Paradox, Belief Change and Fixed Points:

How to Avoid Unexpected Exams

Sonja Smets and Alexandru Baltag

(ILLC, University of Amsterdam)

The Surprise Examination Paradox

It is **known** that the date of the exam has been fixed in **one of the five (working) days** of next week. But the students don't know in which day.

Then the Teacher announces her students that the **exam's date will be a surprise**:
even in the evening before the exam, the students **will still not be sure** that the exam is tomorrow.

"*Be sure*" is here to be interpreted as "**believing with (high enough) certainty**".

Paradoxical Argumentation

Intuitively, one can prove (by backward induction, starting with Friday) that, IF this announcement is true, then the exam cannot take place in any day of the week.

So, using this argument, the students come to "know" that the announcement is false: the exam CANNOT be a surprise.

GIVEN THIS, they dismiss the announcement, and... THEN, whenever the exam will come (say, on Tuesday) it WILL indeed be a complete surprise!

MODELS FOR BELIEF: Plausibility Models

A **plausibility model** consists of:

- a *finite set* S of "**states**" (or "*possible worlds*");
- a **total preorder** $\leq \subseteq S \times S$, called *plausibility relation*;
- a **valuation map**, assigning to each atomic sentence p (in a given set At of atomic sentences) some set $\|p\| \subseteq S$.

"**Preorder**": *reflexive* and *transitive*.
Total (="Connected"): $\forall s, t(s \leq t \vee t \leq s)$.

Totality will not play an essential role here, and can be dropped.

Interpretation

A plausibility model encodes exactly what people Belief Revision call **the "epistemic state" of the agent**:

her beliefs, knowledge, belief-revision plans etc,
i.e. *all that is available to the agent by pure introspection.*

Read $s \leq t$ as "**state s is at least as plausible as state t**".

Information as "(Irrevocable) Knowledge"

We say that **the agent has the (implicit) information that** φ iff φ **is true in all the possible worlds of the model**:

$$\|\varphi\|_{\mathbf{S}} = S.$$

Following Game Theorists and Computer Scientists, we may call this implicit possession of information "**knowledge**", and denote by $K\varphi$.

But note that this is an "absolute" sense of knowledge, which doesn't have much to do with the mental state of a real agent: it is an *absolutely certain, necessarily true, absolutely unrevisable* kind of "knowledge", capturing all the information that can be thought (from an objective, external point of view) to be potentially available to the agent.

Belief

A sentence φ is **believed**, and we write $B\varphi$, if φ **is true in all the "most plausible" worlds**; i.e. we have

$$Max_{\leq}S \subseteq \|\varphi\|_{\mathbf{S}},$$

where $MaxS$ is the set of all "maximal" states

$$Max_{\leq}S := \{s \in S : s \not< t \text{ for any } t \in S\}.$$

NOTE: IF we assume the connectedness of $\leq$ and the finiteness of S, then we can simplify this to

$$Max_{\leq}S := \{s \in S : t \leq s \text{ for all } t \in S\}.$$

Conditional Belief

More generally, a sentence φ is **believed conditional on** ψ, in which case we write $\mathbf{B}(\varphi|\psi)$, if φ **is true at all most plausible worlds satisfying** ψ; i.e.

$$Max_{\leq}\|\psi\|_{\mathbf{S}} \subseteq \|\varphi\|_{\mathbf{S}},$$

where $Max_{\leq}\|\psi\|_{\mathbf{S}}$ is the set of all maximal φ-states:

$$Max_{\leq}\|\psi\|_{\mathbf{S}} := \{s \in \|\psi\|_{\mathbf{S}} : s \not\leq t \text{ for any } t \in \|\psi\|_{\mathbf{S}}\}.$$

Again, finiteness and connectedness would allow us to simplify this to:

$$Max_{\leq}\|\psi\|_{\mathbf{S}} := \{s \in \|\psi\|_{\mathbf{S}} : t \leq s \text{ for all } t \in \|\psi\|_{\mathbf{S}}\}.$$

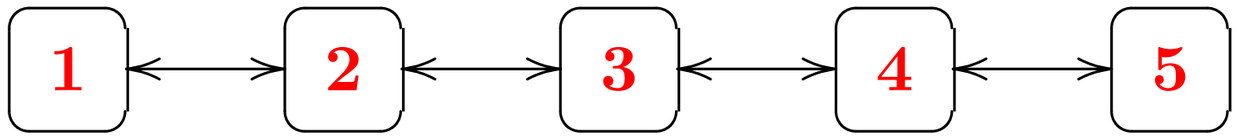
Contingency Plans for Belief Change

We can think of conditional beliefs $B(\varphi|\psi)$ as ‘‘*strategies*”, *or* *“contingency plans” for belief change*:

in case I will find out that ψ was the case, I will believe that φ was the case.

Example 1

An example of a model is

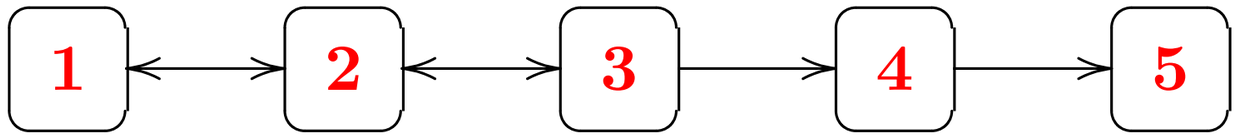


where i means that: the exam takes places in the i-th (working) day of the week.

This encodes an initial situation in which the student considers all days as being *equally plausible* dates for the exam.

Example 2

In the model



the student believes the exam will take place on Friday. But, if given the information that this is not the case, the student would believe the exam will take place on Thursday. If given the information that none of the above is the case, the student just considers all the other days (Monday, Tuesday, Wednesday) as equally plausible.

ACTUAL Belief Change: Upgrades with φ

A **belief upgrade with (a sentence)** φ is a *model transformer* $T\varphi$, that takes *any* (plausibility) model $\mathbf{S}$, and returns a *new* model $T\varphi\,(\mathbf{S})$, having:

- as new set of worlds: some *subset* $S' \subseteq S$,
- as new plausibility relation: some other total preorder $\leq'$.

The transition map associated to an upgrade

Such an upgrade $T\varphi$ induces a **partial map** on the set of states S of any model **S**, map also denoted by $T\varphi : S \to S$, and given by

$$T\varphi(s) = s \ , \ \text{ iff } s \in S',$$

and

$$T\varphi(s) = \ \text{undefined} \ , \ \text{otherwise.}$$

Upgrades on Pointed Models

By putting together the upgrade with the induced transition map, we can think of an upgrade $T\varphi$ as **a partial map on pointed models** $\mathbf{S} = (\mathbf{S}, s)$, given by:

$$T\varphi(\mathbf{S}, s) = (T\varphi(\mathbf{S}), T\varphi(s)).$$

Hard and Soft Upgrades

An upgrade $T\varphi$ is called **soft** if, for every model $\mathbf{S}$, the map $T\varphi : S \to S$ is *total*; i.e. iff

$$S' = S$$

for all $\mathbf{S}$. A soft upgrade *doesn't add anything to the agent's irrevocable knowledge*: it *only conveys "soft information"*, changing only the agent's beliefs or his belief-revision plans.

In contrast, a **hard** upgrade adds new knowledge, by shrinking the state set to a *proper subset* $S' \subset S$.

Dynamic Operators

We can add to the language, in the usual way, **dynamic operators** $[T\varphi]\psi$ to express the fact that ψ **will surely be true** (in the new model) **AFTER the upgrade** $T\varphi$.

But one can go on and introduce "**temporal plausibility models**", which can be identified with **sequences**

$$\mathbf{S}_0, \mathbf{S}_1, \ldots, \mathbf{S}_n, \ldots$$

of plausibility models obtained by successive upgrades $T_0\varphi_0, T_1\varphi_1, \cdots$:

$$\mathbf{S}_{n+1} = T_n\varphi_n(\mathbf{S}_n).$$

Temporal Operators

Given such a (deterministic) temporal plausibility model, one can capture the same thing as $[T_n\varphi_n]\psi$ at a given state $s \in \mathbf{S}_n$ (for some n), without any direct reference to T_n, by using a **temporal "next" operator**:

$$NEXT\,\psi = [T_n\varphi_n]\psi.$$

Dually, one can introduce a **past-tense operator**

$$BEFORE\,\psi$$

which is true at a state s in a plausibility model $\mathbf{S}_n$ iff s satisfies ψ in $\mathbf{S}_{n-1}$.

Examples of Upgrades $T\varphi$ with a sentence φ

(1) Update $!\varphi$:

all the non-φ states are deleted and *the same plausibility order is kept between the remaining states.*

(2) Radical upgrade $\Uparrow \varphi$:

all φ-worlds become "better" (more plausible) than all $\neg\varphi$-worlds, and *within the two zones, the old ordering remains.*

(3) Conservative upgrade $\uparrow \varphi$:

the "best" (most plausible) φ-worlds become better than all other worlds, and *in rest the old order remains.*

Different attitudes towards the new information

These transformations correspond to *different possible attitudes* of the learner towards *the reliability* of the source of information:

- **Update**: an **infallible** source. The source is *"known" (guaranteed) to be always truthful.*
- **Radical upgrade**: strong trust. The source is **fallible, but highly reliable**, or at least *strongly believed to be truthful.*
- **Conservative upgrade**: the source is **trusted, but only "barely"**. The source is *("simply") believed to be truthful*; but this belief can be easily given up later!

Explanation continued

After a *conservative or a radical upgrade*, the agent only comes to **believe** that φ (was the case), **unless he already knew** (before the upgrade) that φ was false; i.e. we have the validity

$$\neg K \neg \varphi \Rightarrow [\uparrow \varphi] B(BEFORE\, \varphi)$$

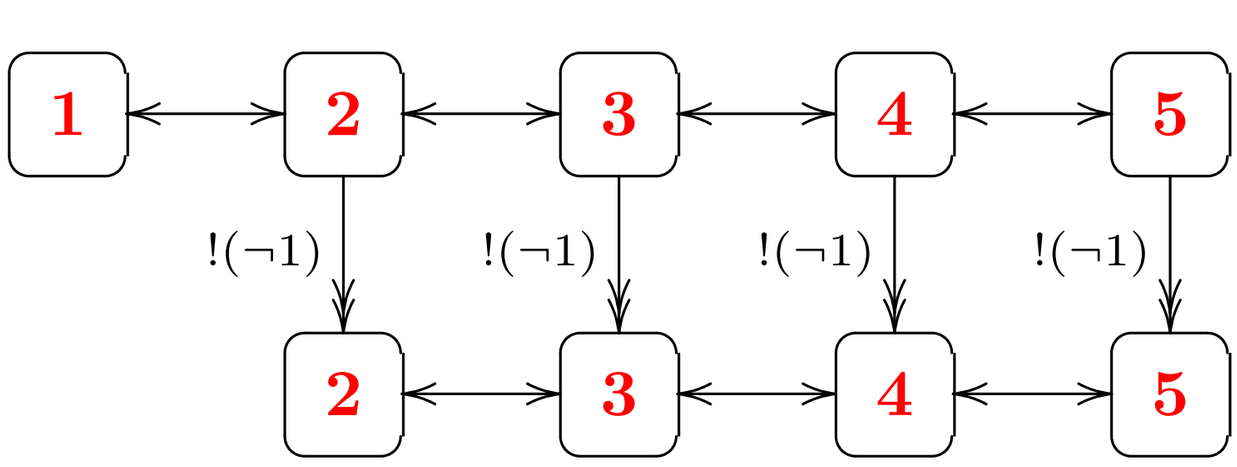
After an *update*, the agent comes to "**know**" φ, so that *all non-φ possibilities are forever eliminated*: we have the validity

$$[!\varphi] K(BEFORE\, \varphi).$$

Finally, after any *negative* update/upgrade, the agent comes to know/believe that that φ was false.

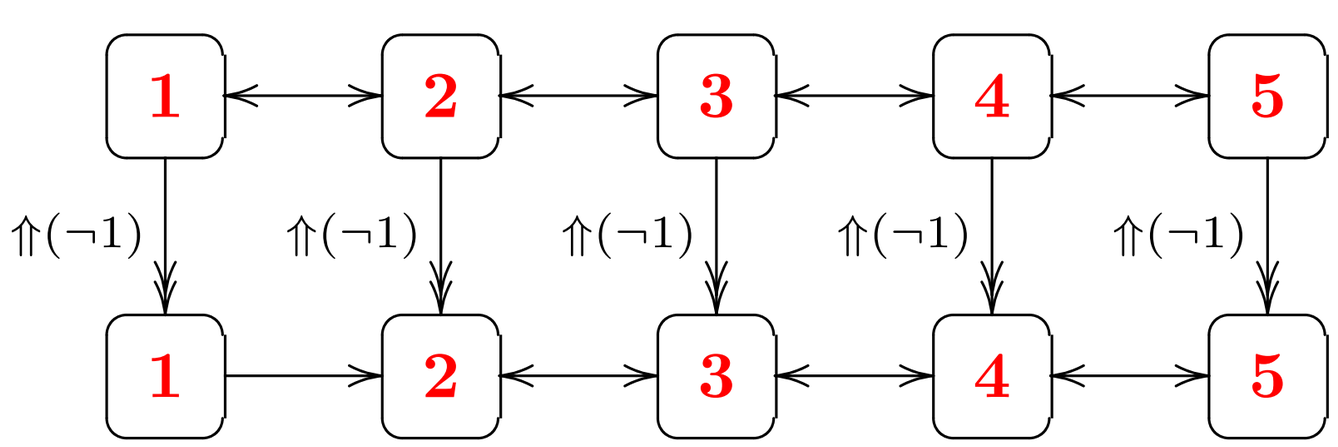
Example 3

Suppose that, in the model in Example 2, no announcement is made by the teacher, but day 1 (Monday) simply passes and no exam has yet taken place. This is an **update** $!(\neg 1)$, inducing a transition $\stackrel{!(\neg 1)}{\Longrightarrow}$ to a new plausibility model with only 4 possible worlds:



Example 4

In contrast, suppose that in the model in Example 2, *a HIGHLY TRUSTED, BUT NOT INFALLIBLE Teacher announces that the exam will not be on Monday.* Then this is a *radical upgrade* $\Uparrow(\neg 1)$, inducing a transition to a model with the same 5 worlds:



Dynamics

The **reduction law for belief under updates** makes essential use of *conditional belief*:

$$[!\psi]B\varphi \quad \longleftrightarrow \quad \psi \to B^{\psi}[!\psi]\varphi\,.$$

This is **an example of the "pre-encoding strategy"**:

to get reduction laws for (simple) belief, we needed to pre-encode its behavior under updates in the initial (static) logic, by extend it with conditional belief operators.

A similar strategy works for other upgrades.

The Logic of updates and conditional beliefs

Of course, for a complete axiomatization, we need to get in turn a *reduction law for conditional belief*!

Luckily, the buck stops here: **conditional belief can pre-encode its own behavior**.

The **logic of updates and conditional beliefs** can be completely axiomatized by adding the standard Reduction Laws for atoms, negation, conjunction and knowledge, together with the following reduction axiom for conditional belief:

$$[!\psi]B^{\theta}\varphi \quad \longleftrightarrow \quad \psi \rightarrow B^{\psi \wedge [!\psi]\theta}[!\psi]\varphi\,.$$

Doxastic Attitudes: "static" and "dynamic"

A **"static" doxastic attitude** $A\varphi$ captures an agent's *opinion about some sentence* φ; e.g. the agent **knows** φ: $K\varphi$; the agent **believes** φ: $B\varphi$; the agent **strongly believes** φ: $Sb\varphi$ etc.

There are other possible attitudes: **defeasible knowledge** (=*safe belief*) $\Box\varphi$, *knowledge-to-the-contrary* $K_i\neg\varphi$ etc.

But there also exist doxastic attitudes of a **"dynamic" nature**, governing an *agent's belief revision policy towards any information coming from a given source.*

Dynamic Attitudes

We saw that the way an agent *revises her beliefs after receiving some new information* depends on the agent's doxastic **attitude towards the source of information**.

Hence, such a "dynamic" attitude captures the **agent's opinion about the reliability of information coming from this particular source.**

Dynamic Attitudes "are" Types of Upgrades

Each dynamic-doxastic attitude will thus correspond to a specific "type" of doxastic transformer: i.e. a map

$$\varphi \ \mapsto \ T\varphi$$

associating to any input-sentence φ some doxastic upgrade $T\varphi$.

The meaning of this is that, whenever receiving information φ from this specific source, our agent will revise her beliefs by applying transformation $\tau\varphi$ to her plausibility relation $\leq$.

Examples of "Positive" Attitudes

We saw that **(1) update, (2) radical upgrade** and **(3) conservative upgrade** correspond to *three different possible attitudes* of the learners towards *the reliability* of the source.

But, intuitively, all of them were **positive** attitudes: the new information was accepted (unless it contradicted previous knowledge).

In contrast, there also exist **negative** ones:

(4) **Negative Update** $!^{-}\varphi$: an **infallible source of falsehoods**. The source is *"known" (guaranteed) to be always lying.*

In this case, **all the** φ **states are deleted** and *the same order is kept between the remaining states.*

More Examples: Negative Attitudes

(5) Negative Radical upgrade $\Uparrow^{-} \varphi$:
all $\neg\varphi$-worlds become "better" (more plausible) than all φ-worlds, and *within the two zones, the old ordering remains.*

This reflects **strong distrust**: **the listener strongly believes the speaker is lying**.

(6) Negative Conservative upgrade $\uparrow^{-} \varphi$:
the "best" $\neg\varphi$-worlds become better than all other worlds, and *in rest the old order remains.*

This reflects **relative distrust**: the listener barely believes the speaker is lying.

Another Example: Neutrality

(7) Doxastic Neutrality $id\varphi$ is the attitude according to which the source cannot be trusted nor distrusted: the listener **simply ignores** the new information φ, keeping her old plausibility order as before.

This is the **identity map** id on plausibility models.

So the source is **neither believed nor dis-believed, but simply ignored**. The agent *keeps the old beliefs.*

Other Examples: Mixed Attitudes

An agent's attitude towards a source of information might **depend on the type of information** received from that source: she might treat differently different types of information. She might **mix** two or more basic transformers, using **semantic or syntactic conditions** to decide which to apply.

For instance, the agent may strongly trust the source to be right about sentences belonging to a given sublanguage L_0, while she may only barely trust it with respect to any other announcements. This attitude could be denoted by $\Uparrow_{L_0}\uparrow$.

Example

If the "source" is a well-known Professor of Mathematics, our agent may accept him as *an infallible source of mathematical statements*, and thus perform an update $!\varphi$ whenever the professor announces a sentence φ about Mathematics.

In any other case, our agent might treat the new information coming from the professor more cautiously (say, barely believing it $\uparrow \varphi$, or even ignoring it, and thus applying id): indeed, a typical mathematician may be utterly unreliable concerning any other area of conversation except for Mathematics!

What is a "Positive" Attitude?

When exactly can we say that the agent adopts a "positive" attitude?

What is the demarcation line between "positive" attitudes and the others ("neutral", "negative" or "mixed" ones)?

Coming to Believe?

All the above "positive" types of upgrade $T\varphi$ have the property that, after the upgrade, the agent **comes to believe** that φ was true (before the upgrade), **unless he already knew** (before the upgrade) that φ was **false**:

$$\neg \mathbf{K} \neg \varphi \Longrightarrow [\mathbf{T}\varphi]\mathbf{B}(\mathbf{BEFORE}\,\varphi).$$

Semantically, this corresponds to requiring that the upgrade $T(S, \leq) = (S, \leq')$ satisfies:

$$\|\varphi\|_{\mathbf{S}} \neq \emptyset \Longrightarrow \mathbf{Max}_{\leq'}\mathbf{S}' \subseteq \|\varphi\|_{\mathbf{S}}\,.$$

Is **this** the characteristic of "positive" attitudes?

Counterexample

Let's assume that Alice has a generally positive attitude towards Bob as a source of information: she is in general willing to believe that whatever Bob tells her is true. (Bob is a good friend, and she has no reason to suspect that he will be lying to her or pulling her leg.) Also, Alice has a good knowledge of Romanian language, but she doesn't know whether or not Bob speaks any Romanian.

Then Bob comes and tells her:

> "Nu stiu nici o vorba Romaneasca!"
>
> (I don't know any Romanian word)

Explanation

The sentence is *perfectly consistent with Alice's prior knowledge.*

Nevertheless, Alice obviously *CANNOT come to believe this sentence, despite her willingness to believe*: the fact that this sentence is uttered by Bob proves to Alice that the sentence was false!

Note, unlike Moore sentences, *this sentence does not change its truth value.* But, when our agent receives this information from this particular source (Bob), she instantly learns the opposite of what the sentence claims!

Positive attitudes

So **certain sentences, by their utterance, may become known to be false, even if they we were not so known before**. Somehow they are proven to be false during (and due to) the very attempt (by a trusting agent) to revise with them.

So what we should require for "positive" attitudes is that the new φ **is BELIEVED AFTER the upgrade, UNLESS the agent COMES TO BE KNOWN that it's FALSE**. the following logical law:

$$[\mathbf{T}\varphi]\neg\mathbf{K}\neg(\mathbf{BEFORE}\,\varphi) \Longrightarrow [\mathbf{T}\varphi]\mathbf{B}(\mathbf{BEFORE}\,\varphi).$$

Willingness to revise

This logical validity express a *willingness to revise*:
after the upgrade $T\varphi$, the new information is believed (to have been true at the moment when it was learnt) **UNLESS this is inconsistent with the agent's knowledge (AFTER the upgrade)**.

Essentially, this means that the agent does AT LEAST **ATTEMPTS** to perform a belief revision with φ: he has AT LEAST a *MINIMAL trust* in the new information: and so he comes to **believe** that it was correct **unless** he comes to positively **know** that it was incorrect.

"Positive" (Dynamic) Attitude

An agent is said to have a **positive (dynamic) attitude** towards a source of information if, whenever any piece of information φ is received from that source, the agent's prior plausibility structure $(S, \leq)$ is changed to a new structure $(S', \leq')$ by applying a doxastic *upgrade* (model transformer) $T\varphi$ satisfying

$$\|\varphi\|_{\mathbf{S}} \cap \mathbf{S'} \neq \emptyset \implies \mathbf{Max}_{\leq'}\mathbf{S'} \subseteq \|\varphi\|_{\mathbf{S}}$$

Static Attitudes as Fixed Points

To each dynamic attitude T we can associate in a canonical way a static attitude T^∞, defined by:

$T^\infty\varphi$ holds at a state s in a model $\mathbf{S}$ iff the pointed model $(\mathbf{S}, s)$ is a **fixed point** for $T\varphi$; i.e. if the upgraded model is **bisimilar** to the initial one:

$$s \in \|T^\infty\varphi\|_{\mathbf{S}} \text{ iff } T\varphi(\mathbf{S}, s) \sim (\mathbf{S}, s).$$

Informally, $\mathbf{S}$ is a fixed point of an upgrade $T\varphi$ iff $T\varphi$ is **not an "informative" upgrade** (but a "*redundant*" one): the agent doesn't actually learn anything, all her conditional beliefs and her knowledge stay the same.

Positive and Negative Fixed Points

For an upgrade $T\varphi$, a fixed-point model $(\mathbf{S}, s)$ is a "**positive**" fixed point iff it is a fixed point in which *phi* is **true** (at s in $\mathbf{S}$).

A **negative** fixed point $(\mathbf{S}, s)$ is a fixed point in which φ is **false** (at s in $\mathbf{S}$).

Fixed Points of Update: Knowing the Answer

Intuitively, asking a question is redundant iff you **already know** the answer.

This is indeed the *correct characterization for the fixed points of an update*:

A model S is a fixed point of an update $!\varphi$ iff φ is known in the model S; i.e. iff

$$\mathbf{S} \models K\varphi.$$

Fixed Points of $\uparrow\varphi$: Believing the Answer

For other kinds of upgrades though, knowledge is too much to ask.

For instance, the *characterization for the fixed points of a conservative upgrade* is:

A model S is a "positive" fixed point of a conservative upgrade $\uparrow\varphi$ iff φ is believed in (every state of) the model S; i.e. iff

$$\mathbf{S} \models B\varphi.$$

S is a "negative" fixed point of $\uparrow\varphi$ iff φ is known to be false in the model S; i.e. iff

$$\mathbf{S} \models K\neg\varphi.$$

Fixed Points of $\Uparrow\varphi$: Strong Belief

A model S is a fixed point of a radical upgrade $\Uparrow\varphi$ iff φ is "strongly believed" in the model S, i.e. iff all φ-worlds are more plausible than all non-φ-worlds:

$$s < t \text{ , for all } s \in \|\varphi\|_{\mathbf{S}} \text{ and all } t \in S \setminus \|\varphi\|_{\mathbf{S}}.$$

S is a "positive" fixed point of $\Uparrow\varphi$ iff φ is strongly believed in the model S.

S is a "negative" fixed point of $\Uparrow\varphi$ iff φ is known to be false in the model S (i.e. if $\|\varphi\|_{\mathbf{S}} = \emptyset$).

Solutions to Surprise Exam: Gerbrandy's Solution

Jelle Gerbrandy proposed a nice solution to the Surprise Exam puzzle.

Gerbrandy interprets the announcement itself as an "*update*" (="public announcement", conveying **"hard" information**) with the above sentence.

According to this solution, the students' correct conclusion should be only that (if the Teacher tells the truth, then) the exam **won't take place on Friday**; but **none of the previous days can be excluded further**!

"Surprise" according to Gerbrandy

Gerbrandy's interpretation of "surprise" can be encoded as:

$$surprise \;=\; \bigwedge_{1\leq i\leq 5}\left(i \rightarrow [!(\bigwedge_{1\leq j<i} \neg j)]\neg Bi\right).$$

By the Reduction Laws for updates, this is is equivalent to:

$$surprise \;=\; \bigwedge_{1\leq i\leq 5}\left(i \rightarrow \neg B(i|\bigwedge_{1\leq j<i} \neg j\,)\right).$$

In English, this reading of "surprise" is given by:
"**if the exam is in day i then at the end of day $i-1$, the student will still not BELIEVE that the exam is tomorrow**."

surprise is a Moore sentence

It is easy to see that, given the assumption of the story (that it is known that the exam will take place in one of the days), the sentence *surprise* is a Moore-type sentence: even if it is true, **it cannot be believed** (by the Student).

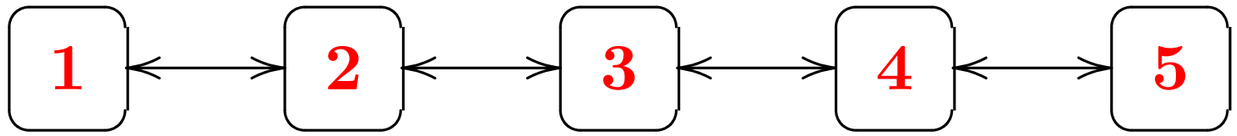
Indeed, the following is a logical validity

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

PROOF: by backward induction (starting with Friday).

NO SURPRISE ON FRIDAY

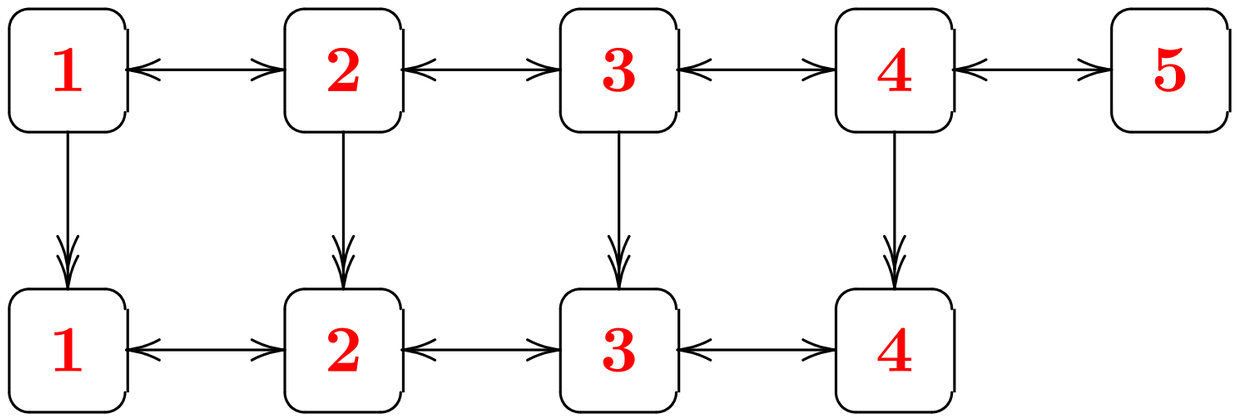
It is easy to see that if initially the Students has complete uncertainty, i.e. we start with the model



then *surprise* is *true at all worlds except Friday* (the world satisfying 5).

Gerbrandy's Solution

Gerbrandy interprets the teacher's announcement as an **update** !(*surprise*) with the sentence *surprise*. So this induces a transition



The conclusion is: if Teacher didn't lie then the exam cannot be on Friday. But this reasoning cannot be iterated: none of the other days can be excluded!

Conclusion?

In conclusion, the Teacher's announcement had a clear truth-value: it was true iff the exam will be in any other day but Friday.

But note that, if say the exam will be on Thursday, then the announcement is NO LONGER true imediately after it was announced:
the sentence "surprise" changed its truth value by being announced!

Indeed, on Wednesday evening, the Student will know that the exam is going to be on Thursday!

If accused of lying, the Teacher can later claim that he DIDN'T mean to say that exam will still be a surprise EVEN after he announced that: he only meant that this was true before the announcement!

Unsatisfactory Solution!

This clearly sounds like cheating to me.

Essentially, Gerbrandy's solution corresponds to interpreting the expression "it will be a surprise" as:
"**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**".

This eliminates the paradox, but only by "cheating". Most people would say that the Teacher lied: his sentence, interpreted in the "natural" way, turned out to be false.

The "self-referential" interpretation

Most people think the natural interpretation of Teacher's announcement is:
"**The exam will be a surprise (even) after I'm telling you ALL THIS**".

But this is a **self-referential** sentence. How can we give it a **meaning**?

Iterated Updates

A way to interpret this self-referential announcement is as being equivalent to **an infinite sequence of (non-self-referential) announcements**

$$!surprise; !surprise; !surprise; \cdots ,$$

i.e. *first* the teacher says "the exam would have been a surprise if I didn't make this very announcement";
then she says "even after the previous announcement, the exam would still have been a surprise if I didn't make this second announcement";
then she repeats *this*, *etc.*

But it's easy to see that **this infinite sequential composition of updates is an "impossible" event, since it leads to paradox**:

- the first update **deletes Friday** from the model,
- the second **deletes Thursday**,
- the third **deletes Wednesday**,
- the fourth **deletes Tuesday**,

- hence **the fifth update will be impossible**;
  since, if possible, it would delete the last world left (Monday);
  thus **contradicting the student's background knowledge**
  (that there will be an exam in one of the week's days). Paradox!

Conclusion

There was an underlying **assumption**: by modelling the announcements as updates, we assumed that the Student has an **absolute trust in the Teacher**, i.e. he considers her as an **infallible source** of (**always truthful**) information.

The contradiction we reached shows this assumption was an error:

a teacher who makes such a self-referential announcement CANNOT be an infallible source;

she *might* tell the truth, but her announcement does NOT come with any inherent warranty of truthfulness.

Lowering Your Trust in Your Teacher

But maybe the student can *lower* his degree of trust?

Then we should interpret this announcement as some other kind of belief **upgrade**, rather than an update?

The Trivial Solution: Unwillingness to revise

The simplest solution (due to Quine) is that the Student adopts the **neutral attitude**, given by the identity upgrade *id*: he will simply **refuses** to revise even if this is consistent with his knowledge. So he'll completely distrust the source of "information" to start with, and hence dismiss the *surprise* announcement out of hand.

So the students sticks ith whatever he believed before, no matter what: assuming he started with total lack of information (considering all days equally plausible), then he keeps this position after Teacher's announcement.

The sentence *surprise* will then be **true** unless the exam is on Friday.

But it will be true for **trivial** reasons: the student didn't learn anything from Teacher, so of course he'll be surprised by the exam.

CONCLUSION: Such a total distrust IS indeed a solution to the Surprise Examination Puzzle, but a completely trivial one.

Non-triviality: willingness to revise

We will henceforth assume that this is NOT the case: we assume the student **starts** with **some (moderate or even minimal) trust** in the Teacher, i.e. he adopts a **POSITIVE attitude towards the Teacher** (as a source of information).

For instance, let us assume that the Student **strongly trusts** the Teacher: so he'll do a **radical upgrade** with the Teacher's announcement.

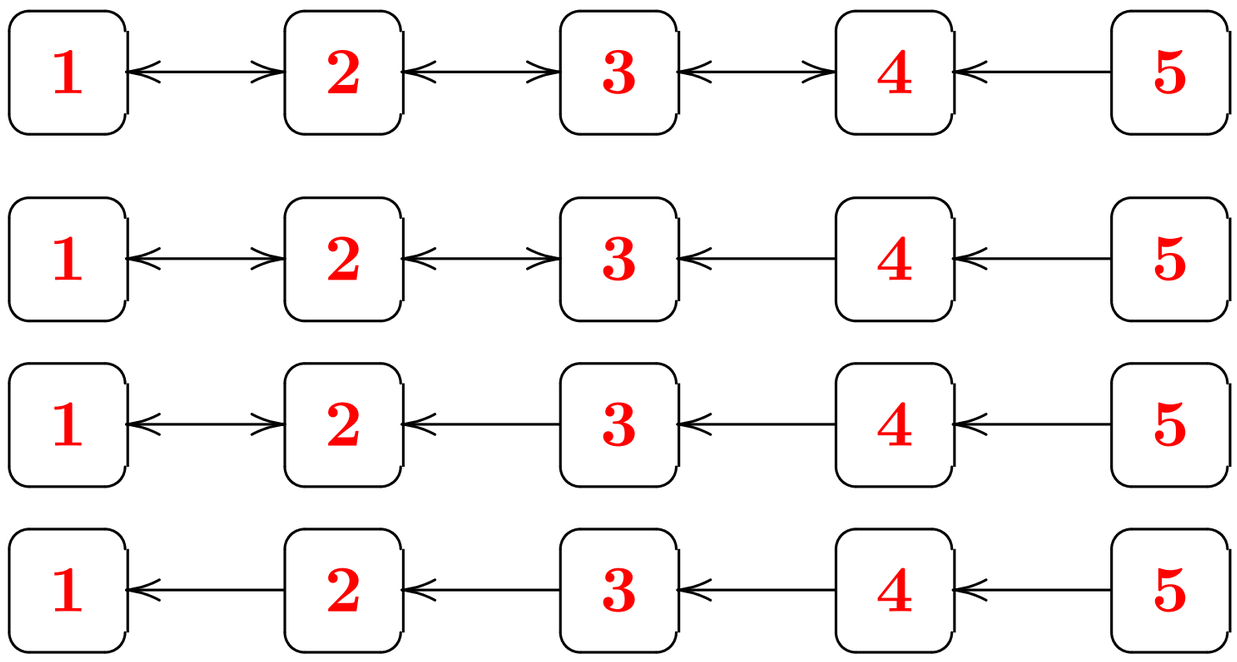
As before, since the announcement was *self-referential* ("The exam will be a surprise EVEN after I am telling you all this"), the Student will have to **iterate** this upgrade.

The "True" Solution: First Proof (by Iteration)

So let's treat this as an *iterated radical upgrade*

$$\Uparrow(surprise); \Uparrow(surprise); \Uparrow(surprise); \cdots$$

applied to an initial model with total lack of info (all days equally plausible). The successive upgrades produce the models:



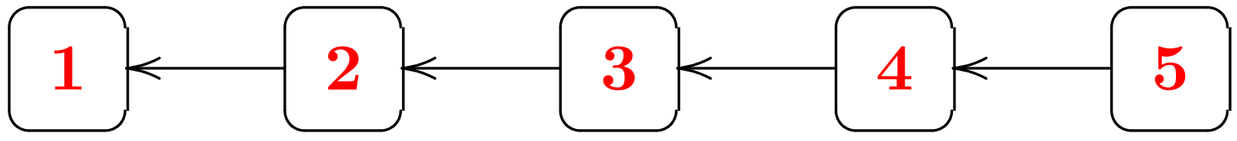
After this, any further iteration leaves the model unchanged!

Note that (unlike the case of iterated hard updates !(*surprise*)), **the infinitely many future upgrades CAN be performed**: they NEVER fail, but they simply stop changing the model. A *fixed point* has been reached.

But **this fixed point is "negative"**: after the fourth iteration, the sentence *surprise* is **known to be false**. **Any further announcements** $\Uparrow(surprise)$ (although executable) **cannot be accepted (believed) by the student**.

Any Positive Upgrade on ANY initial belief model will Do!

Moreover, the conclusion does NOT depend on on the *original plausibility relation* nor on the *the type of positive upgrade* that is being iterated:
starting with ANY initial plausibility relation on the five-day model, and iteratively applying either radical upgrades $\Uparrow(surprise)$ or conservative upgrades $\uparrow(surprise)$, we always eventually reach **the same fixed point**:



Our Solution is "Canonical"

Hence, there is *no ad hoc element*, no arbitrariness in our solution: it is indeed **the ONLY solution (to the self-referential** *surprise* **announcement) compatible with the Student having a positive attitude towards the Teacher**!

Conclusions

Note that, in this last model, the student KNOWS that *the teacher lied*: he knows NOW that *the exam CANNOT be a surprise*. Indeed, no matter what day the exam will be, at the end of the previous day the student will *correctly believe* that the exam is tomorrow! (Moreover, one can show that in fact the student will then "defeasibly know" that the exam is tomorrow!)

QUESTION: Given this conclusion, why can't the student just dismiss the announcement and revert to his original plausibility order?

ANSWER: Of course he can, BUT in this way, he would cancel the reasons behind his own previous conclusion!

Indeed, if he reverts to the original order, then he does NOT know anymore that the teacher lied: the exam might then be a surprise!

YES, he CAN disregard this possibility and stick to the unwarranted belief that the exam won't be a surprise.

But he'd do this only at his own peril: *any retroactive dismissal of the announcement is unwarranted*!

There is NO justification for going back to the original beliefs.

The ONLY way for the student to prevent the exam from being a surprise is **to perform the above upgrade**, and **stick with its conclusion**: the Teacher lied, but nevertheless this is only because his lie DID have an effect on the student (triggering the above upgrade), and this effect WAS justified by the student's initial (modest) trust in the teacher.

There is NO justification for undoing this upgrade.

The *(correct) conclusion that the teacher lied is NOT a warranty for dismissing his announcement* altogether, since *this conclusion was ONLY ensured by the student's change of belief* order as triggered by the announcement.

SECOND ARGUMENT

Another way to reach the same conclusion is to assume that the Student has SOME (yet to be determined!) KIND OF POSITIVE ATITUDE towards Teacher (given by SOME SOFT POSITIVE UPGRADE) T, then ask the QUESTION:

What kind of soft positive upgrade is consistent with Teacher's self-referential announcement?

It is easy to see that T **CANNOT be given by radical NOR conservative upgrade**: both $\Uparrow(NEXT\, surprise)$ and $\uparrow(NEXT\, surprise)$ **lead to paradox**. Given the initial assumptions, these upgrades are "impossible events", similarly to $!(NEXT\, surprise)$.

Impossibility Proof

Put $\varphi := NEXT\, surprise$. The contradiction is reached by recalling that all upgrades $T \in \{!, \Uparrow, \uparrow\}$ build **belief** in $BEFORE\varphi$, given Student's initial lack of knowledge:

$$K(\bigvee_{1\leq i\leq 5} i) \wedge \bigwedge_{1\leq i\leq 5} (\neg Ki \wedge \neg K\neg i) \;\Rightarrow\; [T\varphi]B(BEFORE\varphi).$$

Putting this together with the equivalence

$$BEFORE(NEXT\, surprise) \;\Leftrightarrow\; surprise$$

we get

$$K(\bigvee_{1\leq i\leq 5} i) \wedge \bigwedge_{1\leq i\leq 5} (\neg Ki \wedge \neg K\neg i) \;\Rightarrow\; [T\varphi]B surprise$$

This, combined with the earlier proved validity

$$K(\bigvee_{1\leq i\leq 5} i) \;\Rightarrow\; [T\varphi]\neg Bsurprise\,,$$

gives us

$$K(\bigvee_{1\leq i\leq 5} i) \;\Rightarrow\; [T\varphi]FALSE.$$

Proposed Solution

ANSWER: There **does exist** a soft positive attitude T, leading to an executable upgrade $T(NEXT\,surprise)$ with the sentence $NEXT\,surprise$.

In a sense, this is by now obvious: such an upgrade can be defined e.g. via *the limit of the above-mentioned infinite sequence of iterated upgrades*.

So, in explicit terms, $T(NEXT\,surprise)$ is the upgrade which, when applied to **any** model **S**, induces a plausibility relation **making earlier days to be more plausible than later ones**.

Our Solution is "Canonical"

Moreover, one can show that this answer is **UNIQUE**: **there exists ONLY ONE such soft positive upgrade** with the sentence *NEXT surprise*.

Hence, there is *no ad hoc element*, no arbitrariness in our solution: it is indeed **the ONLY solution for the "soft" version of the puzzle!**

Proof

Let $T\varphi$ be such a soft positive upgrade with the sentence $\varphi := NEXT\, surprise$.

By **our definition of positivity** of an attitude (the "willingness to revise" condition), we obtain that:

$$[T\varphi]\neg K\neg surprise_B \implies [T\varphi]Bsurprise.$$

This, combined with the earlier proved validity

$$K(\bigvee_{1\leq i\leq 5} i) \;\Rightarrow\; [T\varphi]\neg Bsurprise\,,$$

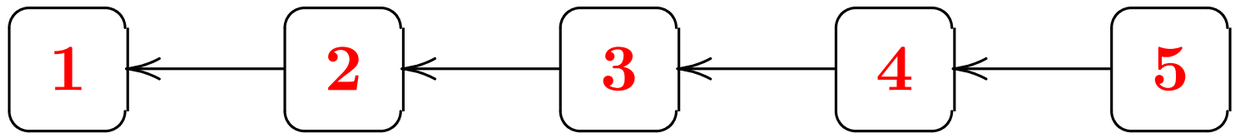
gives us

$$K(\bigvee_{1\leq i\leq 5} i) \wedge [T\varphi]\neg K\neg surprise_B \;\Rightarrow\; [T\varphi]FALSE.$$

This means that such an upgrade $T\varphi$ is **possible (executable) ONLY if** $\neg[T\varphi]\neg K\neg surprise$ holds in the original model, i.e. if $K\neg surprise$ **holds in the new model after the upgrade**.

Finally, it is easy to see that $K\neg surprise$ **holds in a model IFF all the surviving worlds are ordered in their reverse temporal order**: $1 > 2 > 3 > 4 > 5$.

Since T is a *soft* upgrade, the result of applying T to ANY initial model with $S = \{1, 2, 3, 4, 5\}$ is



A THIRD ARGUMENT:

Surprise Examination as a game of imperfect information

Yet another way to argue for the same conclusion is by encoding the "belief" version of Surprise Examination as a game.

The Game

On Sunday evening, player 1 (Student) *forms a belief*: either the "exam is tomorrow" (t) or not ($\bar{t}$).
Next (on Monday morning) player 2 (Teacher) chooses an action, **without knowing the Student's previous move(s)**: either "gives the exam" today (e) or he doesn't ($\bar{e}$). Teacher's moves are visible to all.

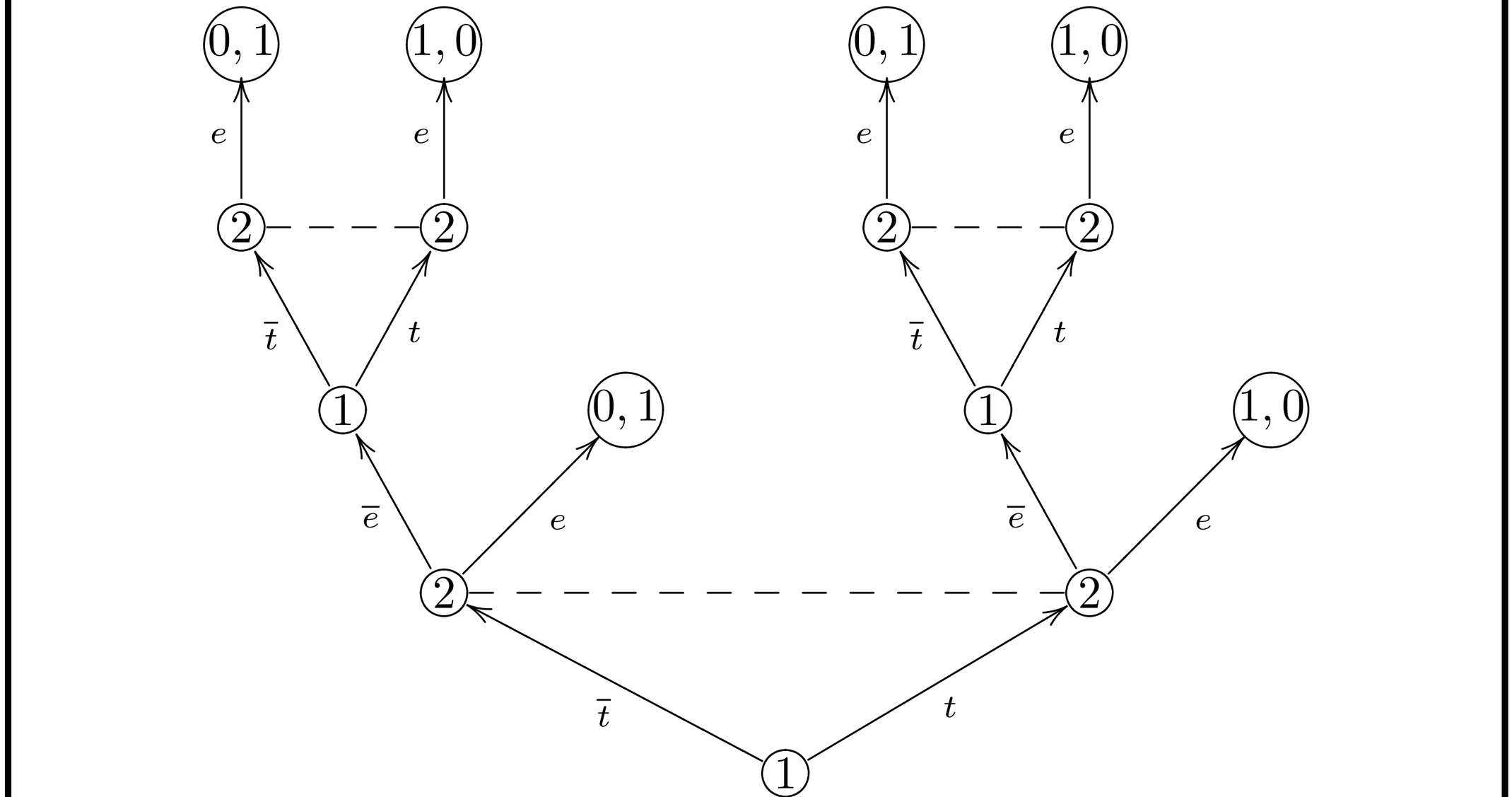
If he gives the exam, the game ends; if not, then player 1 has choose again (on Monday evening) between t and $\bar{t}$ etc.
In the last day (Friday), player 2 has only one choice: to give the exam today (e).

So **the game ends if and only if player 2 chooses** e.

In that case, **the payoff is** $(1, 0)$ **iff the last move by player 1 was** t: Student wins, since he correctly anticipated the day of the exam.
Otherwise, the payoff is $(0, 1)$: Teacher wins.

The announced sentence is true iff there is a winning strategy for Teacher.

I represent the game tree **for 2 days only**, but use it to reason about the whole (5-day) game:


0, 1
1, 0
e
e
2
2
$\overline{t}$
t
1
0, 1
$\overline{e}$
e
2
$\overline{t}$
t
1
0, 1
1, 0
e
e
2
2
$\overline{t}$
t
1
1, 0
$\overline{e}$
e
2
$\overline{t}$
t
1

Student's Winning Strategy

It is easy to see that, not only **the announced sentence is false**, but Student can get to **know that it is false**:
this is captured by the fact that **there is a winning strategy** *tttt* **for Student** (player 1).

This is in fact **the only winning strategy**, and encodes exactly the beliefs and belief-revision process captured by our solution:
$Mon > Tue > Wed > Thu > Fri$.

CONCLUSION: What Does All This Mean?

Recall that we started by assuming that the Teacher's announcement induces a "belief upgrade". This means that we assumed that the student *starts by TRUSTING the Teacher*, at least minimally: he is willing to revise with the information she provides, as long as this doesn't contradict his "hard" knowledge.

Assuming this, we showed that the belief-revision induced by Teacher's future-oriented announcement CANNOT be *an "update", but only an "upgrade" with "soft" information; i.e. the teacher may be a trusted, but NOT an infallible, source of information.*

MOREOVER, we also showed that this revision CANNOT be a lexicographic or conservative upgrade either; i.e. **the teacher cannot be strongly trusted, or even conservatively trusted**.

What this upgrade turned out to be was **one that made the content of the announcement to come to be KNOWN to be false**, although the announcement itself (as a belief-revising action) **WAS still "efficacious"** (and thus *it COULD NOT be dismissed*).

This reinforces the lesson that DEL has learnt from the older Dynamic Semantics (for natural language):
the meaning of an announcement cannot be reduced to its truth conditions;
in addition to being true or false, **an announcement "does" something**: *it changes the hearer's doxastic state in a certain way.*

Even if known to be false, the announcement can still change one's beliefs. The "real" meaning of an announcement is given by this doxastic change.

All the three proposed approaches come essentially to the same conclusion:

IF the student starts by trusting the Teacher, *then after the announcement* **he should come to regard as more plausible (or more probable) that the exam will take place in any particular day than that it will take place in the next day**.

As a result, every evening the student will *believe the exam is tomorrow*.

So, whenever the exam comes, it **will NOT** be a surprise.

But... the ONLY way for him to "prove" the teacher wrong is to be prepared for the exam at any given moment!

There is no going back: the doxastic move that made the student be prepared every day is precisely the one that falsified Teacher's statement.

But...I find it very pleasant (and rather ironical) that this solution agrees with (what should be every) Teacher's *true intentions*: **what more can she expect to achieve** with such a future-oriented "surprise" announcement, **but to make the student be always prepared for the exam**?!