine's analysis (. . .) commits us to an unacceptable sort of skepticism. (. . .) Quine maintains that the elimination argument simply shows that the judge's declaration[10] is insufficient evidence for K[11] to know that he will be hung on an unforeseen day. (. . .) [However,] K confronts an epistemologically ideal judge in epistemologically ideal circumstances.[12]

What Sorensen is saying is 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 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, I believe, against Baltag's solution: nothing in the scenario seems to indicate that the teacher might be deceitful (lying) or that he may be wrong. He is an epistemologically ideal teacher, *ipso facto* excluding both lying and saying something false [13]. A possible solution to this problem might be to argue that the agents in Baltag's scenario are not ideal. Nevertheless, this would not solve the problem because in a real life scenario also the students will no longer be ideal. And being confronted with a contradiction between their beliefs and an announcement by the teacher, it would be more rational for them to doubt their reasoning rather than the teacher's: the students should abound (given that the teacher is a good teacher) in situations in which what the teacher announced ended up being true, even when they did not believe that. So it would be more rational for them to blame themselves rather than to believe the teacher has lied (for what Baltag proposes to work we should confront a real-life teacher with ideal students). I believe this is a good enough reason to follow Sorensen and assume that all the agents in the surprise examination scenario are ideal and that ideal means reliable, and look for a solution that does not make the teacher a liar.[14]

However, claiming that the teacher and the students are ideal raises yet another problem: do ideal agents have anything less than knowledge? Binkley answers this question elegantly:

[10] Quine (1953) uses the hangman version of the surprise examination paradox

[11] Quine denotes the agent in his version of the paradox by K.

[12] Sorensen (1988), p. 310.

[13] This does not mean that he knows all truths, just that his inferences given his circumstances are flawless and that he is aware of his own ignorance.

[14] I hope to offer such a solution later.

> [I]t should be noted that we are concerned with an ideal seeker after knowledge, not necessarily someone who already possesses knowledge, and that consequently the ideal knower must be defined in terms of what he judges or believes, not in terms of what he knows. He is meant to be an ideal of rationality, and circumstances may conspire to prevent a rational man from acquiring knowledge, and perhaps may even lead him into false belief. [15]

Summing up, one important reason for rejecting a solution like Baltag's is that according to it, if the agents are ideal, and if this means that they are reliable, or if simply their hierarchy of trust does not include MAGM, then there is no way out of the paradox: the paradox is truly paradoxical. This cannot be the "real" solution! However, this reason is external to Baltag's framework and it depends more on the choices the modeler makes: (Quine and) Baltag consider(s) that the teacher might be lying, whereas Sorensen considers that this possibility is excluded by the ideal nature of the agents, or just put differently, Baltag assigns MAGM to the students' hierarchy of norms, whereas others might not. So far, nothing has been said about the particulars of Baltag's solution, so one might believe that working within Baltag's modeling choices his solution works. I believe that this is not the case and in the next section I will attempt to argue against Baltag's solution form an internal point of view.

7.2 Lying is never the answer

Let us put these worries raised above aside for now and accept Baltag's solution[16] to be applicable to the surprise examination paradox. First of all, Baltag's strategy in tackling with the surprise examination scenario again: The teacher makes an announcement. The students try to trust the teacher as much as possible, which means that they first try to update with the information coming from the teacher. If updating leads to contradiction, then they try to lexicographically upgrade with the same information, which would be the next best thing in terms of trust towards the teacher, and so on (with the proviso that anything else than hard update and lexicographical upgrade means the students do not really trust the teacher). Since they have a strict hierarchy of all the ways in which they can upgrade with the information they receive from the teacher in terms of their trust towards him, once they reach a non-contradictory way of upgrading they stop and integrate that information in the way dictated by that upgrade; where the hierarchy is:

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

and

$$O(FAGM) \wedge O(MAGM|\neg FAGM)$$

[15] Binkley 1968, p. 128.

[16] For clarity, although it might have been clear from the context, I understand a solution here as a strategy of dealing with the scenario of the paradox which does not yield a contradiction.

Now consider the following reformulation of the surprise examination paradox:

Example 7.1 (The Surprise *Examination*[17])**.** *In the kind of school where every exam comes as a surprise and the number of exams students may receive during a n-day semester varies from 0 to n (the evaluation of the students is not made in terms of performance in exams), a teacher announces to his class: "Next week, there will be an exam (and only one!)." It is commonly understood that an exam comes as a surprise if you do not believe, the evening before, that it is given the next day.*[18]

What the students know is a conditional sentence: if there is an exam then it will be a surprise. But this is not expressive enough, as by a carefully chosen announcement the teacher can then change their beliefs in such a way that they loose their knowledge of the surprise sentence in the new model. The scenario suggests that this cannot happen (the rules of the school cannot change and nothing the teacher says should make the students change their epistemic attitude towards them). So what the students know is that even after the announcement of the teacher they will still be surprised when the exam will come, if it does. Therefore, after the announcement (that there will actually be an exam) they will be able to reason in the following manner: if there is no exam by Thursday evening, then there has to be an exam on Friday, but since exams come always as a surprise (i.e. you do not believe, the evening before, that it is given the next day) there can be no exam on Friday. If there is no exam by Wednesday evening, then there has to be an exam on Thursday, since by the preceding argument it cannot be on Friday. But then, since exams always come as a surprise there can be no exam on Thursday. And so on until all days of the week are eliminated. Nevertheless an exam does come on Wednesday, say, and the students are surprised. What went wrong with the students' reasoning?

Remark that **The Surprise *Examination*** is different from the scenario Baltag and Gerbrandy use, namely **The *Surprise* Examination**. The main difference is that whereas in the latter (used by Baltag and Gerbrandy) the students know that there is an exam every week, the teacher announcing them that the exam in a particular week is going to be a surprise one, in the former (above) the students know that if there is an exam on a particular day, then that exam will come as a surprise to them; the teacher now simply announces that a particular week contains one exam day (and only one!). Despite the differences between **The Surprise *Examination*** and **The *Surprise* Examination**, the ways in which the students reason are so similar that I believe it is safe to assume that all who read the scenarios agree that they have to be instantiations of the same underlying paradox. (In the future I will like to be able to argue for this idea in a more precise way.)

In what follows I will apply Baltag's strategy to the **The Surprise *Examination***.

[17] I suggest which sentence is announced by the teacher by emphasizing it in the name of the scenario.

[18] One might notice the similarity in structure with the scenario used by Gerbrandy.

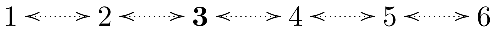
Definition 7.1. *The meanings of the exam and of the surprise sentence and of what the students know in the initial model, respectively, are rendered by the following formulas:*

1. $surprise := \bigwedge_{1 \leq i \leq 5}(i \to [!(\bigwedge_{1 \leq j < 5} \neg j)] \neg Bi)$
2. $exam := \bigvee_{1 \leq i \leq 5} i$
3. $NEXTsurprise := [*\varphi_n]surprise$*, where* $NEXT$ *is a temporal operator over sequences of models defined as below.*

Definition 7.1 (Temporal operators)**.** *We can identify a sequence of plausibility models,* $M_0, M_1, M_2 \ldots$ *obtained by successive upgrades* $*\varphi_0, *\varphi_1, *\varphi_2 \ldots$*,* $M_{n+1} = *\varphi_n(M_n)$*, by means of a temporal operator "*$NEXT$*":*

$$M_n, w \vDash NEXT\psi \Leftrightarrow M_n, w \vDash [*\varphi_n]\psi$$

Using Baltag's formalism, the initial situation (from the point of view of the students, excluding transitive and reflexive arrows, and assuming all states to be epistemically indistinguishable) before the announcement of the teacher is $M, 3$:



The state labeled 6 is a state in which no exam takes place. The other states get their usual readings, e.g. the world labeled 3 represents the world in which there is an exam on the third day of the week. The arrows depict plausibility relations between states. In this initial model, things are simple: all worlds are equally plausible. The emphasized state represents the actual state of the world. That means that an exam will actually come on the third day.

Proposition 7.1. *If* $M \vDash K_{students}NEXTsurprise$*, then* $M \vDash [!exam]\bot \wedge [\Uparrow exam]\bot \wedge [\uparrow exam]\bot \wedge [*exam]\bot$*, for any soft upgrade* $*exam$ *respecting FAGM.*

To prove this, the following proposition will be needed:

Proposition 7.1. $Bexam \Rightarrow \neg Ksurprise$

Proof. Let $\mathcal{D}(M) = \{s_1, s_2, \ldots s_n\}$. The subindices of the states indicate their ordering in the plausibility relation: $s_1 \leq s_2 \leq \ldots s_n$. Remember that we work over a sequence of models defined as follows:

$$M_1 = M$$

$$M_n = M_{n-1}|\neg s_j$$

This sequence corresponds to the elimination of days that comes with the flow of time. M_n is obtained by deleting the state at which s_j is satisfied, and not the one at which s_{n-1} is satisfied since it is not necessary that the plausibility ordering and the elimination ordering match. Also, the plausibility ordering might not be able to distinguish between more maximal states. For now, I will assume the ordering to yield only one maximal state, namely s_n. Assume the following convention: when s_i is on the left of the turnstile ($\vDash$) then it denotes a state in the model. When it is on the right of the turnstile it has the following meaning: "At state s_i an exam is going to be given". Also, $M, s_i \vDash s_i \vee \neg s_i$, so the agents consider it possible that no exam takes place, but $M, s_i \nvDash s_j$, for any $j \neq i$.

Assume that $M \vDash Bexam$. Then $M, s_n \vDash exam \Rightarrow M, s_n \vDash s_n$. Now remark that if we eliminate any state different than s_n, say s_i (i.e. if we perform any hard update with $\neg s_i$), then in the new model, Bs_n will also hold. That is, $M \vDash Bs_n \Rightarrow M|s_i \vDash Bs_n$. So eliminating anything other that s_n does not change the (initial doxastic) fact Bs_n. Assume then that in the elimination ordering s_n is eliminated after the sequence $s_k, \ldots, s_m$. Then $M|s_k, \ldots, s_m \vDash Bs_n$, and obviously $M|s_k, \ldots, s_m, s_n \vDash Bs_n$. But then $M, s_n \vDash [! \bigwedge_{k \leq i \leq m} \neg s_i] Bs_n$ (since all of $k, \ldots, m$ are different from n). Assume that $M, s_n \vDash surprise$. Then $M, s_n \vDash [! \bigwedge_{k \leq i \leq m} \neg s_i] \neg Bs_n$. So then $M, s_n \vDash [! \bigwedge_{k \leq i \leq m} \neg s_i] \bot \leftrightarrow M, s_n \vDash \bigwedge_{k \leq i \leq m} \neg s_i \rightarrow \bot$. Therefore $M, s_n \vDash \bigvee_{k \leq i \leq m} s_i$. Contradiction! Therefore $Bexam \Rightarrow \neg Ksurprise$

Assume now that the plausibility ordering yields more than one maximal state. Let us denote these states as $s_{n_1}, \ldots, s_{n_l}$. These states are however, eliminated in a strict order, and let us assume that order to be $s_{n_l}, \ldots, s_{n_1}$. Then after the elimination of $s_{n_1}, \ldots, s_{n_{l-1}}$, we can repeat the argument above only that we will substitute s_{n_l} for s_n. The same conclusion follows.

□

Proof. Assume $M \vDash K_{students} NEXT surprise$,

1. $M|exam \vDash Ksurprise \wedge Bexam \Rightarrow M|exam \vDash \bot$

2. $M \Uparrow exam \vDash Ksurprise \wedge Bexam \Rightarrow M \Uparrow exam \vDash \bot$

3. $M \uparrow exam \vDash Ksurprise \wedge Bexam \Rightarrow M \uparrow exam \vDash \bot$

4. Since $M \nvDash K \neg exam$, then $M * exam \vDash Ksurprise \wedge Bexam \Rightarrow M * exam \vDash \bot$

□

Proposition 7.1. *There is a MAGM-governed revision policy, *exam which does not lead to a contradiction.*

Proof. Assume $M \vDash K_{students} NEXT surprise$. We can write the MAGM norm as: $[*exam] \neg K \neg exam \Rightarrow [*exam] Bexam$. However, $M * exam \vDash Ksurprise \wedge Bexam \Rightarrow M * exam \vDash \bot$. So, $M \vDash \neg [*exam] \neg K \neg exam \Leftrightarrow M \vDash \langle *exam \rangle K \neg exam$. This no longer contradicts the assumption that $M \vDash K_{students} NEXT surprise$, and obviously there is only one such revision possible. Denote it by $!!\varphi$, and define $M' = M!!\varphi$ as:

$$\mathcal{D}(M') = \mathcal{D}(M) - \{s \in \mathcal{D}(M) : M, s \vDash \varphi\}$$

$$R' = R \cap \{(s,t) : s \in \mathcal{D}(M') \wedge t \in \mathcal{D}(M')\}$$

$$V' = V$$

□

Therefore there is a way for the students to make sense of the teacher's announcement, while respecting their norm hierarchy, and this is to know that he has lied and that there will be no exam. Nevertheless, an exam does come on the third day (as we assumed), and the students really are surprised. What is more, they even deleted the actual world from their information state. So, what went wrong? Baltag's strategy cannot offer an answer to this question.

For making it clear that this strategy is exactly Baltag's strategy I will highlight the corresponding steps in the following table.

The *Surprise* Examination	The Surprise *Examination*
$K_{students} exam$	$K_{students} NEXTsurprise$
$Bsurprise \Rightarrow \neg Kexam$	$Bexam \Rightarrow \neg Ksurprise$
Teacher announces $\ast NEXTsurprise$	Teacher announces $\ast exam$
$[!NEXTsurprise]FALSE$	$[!exam]FALSE$
$[\Uparrow NEXTsurprise]FALSE$	$[\Uparrow exam]FALSE$
$[\uparrow NEXTsurprise]FALSE$	$[\uparrow exam]FALSE$
$[\ast_{FAGM}(NEXTsurprise)]FALSE$	$[\ast_{FAGM} exam]FALSE$
$\langle \ast_{MAGM}(NEXTsurprise)\rangle K\neg surprise$	$\langle \ast_{MAGM} exam\rangle K\neg exam$
Only one such upgrade: T	Only one such upgrade: !!
$K_{students}\neg surprise$	$K_{students}\neg exam$
NO PARADOX!	PARADOX!

The moral of **Example 1** is then that the order in which the *exam* and *surprise* sentences are announced to the students does not make a difference. Baltag's solution, however, is not flexible enough to prevent the paradox from emerging in the **The Surprise *Examination***, despite being able to do so in the case of **The *Surprise* Examination**. Therefore, by **ALL**, Baltag's solution is not the ”real” solution to the Surprise Examination Paradox.

***

In conclusion, I have formulated two arguments against Baltag's solution. One the one hand, there is a motivation for not being satisfied with Baltag's solution from an external point of view since his solution implies that either the paradox is paradoxical or that the teacher is not ideal. On the other hand, there is a more internal motivation since his strategy of solving The Surprise Examination Paradox does not apply to all its variations, in particular to **The Surprise *Examination***.

7.3 Gerbrandy's solution face to face with The Surprise *Examination* Paradox

A question that can be asked is how Gerbrandy's solution would deal with **The Surprise *Examination* Paradox**. Of course, this problem is of purely theoretical interest. Even if we establish that Gerbrandy's solution does solve **The Surprise *Examination* Paradox**, it would still be the case that his solution is wrong due to its inability to solve **Tomorrow's *Surprise* Examination**.

I believe that Gerbrandy's solution to **The Surprise *Examination* Paradox** corresponds to Baltag's solution presented above just that Gerbrandy's solution stops at $[!exam]\bot$. The only problematic thing is that in order to get the previous formula we assumed that $K_{students}surprise$ and Gerbrandy's formalization of surprise is a bit different. However, I believe that once faced with this scenario, Gerbrandy too would agree that the students will have to continue knowing the rules of the school even after the teacher's announcement, and that moreover, any announcement that the teacher makes the rules still apply. Hence, I believe that it is safe to assume that even Gerbrandy would agree that $K_{students}surprise$. But then **The Surprise *Examination* Paradox** is indeed a counterexample to Gerbrandy's solution also.

8 Conclusions and further work

So far, I believe that it is clear that both Gerbrandy and Baltag's solutions fail to be general enough to solve all the scenarios that philosopher's intuitions identify as the surprise examination paradox. By ALL, this means that they fail to be solutions to the surprise examination paradox. However, I believe that more than this can be said and I will give, below, some further ideas on the topic.

8.1 The Cognitive Structure of Surprise

Gerbrandy and Baltag consider surprise as the clash between not believing φ and φ actually being the case. But this conception of surprise is too wide. For example, I believe neither that Alice in Wonderland is playing at the Pathe cinema in Groningen, nor that it is not playing. Visiting Groningen, I will notice that it is actually playing. Will I be surprised? I think not. This phenomenon extends to most things we learn everyday. By Gerbrandy and

Baltag's and way of defining 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 a certain (true) information σ is new then, obviously, $\sigma \wedge \neg B\sigma$.

However, it is hardly the case that things we consider possible surprise us when we learn that they are actually true. Therefore, surprise needs a more narrow characterization, and the most suitable candidate is the clash between believing that $\neg\varphi$ and φ actually ocurring. If I believe that Alice in Wonderland is not playing at the Pathe in Groningen (say because I believe that the cinemas in Groningen only show Dutch movies and Alice in Wonderland is not Dutch), I will truly be surprised to learn that it is actually playing there. Also, learning new information will no longer be surprising unless that information seemed impossible before learning it.

I consider this to be a correct definition of surprise. It also fits nicely with the analysis of Lorini and Castelfranchi (2007), who based on the cognitive literature on surprise attempt to offer a logical framework to express its meaning. Their agents are resource-bounded, but if one lifts the constraints they put on their agents, and transforms them in ideal agents, their definition of Mismatch-Based Surprise would be exactly $\varphi \wedge B\neg\varphi$.

8.2 Logics for inferences

The above argument against Baltag's solution is weakened by a general shortcoming of this thesis, and for which I do not yet have a solution. As I mentioned before I do not present any real argument that **The Surprise *Examination*** is the same as **The *Surprise* Examination** or that they at least have to be solved in the same manner. Instead, I am taking the philosophical high-road of referring you to your intuitions. This is unpleasant to me even as a (would-be) philosopher. The reason for this shortcoming is that offering a clear criterion that could match our intuitions on which scenario is the same as the Surprise Examination Paradox scenario is tricky business. The differences between them is huge, but still intuitions tell us that they are in some respect the same. I offer some examples below. My best guess at this point is that what stays more or less the same is the reasoning of the students. And I am currently thinking whether Fernando Velazquez-Quesada's logic for inferences can express the way in which the students reason. If this is the case, then I see two possible outcomes, not necessarily disjoint: (i) by (fully) formalizing the reasoning of the students we come upon a wrong step, and then we have a very nice solution to the paradox, namely that the students err when they infer x from y; or (ii) we find a very nice way of characterizing the paradox, and even come up with a nice methodology of characterizing paradoxes by the reasoning they induce in the agents confronted with the scenario of the paradox.

8.3 Some strange way of upgrading

Another way of solving the paradox is by assigning the students a doxastic norm different from the AGM-type norms Baltag is assigning them. Such a norm might be “Do anything it takes to come to know what the teacher is announcing”. Such a norm would then correspond to full trust in the teacher. An upgrade that would fulfill this norm would be $\circledR\varphi$:

$M, w \vDash [\circledR\varphi]\psi$ iff $M\circledR\varphi, w \vDash \psi$, where

1. $\mathcal{D}(M\circledR\varphi) = \mathcal{D}(M)$
2. $R^{M\circledR\varphi} = R^M - \{(w,w) \in \mathcal{D}(M) \times \mathcal{D}(M) : M, w \nvDash \varphi\}$
3. $\mathcal{V}^{M\circledR\varphi} = \mathcal{V}^M$

However, this is not the only possible way, so there is a problem with what to do with the other upgrades that would work. Also, after this upgrade, the new model is based on a plausibility frame which is not reflexive. This might be problematic for two reasons: (i) if you define epistemic accessibility, as Baltag usually does, as $\sim=\leq \vee \geq$, then knowledge comes out non-factive; (ii) if you define knowledge separately, then if you want the principle: $B\varphi \rightarrow KB\varphi$, the same problem as before arises.

References

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

[2] A. Baltag 2003 Logics for Communication: reasoning about information flow in dialogue games, Course at *NASSLLI 2003*, http://www.indiana.edu/ nasslli/2003/program.html, 153 slides

[3] A. Baltag 2009 SURPRISE!? An Answer to the Hangman, or How to Avoid Unexpected Exams!, Lecture presented at the *Logic ad Interactive Rationality Seminar* in Groningen, 69 slides

[4] A. Baltag 2010 Dynamic-doxastic norms versus doxastic-norm dynamics, Lecture presented at *Formal Models of Norm Change 2* in Amsterdam, http://www.cs.uu.nl/events/normchange2/program.html, 82 slides

[5] J.F.A.K. van Benthem Forthcoming *Modal logic for open minds*, CSLI

[6] J.F.A.K. van Benthem Forthcomong *Logical Dynamics of Information and Interaction*

[7] R. Binkley 1965 The Surprise Examination in Modal Logic *The Journal of Philosophy* 65: 127-136

[8] T. Chow 1998 The Surprise Examination or Unexpected Hanging Paradox *The American Mathematical Monthly* 105: 41-51

[9] H.v. Ditmarsch, B. Kooi 2006 The secret of my success *Synthese* 153: 201-232

[10] H.v. Ditmarsch, B. Kooi, W.v.d. Hoek 2007 *Dynamic Epistemic Logic*, Springer

[11] J. Gerbrandy 1999 *Bisimulations on Planet Kripke*, PhD Thesis

[12] J. Gerbrandy 2007 The Surprise Examination in Dynamic Epistemic Logic *Synthese* 155: 21-33

[13] D. Kaplan, R. Montague 1960 A Paradox Regained *Notre Dame Journal of Formal Logic* 1: 79-90

[14] K. Levy 2009 The Solution to the Surprise Exam Paradox *The Southern Journal of Philosophy* ?: 131-158

[15] E. Lorini, C. Castelfranchi 2006 The unexpected aspects of surprise *International Journal of Pattern Recognition and Artificial Intelligence* 20: 817-833

[16] E. Lorini, C. Castelfranchi 2007 The cognitive structure of surprise: looking for basic principles *Topoi* 26: 133-149

[17] W.v.O. Quine 1953 On a So-Called Paradox *Mind* 62: 65-67

[18] R. Shaw 1958 The paradox of the unexpected examination *Mind* 67: 382-384

[19] R. Sorensen 1988 *Blindspots* Clarendon Press

[20] T. Williamson 2000 *Knowledge and its limits* Oxford University Press

[21] C. Wright, A. Sudbury 1977 The paradox of the unexpected examination *Australasian Journal of Philosophy* 55: 41 58