Surprise! Or How to Avoid Unexpected Exams

Alexandru Baltag
(ILLC, University of Amsterdam)

This is a light introduction to formal tools for use in modern Epistemology.

1.1. Epistemic Puzzle no. 1: Moore Sentences

Our starting example concerns a "*love triangle*": suppose that Alice and Bob are a couple, but Alice has just started an affair with Charles.

At some point, Alice sends to Charles an email, saying:

"**Don't worry, Bob doesn't know about us**".

But suppose now that Bob accidentally reads the message (by, say, secretely breaking into Alice's email account).

Then, paradoxically enough, after seeing (and believing) the message which says he doesn't know..., *he will know*!

So, in this case, **learning the message is a way to falsify it**.

As we'll see, this example shows that **standard belief-revision postulates may fail to hold in such complex learning actions**, in which the message to be learned refers to the knowledge of the hearer.

Epistemic Puzzle no. 2: Self-fulfilling falsehoods

Suppose Alice becomes somehow *convinced that Bob knows everything* (about the affair).

This is false (Bob doesn't have a clue), but nevertheless she's so convinced that she makes an attempt to warn Charles by sending him a message:

"**Bob knows everything about the affair!**".

As before, Bob secretely reads (and believes) the message. While false at the moment of its sending, the message becomes true: *now he knows.*

So, *communicating* a false belief (i.e. Alice's action) might be a self-fulfilling prophecy: **Alice's false belief, once communicated, becomes true**.

In the same time, the action of (reading and) *believing* a falsehood (i.e. Bob's action) can be self-fulfilling: **the false message, once believed, becomes true**.

Puzzle no. 3: The Surprise Examination Paradox

The Student knows for sure that the date of the exam has been fixed in one of the five (working) days of next week. But he doesn't know in which day.

But 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.

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.

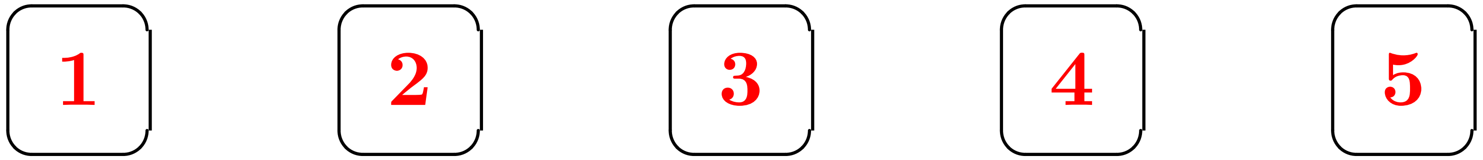
So, using this argument, the students come to “know” that the announcement is false: the exam CANNOT be a surprise.

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 Knowledge

We are given a set of “possible worlds”, meant to represent **all the relevant possibilities** in a certain situation.

EXAMPLE 1: A possible model for the initial knowledge of the Student in the Surprise Exam story:



1 **2** **3** **4** **5**

Here, we denoted by i (with $1 \leq i \leq 5$) the “possible world” in which **the exam will happen in day** i of the week.

Knowledge

The universal quantifier over the domain of possibilities is interpreted as **knowledge** by the (implicit) agent.

So we say the agent **knows** a sentence φ if φ is **true in all the possible worlds** of the model.

In the previous example, the Student **doesn't know the day of the exam, but he knows that it will happen in one of the five working days of the week**:

$$\left(\bigwedge_{1 \leq i \leq 5} \neg K i \right) \wedge K \left(\bigvee_{1 \leq i \leq 5} i \right)$$

Modelling Beliefs

What about beliefs? Unlike knowledge, **beliefs can be wrong**.

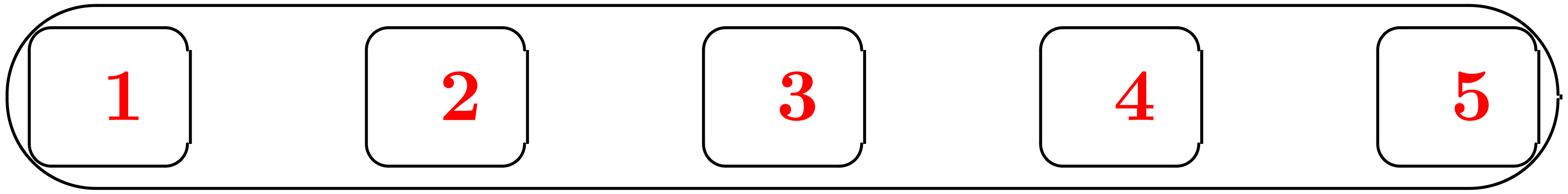
This is because belief *goes beyond what is known*.

To point out which worlds are **believed to be possible** by the agent we **encircle them**: these worlds form the "**sphere of beliefs**".

"*Belief*" now **quantifies (universally) ONLY over the worlds in this sphere**, while "*knowledge*" still quantifies over **ALL** possible worlds.

Example 2

In the Surprise Exam story, a possible initial situation (BEFORE the Teacher's announcement) might be given by:



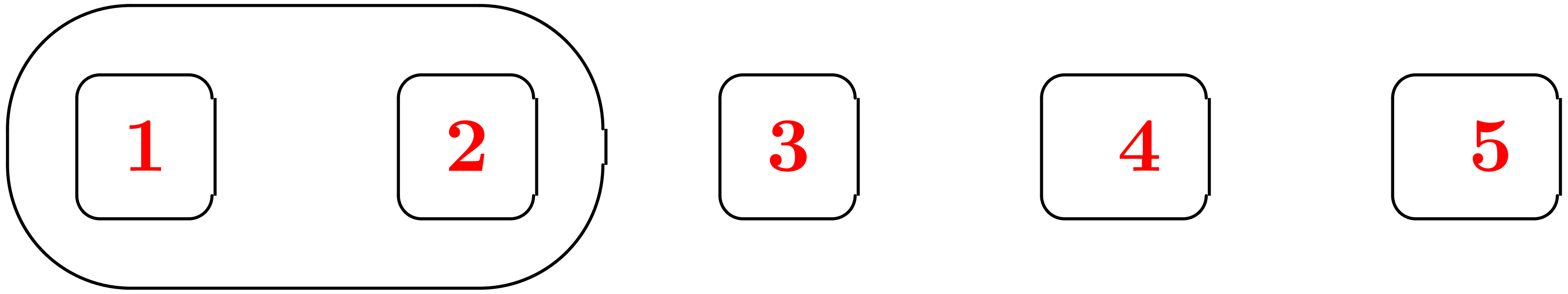
where i means that: the exam takes places in the i-th (working) day of the week.

This encodes an initial situation in which the student **knows that there will be an exam** in (exactly) one of the days, but he **doesn't know the day**, and moreover he **doesn't have any special belief about this**: he considers all days as being *possible*.

Beliefs

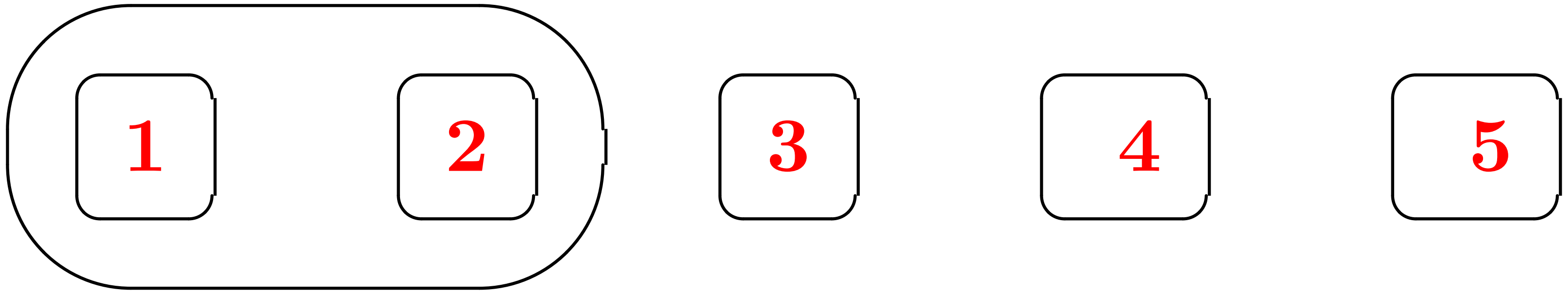
EXAMPLE 2':

If however, the Student **believes** (for some reason or another) that the exam will take place either Monday or Tuesday, then the correct representation is:

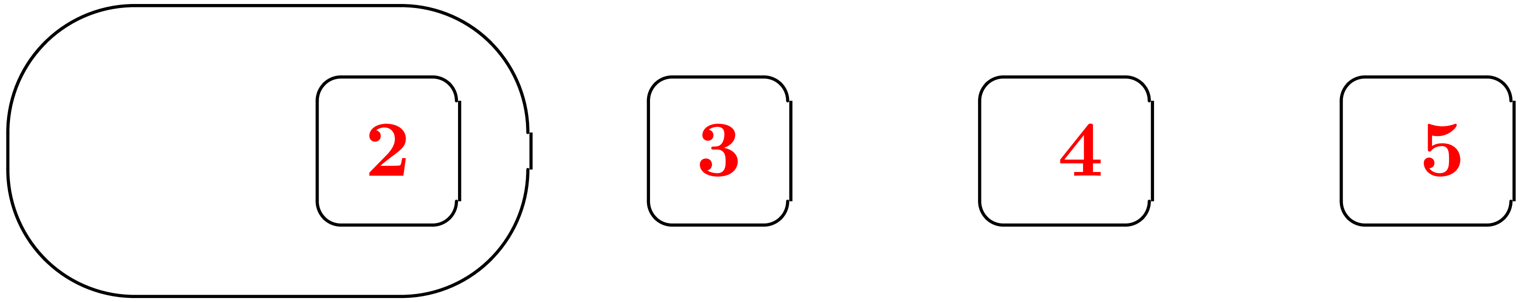


Learning New Information

Suppose we start in the situation from Example 2'



Suppose *no announcement is made by the Teacher.* Still, after the **passing of Monday (with no exam!)**, the student learns that the exam was not on Monday: so he can **eliminate** world 1, as *no longer possible.* The new model is:



Update as World Elimination

This kind of learning is called an **update**.

In general, **updating corresponds to world elimination**:

an **update !φ with a sentence** φ is simply the operation of **deleting all the non-φ possibilities**

After the update, the worlds not satisfying φ are no longer possible: the actual world is known not to be among them.

Dynamic Update Modality

We can introduce a **dynamic update modality** $[!\varphi]$.

The sentence

$$[!\varphi]\psi$$

means that: **after updating with φ, sentence ψ will become true**.

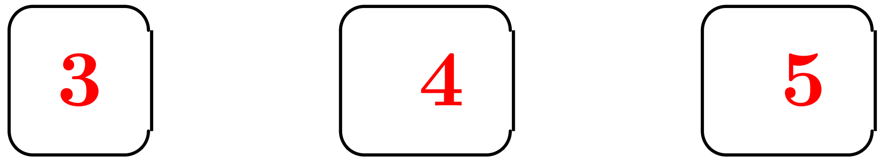
EXAMPLE: Intuitively, after the above update $!(\neg 1)$, the Students comes to **know that the exam is not on Monday**:

indeed, $K\neg 1$ is *true in the **new** model (after the update)*; hence, $[!\neg 1]K\neg 1$ *was already true* in the **original** model (**before** update).

Similarly, in Example 2', $B2$ is *true in the **new** model* (**after** the update);
hence $[!\neg 1]B2$ *was already true* in the **original** model.

Another Day Passes

After the *passing of Tuesday* (again, with no exam happening), let us try to model the result by performing an update $!(\neg 2)$ on the previous model:



3 4 5

The sphere of beliefs **disappeared**!

Or if you like, our new model has an **empty** sphere of beliefs...

The Problem of Belief Revision

If we apply our definition of *belief* to a "model" having an empty sphere of beliefs, it would follow that the agent has come to **believe everything**:

(s)he has *inconsistent beliefs*!

Indeed, the following holds in our original model (from Example 2'):

$$[!\neg 1][!\neg 2]B(2 \wedge \neg 2)$$

Our student has simply gone crazy!

The Problem of Belief Revision: syntactic version

What happens if I learn a new fact φ (e.g. that the exam is neither on Monday nor on Tuesday) **that goes in contradiction to my old beliefs?**

If I accept the fact φ, and put it together with the set T of all the sentences I used to believe, the resulting set $T \cup \{\varphi\}$ is **logically inconsistent**.

So I have to give up some of my old beliefs. But which of them?

Maybe all of them?! No, I should maybe **try to maintain as much as possible of my old beliefs**, while still **accepting the new fact** φ (without arriving to a contradiction).

Solving the Belief Revision Problem

Intuitively, the Student in the Surprise Exam Example and the clean child in the Cheating Muddy Children story should **avoid getting crazy** when faced with information contradicting their prior beliefs, by **falling back onto some weaker, second-level beliefs**.

SOLUTION: **Add more spheres**!

Such sphere would represent "weaker" beliefs, that give the agent a "**contingency plan**":

when the stronger beliefs are contradicted, go with the weaker ones!

The second sphere of beliefs

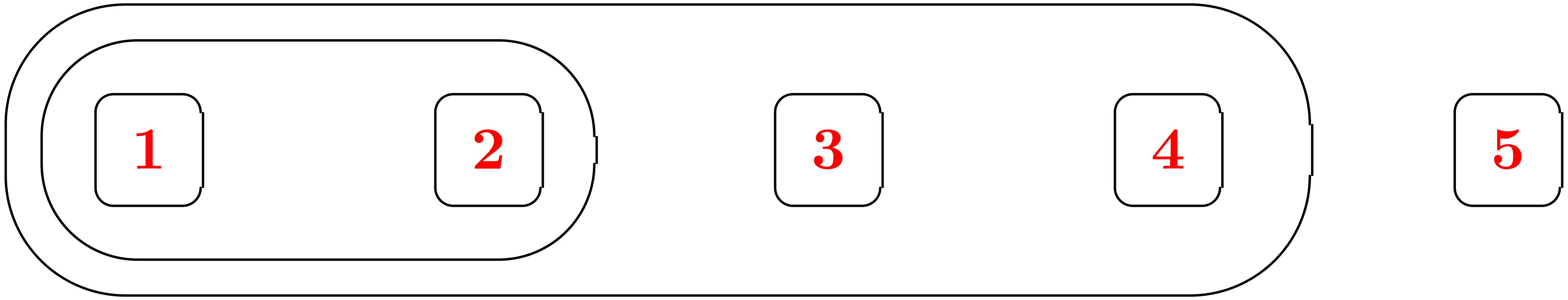
So we should have a **second sphere of beliefs** S_1, consisting of worlds that **are "believed" to be possible, in a weaker sense that the ones in the first sphere** S_0.

Intuitively, *states are believed to be possible in a strong sense should be believed to be possible in a weaker sense* as well:

$$S_0 \subseteq S_1$$

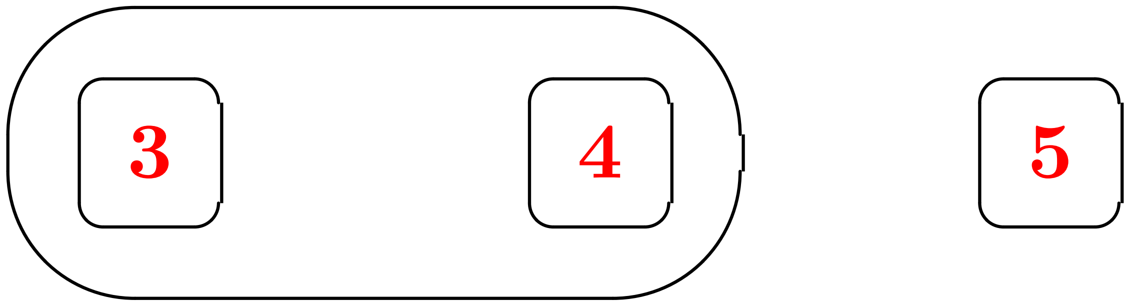
Example

As before, suppose the Student "believes", in the *strongest* sense, that the exam is on **either Monday or Tuesday**; but, in a somewhat *weaker sense*, he believes the exam is **in any day, except for Friday**.



Updating with Information Contradicting Prior Beliefs

We can obtain the *Student's beliefs after the passing of the first two days*, by performing successively an update $!(\neg 1)$ followed by an update $!(\neg 2)$, obtaining

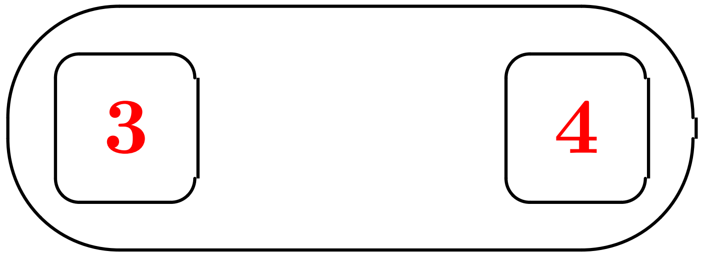


$$[!\neg 1][!\neg 2]B(3 \vee 4)$$

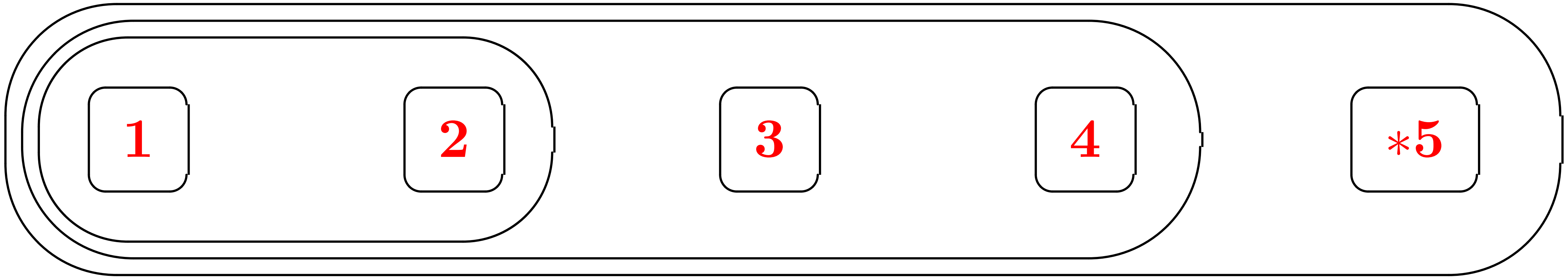
Far from getting crazy, the student has now simply *changed* his beliefs to some *other, still consistent beliefs*: he believes now the exam will be on either Wednesday or Thursday (but NOT on Friday).

Third Sphere of Beliefs

What happens if some of the new beliefs are still false? Say, if the exam is actually on Friday:



Then, after two more day pass, we get the same problem again! To solve it, Student needs **a third sphere**, composed of states believed to be *possible in a still weaker sense*:



From spheres to plausibility order

So we obtained a nested system of spheres

$$S_0 \subset S_1 \subset S_2 \subset \ldots S_n = S$$

ending with the set S of all possible worlds. This is called a **sphere model**.

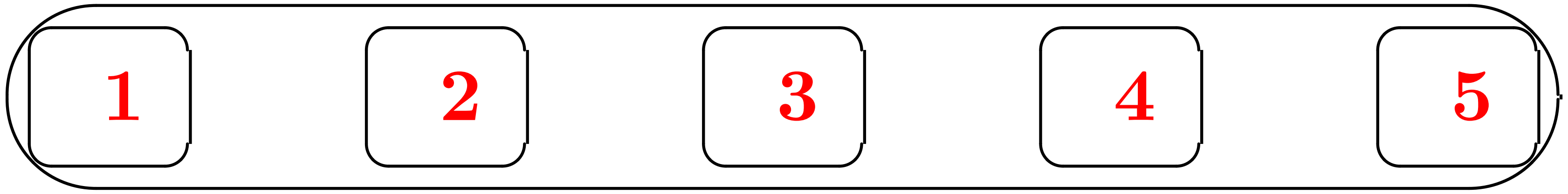
Equivalently, we can encode the same information by introducing a **plausibility (pre)ordering** $\leq$ on worlds:

We write $s \leq t$, and say that **world** t **is at least as plausible as world** s, if *every sphere containing* s *contains* t *as well.*

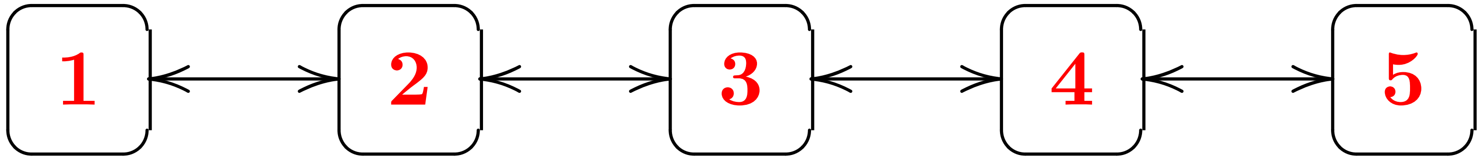
$\mathbf{S} = (\mathbf{S}, \leq)$ is called a **plausibility model**.

Example of plausibility models

The sphere model in Example 2 (in which the student has no beliefs one way or another)

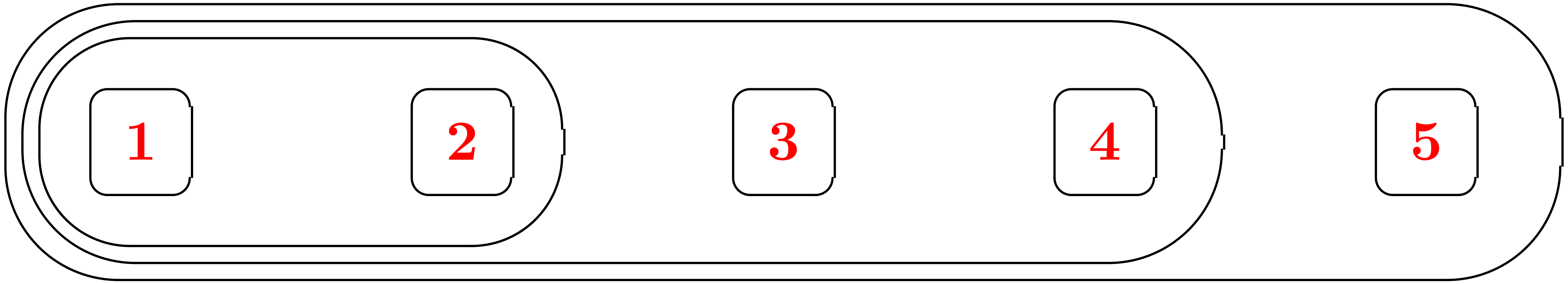


corresponds to considering all days to be **equally plausible**:

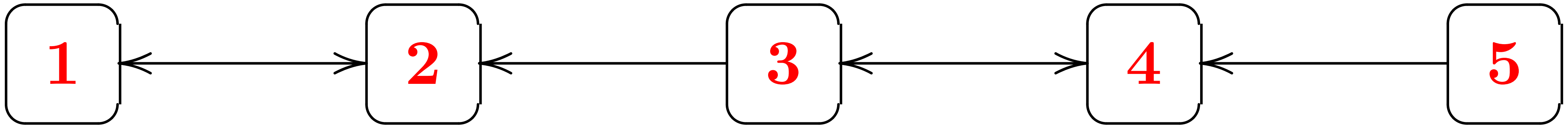


Another Example

The sphere model in Example 2'



can be equivalently represented in terms of plausibility:



The first two worlds (in which the exam is in the first two days) are the most plausible worlds (and equally plausible to each other). The next most plausible are worlds 3 and 4 (corresponding to the exam being on Wednesday or Thursday). Finally, the least plausible is 5 (Friday).

Belief

Our above definition of “BELIEF” (**quantifying universally over the first sphere**) is equivalent to saying that:

φ **is BELIEVED** if φ **is true in (ALL) THE MOST PLAUSIBLE worlds**.

Examples

In the above model, the Student *believes that the exam is either on Monday or on Tuesday*:

$$B(1 \vee 2)$$

But *IF he learns that this is NOT the case, then after this he believes the exam is either on Wednesday or Thursday*:

$$[!\neg(1 \vee 2)]B(3 \vee 4)$$

Belief Revision Theory and the Success Postulate

Alchourrón, Gärdenfors and Makinson made a **list of all the reasonable, intuitively desirable "rationality conditions"**, that govern the way in which any "rational" agent should revise his/her beliefs.

This list comprises the "**AGM Postulates**".

A typical postulate is the following

SUCCESS AXIOM:

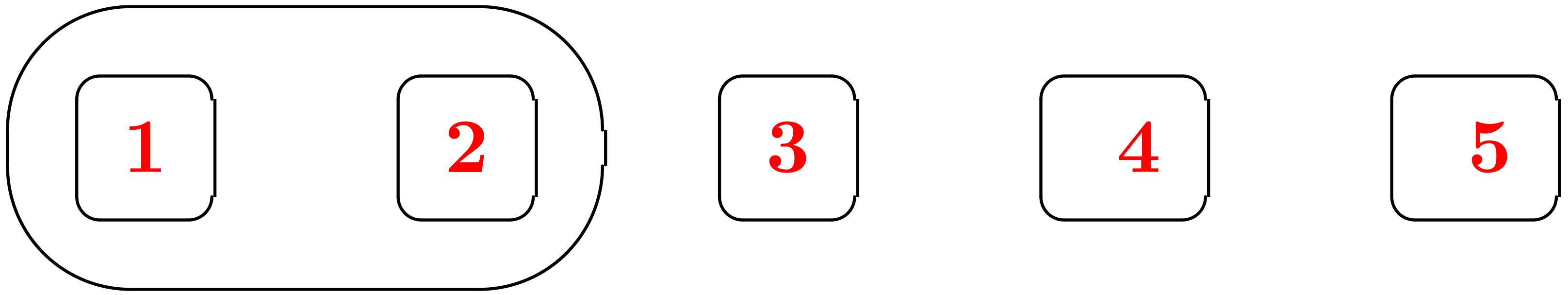
$$[!\varphi]B\varphi$$

After learning φ, the agent believes φ.

What could be more obvious?!

Counterexample

Suppose that in the initial model of Example 2'



the exam is actually planned to take place on Friday. The Teacher announces (truthfully): "**You (Student) believe that the exam is on either Monday or Tuesday, but you are wrong!**"

Moore Sentences

The announced sentence φ is

$$\varphi \ := \ B(1 \vee 2) \wedge \neg(1 \vee 2).$$

This sentence is true, and in fact let's assume for the moment that the Student has absolute trust in Teacher: he knows that the Teacher never lies.

What should the Student believe **after that**?

According to the Success Postulate, he should believe φ.

I.e. **he should believe BOTH that he believes the exam to be on either on Monday or on Tuesday AND that the exam is NOT in any of these days**.

Paradox?!

Moore sentences are un-believable!

The above φ is an example of a *Moore sentence*:

a sentence which can be true, but which cannot be believed (by the Student).

Typically, such sentences are of the form

$\varphi ::= \; Bp \wedge \neg p.$

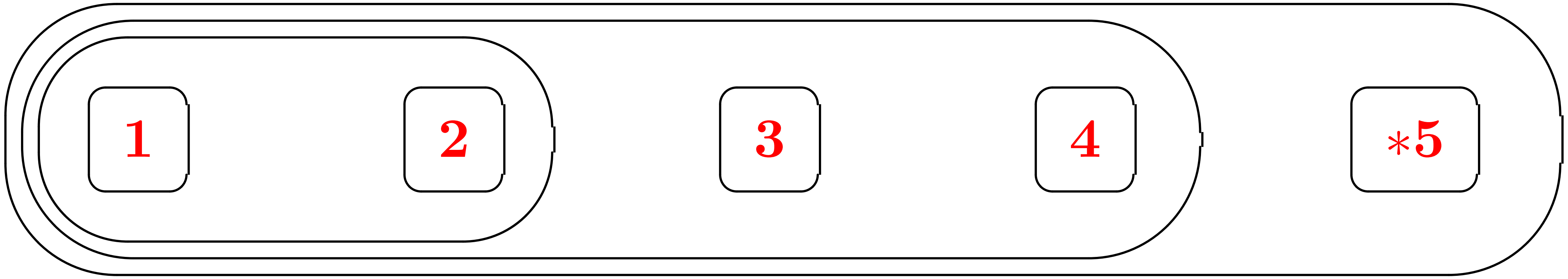
This **can never be believed**: the sentence $B\varphi$, or to spell it out in full

$$B(Bp \wedge \neg p),$$

is **inconsistent**, according to our modelling.

Learning a Moore sentence

Suppose that in the initial situation



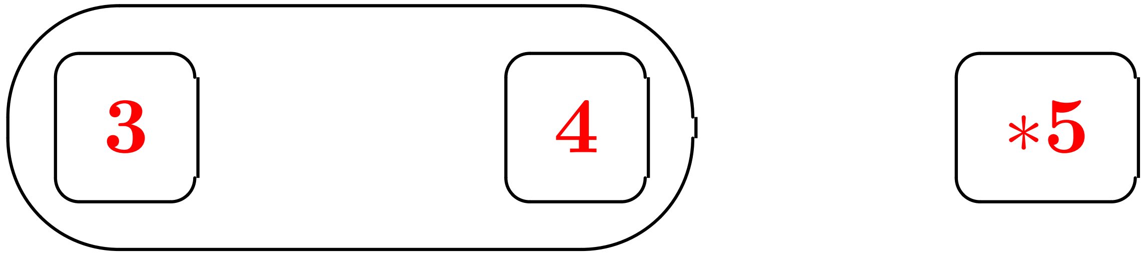
the Teacher (absolutely trusted to always tell the truth) announces the sentence φ:

$$\varphi \ ::= \ B(1 \vee 2) \wedge \neg(1 \vee 2).$$

What will the Student believe next?

Updating with the Moore sentence

Well, if we apply an update with φ, we obtain:



Note that, in this model, the Student does NOT believe φ (as the Success axiom would require)! Instead, he simply believes $3 \vee 4$.

Learning is a way to falsify

Moreover, note that the sentence φ is **false** in this last model.

In other words, φ **changed its truth value**:

it used to be true before being learnt; but it became false after being learnt!

Moore sentence are **self-falsifying**:
they necessarily become false by being learnt.

Moreover, after the Student learns it, he now **knows it to be false**:

so, far from coming to believe φ (as the Success axiom would require), the Students **comes to dis-believe** φ (in fact, to know that it's false) **after learning it**!

Conclusion: Success Postulate FAILS

The conclusion is that the “Success Axiom”

$$[!\varphi]B\varphi$$

fails for Moore sentences.

So this “axiom” is NOT generally true!

GENERALIZATION: Other forms of learning

But what if the student doesn't trust the Teacher to infallibly tell the truth?

In general, the way an agent (e.g. the Student) changes his beliefs after learning some information φ depends on his **attitude (strong trust, distrust etc) towards the source of this new information** (e.g. the Teacher).

To capture this, we need to move from updates to a more general notion: belief **upgrades.**

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 $\leq'$.

Dynamic Upgrade Modalities. As for update, we can introduce dynamic modalities for each type of upgrade $T\varphi$:

$$[T\varphi]\psi$$

means that, *after T-upgrading with φ, sentence ψ will become true.*

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

(0) Identity id:

everything is left the same (same states, same plausibility order.

(1) Update $!\varphi$:

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

(1') Negative Update $!^{-}\varphi$:

all the 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.*

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

- **Identity**: a **neutral** attitude ("indifference"). The source is neither believed nor dis-believed, but simply ignored. The agent keeps all his old beliefs.
- **Update**: an **infallible** source. The source is *"known" (guaranteed) to be always truthful.*
- **Negative update**: an **infallible source of falsehoods**. The source is *"known" (guaranteed) to be always lying.*
- **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!
- **Radical Negative upgrade**: the source is *strongly believed to be lying.*

Explanation continued

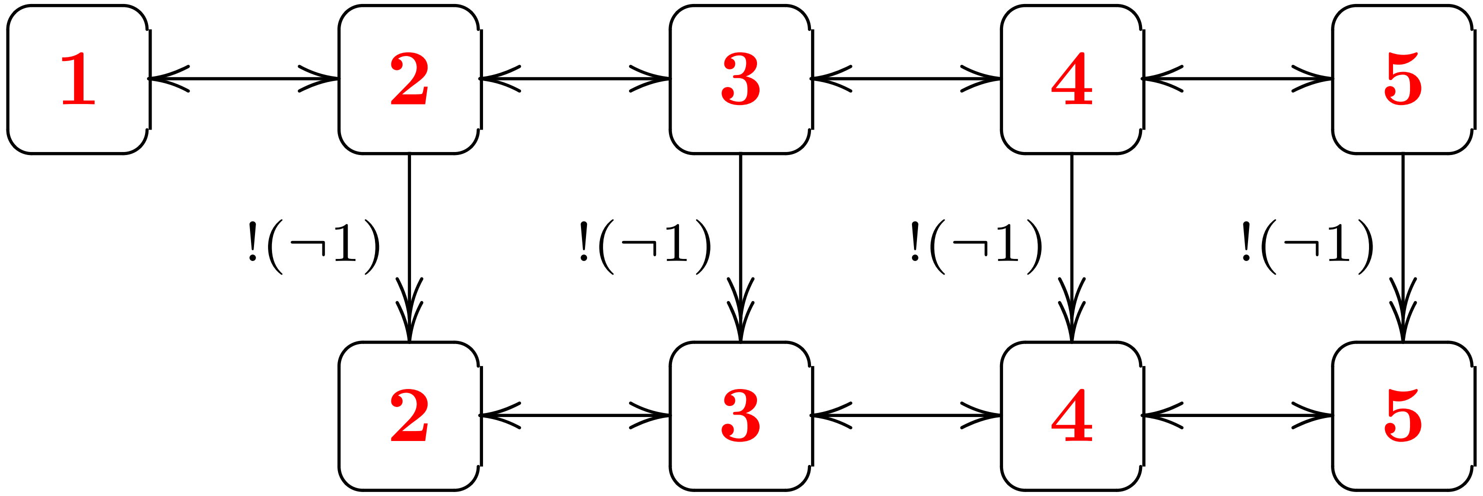
After an **update**, the agent comes to “**know**” that φ (was the case): *all non-φ possibilities are forever eliminated*

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.

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

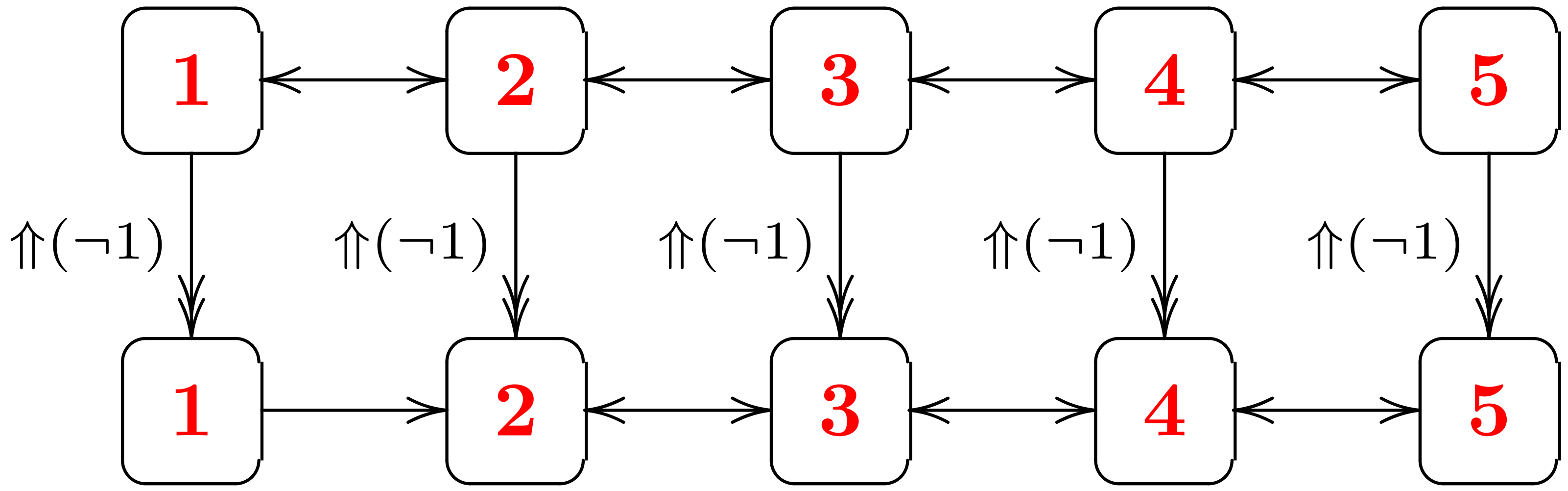
Example 3

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 $\overset{!(\neg 1)}{\Longrightarrow}$ 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:



Solutions to Surprise Exam: (1) Gerbrandy's Solution

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

Gerbrandy interprets the announcement itself as an **update** with the above sentence *surprise*.

So he assumes that Student considers Teacher to be *infallible source of truth*:
in this interpretation, Student *knows* beyond any doubt that Teacher spoke the truth.

Gerbrandy argues that the 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:

$$surprise \;=\; \bigwedge_{1\leq i\leq 5} \left(i \Rightarrow [!\neg 1]\ldots[!\neg(i-1)]\,\neg Bi \right).$$

In English, this reading of “surprise” is given by:
“**(for any i) 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).

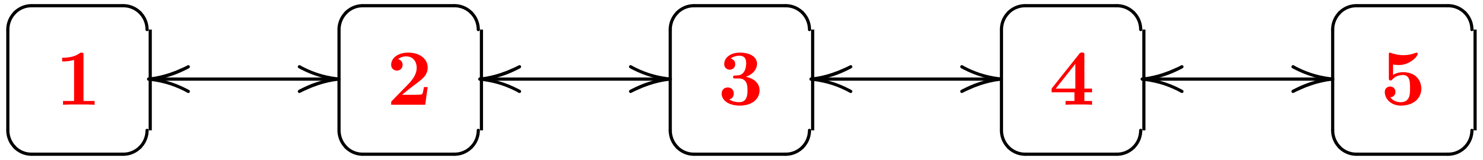
Indeed, the following is a logical validity

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

PROOF: by backward induction (starting with Friday).

NO SURPRISE ON FRIDAY

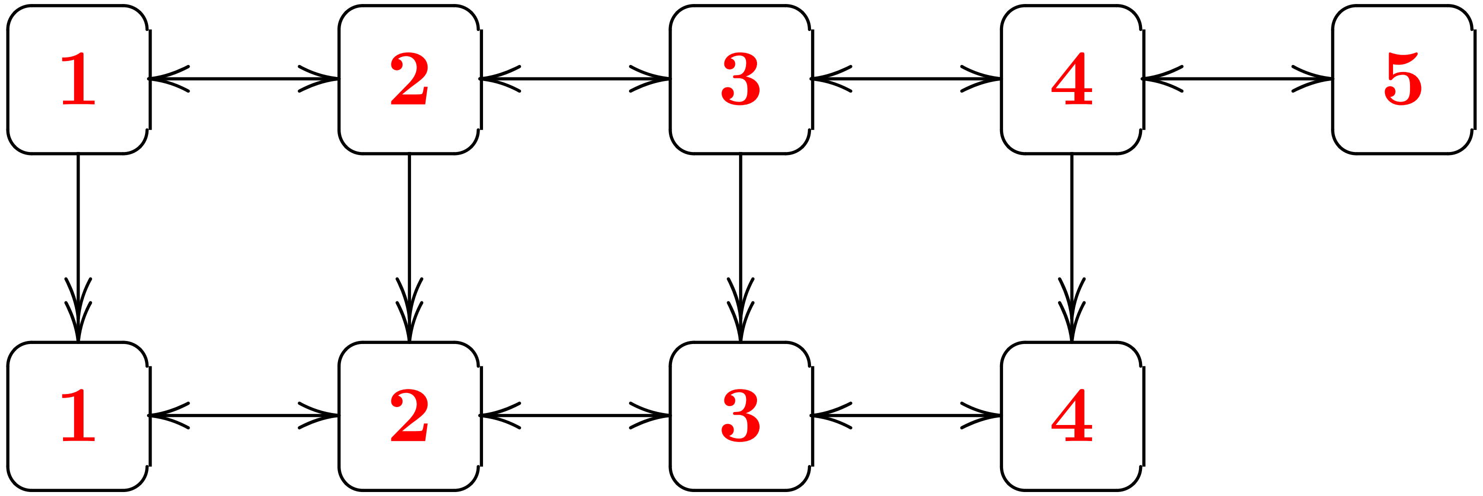
It is easy to see that if initially the Students has complete uncertainty, i.e. we start with the model



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** $!(\mathit{surprise})$ with the sentence $\mathit{surprise}$. 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: Teacher is NOT infallible!

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 (about the Student's future beliefs after hearing her 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?

(2) Quine's 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 **refuse** to revise even if this is consistent with his knowledge.

The Student has no reason to believe or disbelieve the Teacher; so he should just **dismiss** the *surprise* announcement out of hand, sticking with whatever he believed before, no matter what.

Assuming he started by considering all days as equally plausible, he should continue to do so after Teacher's announcement. Thus, 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 indifference to Teacher's announcement 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.

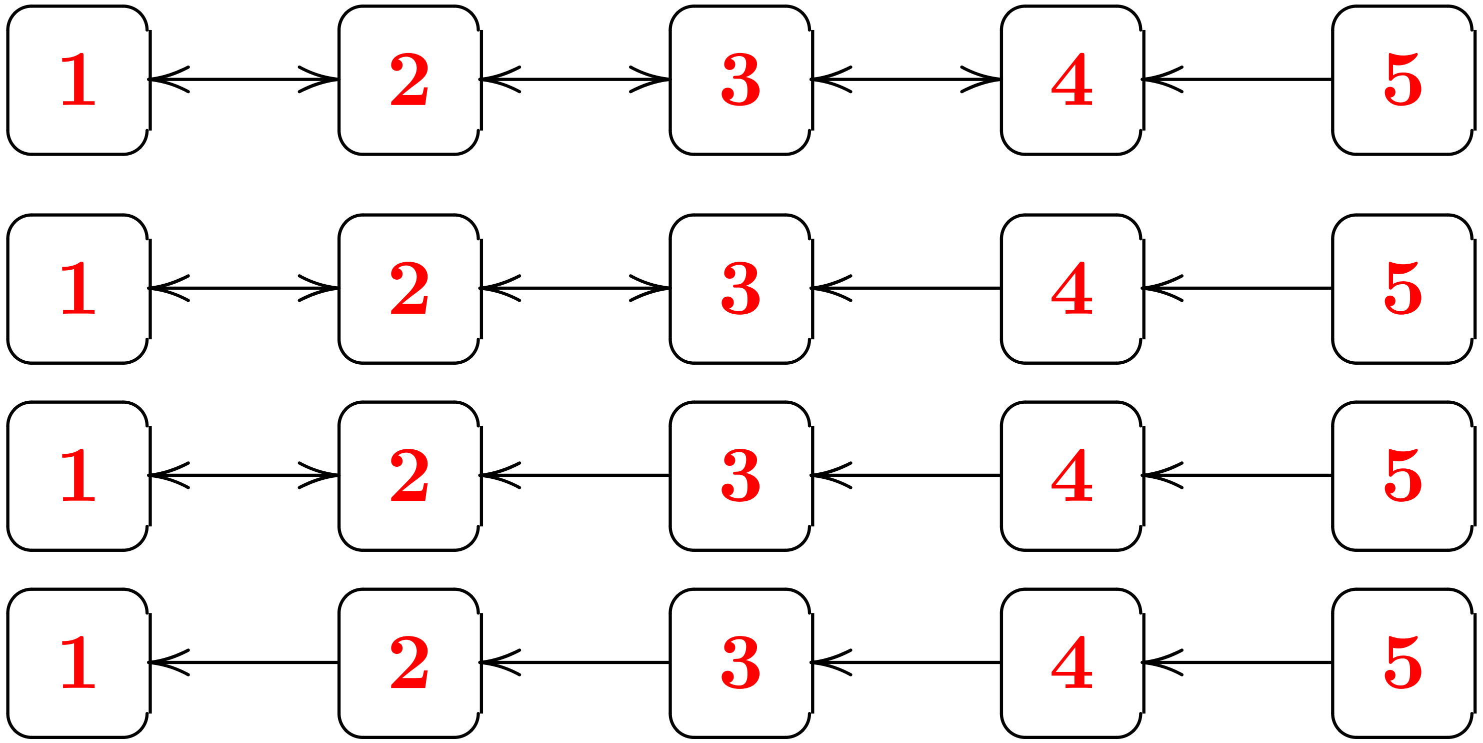
As before, since the announcement was *self-referential* ("The exam will be a surprise EVEN after I am telling you all this"), the Student will have to **iterate** this upgrade.

(3) Our Solution

We (joint work with Sonja Smets) propose to 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:



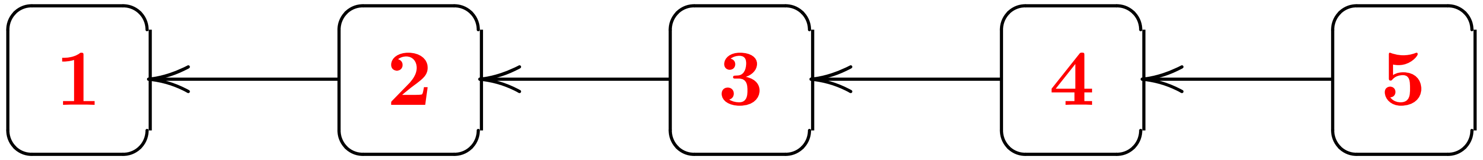
After this, any further iteration leaves the model unchanged!

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 ⇑ ($surprise$) or conservative upgrades ↑ ($surprise$), 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**!

Why not dismiss Teacher's announcement?!

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!

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 IF he does this, THEN 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.

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"**:
the teacher MIGHT be trusted, but NOT as an INFALLIBLE source of truthful information.

MOREOVER, **the only way for the Student to adopt a "positive" attitude (of relative trust) towards the Teacher**'s self-referential announcement is to **perform an iterated upgrade, whose conclusion is that the Teacher LIED**!

However, concluding that the announcement was false is NOT a good reason to dismiss the announcement.

The truth value of Teacher's announcement depends on what the Student will do.

And the only way for the Student to prove Teacher wrong is to do exactly what we assumed he does: keep expecting the exam tomorrow.

LESSON

The lesson is that **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 change of beliefs.

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**?!