Dynamic-doxastic norms versus doxastic-norm norm dynamics

Alexandru Baltag (Oxford)

Using some joint work with Sonja Smets (Groningen), as well as other work by Johan van Benthem (Amsterdam)
Revising one’s beliefs in the face of new information is tricky business.

When the new information contradicts previous beliefs, one cannot just add it to the belief set, since this would lead to **inconsistency**.
There is an implicit **doxastic norm** acting here, namely
the one encoded in the standard axiom $(D)$ of doxastic
logic: *never hold inconsistent beliefs*

\[ \neg B(\varphi \land \neg \varphi) \]

One must give up some beliefs, **but which of them?**
There is no general a priori logical principle to help us
choose.
Example

Suppose I believe two facts $p$ and $q$ and (by logical closure) their conjunction $p \land q$. So my belief base is the following

$$\{p, q, p \land q\}.$$ 

Suppose now that I learn the last sentence was actually false.

Obviously, I have to revise my belief base, eliminating the sentence $p \land q$, and replacing it with its negation: $\neg(p \land q)$. 
But the base

\{p, q, ¬(p \land q)\}

is inconsistent!

So I have to do more!

Obviously, to accommodate the new fact \(¬(p \land q)\), I have to give up either my belief in \(p\) or my belief in \(q\).

But which one?
Belief Revision Theory

Standard Belief Revision Theory, also called AGM theory (from authors Alchourrón, Gärdenfors and Makinson) postulates as given:

- *theories* ("belief sets" or "belief bases") $T$
- *new information* (a formula) $\varphi$
- *a revision operator* $T \ast \varphi$
**Interpretation**

$T * \varphi$ is supposed to represent the *new belief base* ("new theory") theory after learning $\varphi$: the agent’s new set of beliefs, given that the initial set of beliefs was $T$ and that the agent has learned $\varphi$ (and only $\varphi$).
AGM Axioms are “Dynamic-Doxastic Norms”

AGM authors impose a number of axioms on the operation \( * \), which may be called “rationality conditions”, since they are meant to govern the way a rational agent should revise his/her beliefs.

In effect, the AGM axioms can be seen as norms for rational belief revision: what I call “dynamic-doxastic norms”.
Example: the Success axiom

The “Success Axiom” is an example of such an axiom, requiring that the agent should revise in such a way that the new information $\varphi$ is believed after revision:

$$\varphi \in T \ast \varphi$$

Or, in dynamic logic notation:

$$[\ast \varphi]B\varphi$$
Infallible Sources

This seems reasonable, when the source of the new information is **infallible**: if the Pope announces \( \varphi \), then I must come to accept \( \varphi \).

An infallible source comes with a “warranty of truthfulness”: it is common knowledge that any information from this source must be true.
Sometimes, “Success” fails!

However, this norm may bring us into conflict with the consistency norm (axiom D), when it is applied to sentences involving higher-order beliefs:

Suppose I currently believe the sentence $p$ given by

\[ p = "My girlfriend loves me" \]

Then the Pope announces:

“You fool, you wrongly believe your girlfriend loves you!”
The sentence $\varphi$ he uttered is a “Moore sentence”:

$$\varphi := Bp \land \neg p.$$ 

After $\varphi$ is learned, $\varphi$ obviously becomes false!

Indeed, as a side effect, I learn $\neg p$, so I no longer believe $p$. Hence, $\varphi$ becomes false (since it asserts that I believe $p$). More importantly, I now know that $\varphi$ is false (since, by introspection, I know that I don’t believe $p$ anymore).
The Price of “Success”

But the AGM norm of “Success” asks me to believe (after learning \(\varphi\)) that \(\varphi\) is true!

I am now required to somehow believe both that my girlfriend doesn’t love me and that I believe she loves me!

In other words, “Success” requires me (as a principle of rationality!) to acquire false beliefs, in contradiction with \((D)\)!
Who is infallible?!

Besides all this, most new information that we receive comes from sources that are less than infallible: they may be highly trusted, very reliable etc., but they do not come with any absolute warranty of truthfulness.
Revising the believe-revision norms

In other words, default dynamic-doxastic norms such as “Success” may have to be revised themselves: when the source turns out to be fallible, or when the new information involves higher-order beliefs, the “Success” axiom is no longer normative, at least not in its original form.
EXAMPLE: A Weakening of Success

For sentences $\varphi$ involving higher-level beliefs, it is easy to see what is the correctly revised “Success” norm:

$$[*\varphi]B(BEFORE\varphi)$$

where $BEFORE$ is a past-tense operator: after revising with the information $\varphi$ (received from an infallible source), one should believe that $\varphi$ was true before the revision.
The “Infallibility” Norm

AGM Success Axiom, in its original form, seems to accept the “infallibility” of the source of new information;

\[ (*\varphi)B\varphi \]

But, in fact, the infallibility norm is stronger than the Success requirement, since it asks that the new information \( \varphi \), once acquired, will never be revised: it is irrational to revise something that comes with a warranty of truthfulness. In other words: the information received from an infallible source will come to be (not only be believed, but also) known to be true.
**Strong versus Weak Infallibility**

This is the **Strong Infallibility** Principle:

\[ [\ast \varphi]K\varphi \]

In contrast, the AGM Success postulate

\[ [\ast \varphi]B\varphi \]

will be called the **Weak Infallibility** Principle.
Revising the Infallibility Norm

A further weakening of this norm would require that the new information is accepted UNLESS it was already known to be false:

\[ K \neg \varphi \lor [\ast \varphi]B\varphi \]

Combining this with the previous modification, we obtain a version suited for higher-level beliefs:

\[ K \neg \varphi \lor [\ast \varphi]B(BEFORE)\varphi \]

Let us call this “Fallible AGM Norm”. All the belief-revision procedures in the literature satisfy this norm.
Even weaker norms?

It may seem that the last requirement is the weakest possible norm for any “AGM-like” revision: it expresses a “willingness” to revise with $\varphi$ in all cases in which this is consistent with the agent’s prior knowledge.

However, it is in principle possible that the agent, while not having any prior knowledge against $\varphi$, may still acquire knowledge contradicting $\varphi$ during (and due to) the very attempt of revising with $\varphi$.

In that case, the agent will not be able to accept $\varphi$ after revision, as long as he obeys the Consistency norm ($D$).
Minimal AGM: Willingness to Revise

This leads us to an even weaker requirement:

\[[\star \varphi]K \neg \varphi \lor [\star \varphi]B \varphi\]

The version for sentences involving higher-order beliefs is:

\[[\star \varphi]K(\text{BEFORE} \neg \varphi) \lor [\star \varphi]B(\text{BEFORE} \varphi)\]

We call this the “Minimal AGM Norm”. It requires that the agents to be willing to come to believe \(\varphi\), unless he comes to know \(\neg \varphi\).
Different Doxastic Attitudes ⇐⇒ Different Norms

This shows that different belief-revision norms correspond to **different doxastic attitudes** towards the source of information: *higher or lower degrees of trust in the veracity of the source.*

All the previously encountered dynamic-doxastic norms assume at least a minimal degree of trust in the source, or a minimal willingness to revise.
But, of course, these are not the only possible attitudes: the agent might *completely distrust* the source (hence *never revise*).

Or the agent may *condition his acceptance* of new information on some reciprocal action: e.g. the pagan kings were typically willing to convert to Christianity *PROVIDED* that the missionary could perform some miracle.
Different scientific paradigms are embodied in different dynamic-doxastic norms.
EXAMPLES

- **Infallible AGM** (e.g. the Catholic doctrine concerning the Pope): \([*\varphi]B\varphi\)
  
  "The source of new information is infallible. Always revise with it."

- **Fallible AGM**: \(K\neg\varphi \lor [*\varphi]B\varphi\)
  
  "Believe the New Information Unless it is Known to be False."

- **Fundamentalism**: \([*\varphi]B\theta \iff B\theta\)
  
  "Keep your old beliefs no matter what."
Vaguer Normative Principles

There exist also more general, less precise normative principles, such as:

**AGM**: “Minimize the belief change as much as it is possible, given Success and Consistency (D).”

**Learning Theory**: “Apply whatever belief-revision procedure is more likely to lead faster to the ’Truth’.”

As we’ll see (given results by Kelly, and by Baltag and Smets), there is a tension between these two general norms!
SEMANTICS for Revisable Beliefs

A (single-agent, finite) plausibility frame is a finite set $S$ of “states” (or “possible worlds”) together with an equivalence relation $\sim$, called epistemic indistinguishability, and a preorder $\leq \subseteq S \times S$, called plausibility relation, subject to two conditions:

$$s \leq t \implies s \sim t,$$

$$s \sim t \implies s \leq t \lor t \leq s.$$

Read $s \sim t$ as “$s$ and $t$ are indistinguishable” (or “epistemic alternatives to each other”) and $s \leq t$ as “state $s$ is at least as plausible as state $t$”.
A *(single-agent, finite) plausibility model* is just a Kripke model \((S, \sim, \leq, \| \cdot \|)\) having a (finite) plausibility structure as its underlying frame.

I.e. a plausibility frame \((S, \sim, \leq)\) together with a **valuation map**, assigning to each atomic sentence \(p\) (in a given set \(At\) of atomic sentences) some set \(\|p\| \subseteq S\).
A sentence $\varphi$ is ("irrevocably") known at a state $s$ of a plausibility model $S$ if $\varphi$ is true in all the worlds that are epistemically indistinguishable from $s$:

$$s \models K\varphi \iff t \models \varphi \text{ for all } t \sim s.$$ 

In other words, $\varphi$ is known at $s$ iff $s$’s information cell $[s] := \{t \in S : t \sim s\}$ is included in the set $\|\varphi\|_S$ of worlds satisfying $\varphi$:

$$s \in \|K\varphi\|_S \iff [s] \subseteq \|\varphi\|_S.$$
(Conditional) Belief in Plausibility Models

A sentence $\varphi$ is **believed at state** $s$ a plausibility model $(S, \leq)$ if $\varphi$ is true in all the “most plausible” worlds that are indistinguishable from $s$; i.e. in all “maximal” states in the set

$$Max_{\leq}S := \{t \in [s] : w \leq t \text{ for all } w \in [s]\}$$

More generally, a sentence $\varphi$ is **believed conditional on** $\psi$ **at state** $s$ if $\varphi$ is true at all most plausible worlds satisfying $\psi$ that are indistinguishable from $s$; i.e. in all the states in the set

$$Max_{\leq\|\psi\|S} := \{t \in \|\psi\|S \cap [s] : w \leq t \text{ for all } w \in \|\psi\|S \cap [s]\}$$
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 $\psi$ was the case, I will believe that $\varphi$ was the case.

They can also be understood as a subjective (“doxastic”) type of non-monotonic conditionals.
Example 1: A Surprise Exam

A student \textit{knows} that an exam will take place in \textit{exactly one} of the (five) working days of next week. But he \textit{doesn’t know in which day}, and moreover he considers all days as being \textit{equally plausible} dates for the exam.

An example of a model is

\begin{center}
\begin{tikzpicture}
\node[circle,draw] (1) at (0,0) {1};
\node[circle,draw] (2) at (1,0) {2};
\node[circle,draw] (3) at (2,0) {3};
\node[circle,draw] (4) at (3,0) {4};
\node[circle,draw] (5) at (4,0) {5};
\draw[->] (1) -- (2);
\draw[->] (2) -- (3);
\draw[->] (3) -- (4);
\draw[->] (4) -- (5);
\draw[<->] (1) -- (3);
\draw[<->] (3) -- (5);
\end{tikzpicture}
\end{center}

where \textit{i} means that: the exam takes places in the \textit{i}-th (working) day of the week.
Example 2

In the model

1 ← 2 ← 3 → 4 → 5

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.
Modeling Higher-Level Belief Revision

From a *semantic* point of view, higher-level belief revision is about “revising” the whole relational structure: *changing the plausibility relation* (and/or its domain).
Upgrades with $\varphi$

A belief upgrade with (a sentence) $\varphi$ is a model transformer $\ast \varphi$, that takes any plausibility model $S = (S \sim, \leq, \| \cdot \|)$, and returns a new model $\ast \varphi(S) = (S, \sim', \leq', \| \cdot \|)$, having the same set of worlds and valuation, and s.t.:

- **knowledge increases or stays the same**: $s \sim' t \Rightarrow s \sim t$; i.e. $[s]' \subseteq [s]$
  (where $[s] = \{ t : t \sim s \}$, $[s]' = \{ t' : t' \sim' s \}$);

- the new plausibility preorder $\leq'$ satisfies the condition
  $$\| \varphi \|_S \cap [s]' \neq \emptyset \implies \text{Max}_{\leq', [s]'} \subseteq \| \varphi \|_S.$$
This last semantic condition says that: the most plausible worlds of the new information cell are among the ones that satisfied $\varphi$ in the old model (IF there are any such worlds left in the new cell!)

As we will see soon, this corresponds to our **Minimal AGM NORM** ("$\varphi$ is BELIEVED AFTER the upgrade, UNLESS it is KNOWN to be FALSE).

This condition justifies the name "upgrade WITH $\varphi$": it constrains the agent’s doxastic change, to be such that a belief in $\varphi$ is indeed acquired after the change, IF this is permitted by the agent’s posterior ‘hard’ knowledge.
The transition map associated to an upgrade

Such an upgrade $\varphi$ induces a \textbf{partial map} on the set of states $S$ of any model $S$, map also denoted by $\varphi : S \rightarrow S$, and given by

$$
\varphi(s) = s, \text{iff } s \in S',
$$

and

$$
\varphi(s) = \text{undefined}, \text{otherwise}.
$$
Hard and Soft Upgrades

An upgrade $\varphi$ is called **soft** if it doesn’t increase “hard” (irrevocable) knowledge:

$$[s]' = [s]$$

for all $s \in S$.

A soft upgrade *only conveys “soft information”*, changing only the agent’s beliefs or his belief-revision plans.

In contrast, a **hard** upgrade adds new knowledge.
Dynamic Operators

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

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

\[ S_0, S_1, \ldots, S_n, \ldots \]

of plausibility models obtained by successive upgrades \(*\varphi_0, *\varphi_1 \cdots\):

\[ S_{n+1} = *\varphi_n(S_n) . \]
Temporal Operators

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

$$NEXT \psi = [*\varphi_n] \psi.$$

Dually, one can introduce a past-tense operator

$$BEFORE \psi$$

which is true at a state $s$ in a plausibility model $S_n$ iff $s$ satisfies $\psi$ in $S_{n-1}$. 
MAGM: Willingness to Believe New Info

The second semantic condition

$$\|\varphi\|_S \cap [s]' \neq \emptyset \implies \text{Max}_{\leq} [s]' \subseteq \|\varphi\|_S$$

in our definition of belief upgrades can now be seen to “correspond” to (and so to ensure the validity of) the MAGM principle:

$$[*\varphi] \downarrow \rightarrow \text{K} \downarrow \text{BEFORE} \varphi \implies [*\varphi] \downarrow \text{B} \text{BEFORE} \varphi.$$
Examples of Upgrades $T\varphi$ with a sentence $\varphi$

(1) **Update $!\varphi$ (conditionalization with $\varphi$):**
all the non-$\varphi$ states are deleted and the same plausibility order is kept between the remaining states.

(2) **Lexicographic upgrade $\uparrow \varphi$:**
all $\varphi$-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” $\varphi$-worlds become better than all other worlds, and in rest the old order remains.
Different attitudes towards the new information

These correspond to three different possible attitudes of the learners towards the reliability of the source:

- **Update**: an infallible source. The source is “known” (guaranteed) to be truthful.

- **Lexicographic upgrade**: the source is fallible, but highly reliable, or at least very persuasive. The source is 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!
After a conservative or a lexicographic upgrade, the agent only comes to believe that $\varphi$ (was the case), unless he already knew (before the upgrade) that $\varphi$ 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” $\varphi$, so that all non-$\varphi$ possibilities are forever eliminated: we have the validity

$$[!\varphi]K(BEFORE \varphi).$$
Example 3

Suppose that, in the model in Example 1, day 1 (Monday) simply passes and no exam has yet taken place. This is an **update** !(-1), inducing a transition \(!(-1)\) to a new plausibility model with only 4 possible worlds:

![Diagram with nodes 1 to 5 and arrows connecting them, with !(-1) labels at each transition]
Example 4

In contrast, suppose that in the model in Example 1, the teacher announces (on Sunday evening) that the exam will not be on Monday. Assume the teacher is a trusted, but not infallible, source. Then this is a lexicographic upgrade $\uparrow (\neg 1)$, inducing a transition to a model with the same 5 worlds:
Formalizing Doxastic Attitudes

Let us now add to our language two ingredients:

- *dynamic modalities* $[\star \varphi]$ corresponding to announcements from a fixed (unnamed) source;

- *atomic sentences* $!$, $\uparrow$, $\mathcal{A}$, $FAGM$, $MAGM$ etc, encoding the agent’s attitude towards the (unnamed) source of the announcements

We use $\tau$ as generic variable for doxastic attitudes.
Examples

! says that the agent knows the source to be infallible and hence will perform hard upgrades with whatever it is announced,

⇑ says that the agent strongly trusts the source and hence will perform radical upgrades with whatever it is announced,

⇑ says that the agent barely trusts the source and hence will perform conservative upgrades etc.

‘Fallible AGM’ $FAGM$ and ‘Minimal AGM’ $MAGM$ are higher-level of doxastic attitudes, that are consistent with all the attitudes above.
Possible Generalizations

Of course, one could also consider announcements from different sources, in which case we’d have to label the dynamic modalities with the name of the source $i$, as in $[∗φ]_i$, and have atomic sentences $i :!, i :↑, i :↑$ etc expressing the agent’s attitude towards each source $i$.

One can generalize further to a multi-agent setting: then we have plausibility relations labeled by agents $≤_i$, doxastic operators $B_i$ for each agent $i$; the sources are then agents, and for each pair $i, j$ of agents, we have atomic sentences $i, j :!, i, j :↑$, expressing agent $i$’s attitude towards the source $j$. 
Semantic Constraints

We put some natural semantic constraints (on the valuation of the atomic sentences of the form $\tau$, for any doxastic attitude type $\tau$).

The first says that, in any possible world, the agent has some attitude towards the source:

$$\forall s \exists \exists! \tau \text{ such that } s \models \tau$$

The second is an introspection postulate: the agent knows his own doxastic attitude towards the source:

$$s \sim t \implies (s \models \tau \iff t \models \tau).$$
For attitudes that are incompatible, one could add some mutual incompatibility constraints, e.g.

\[ s \models ! \implies s \not\models \top \]

But, in fact, these are not really necessary (since different attitudes will lead to mutually inconsistent beliefs after revision with the same info).
The modality $[\ast \varphi]$ will be interpreted using the agent’s doxastic attitude towards the source: the transformation $\ast \varphi$ is an upgrade that reorders each partition cell $[s]$ by applying the corresponding type of upgrade $\tau$, where $s \models \tau$.

If this reordering is inconsistent for a given type $\tau$, then the corresponding worlds $s \models \tau$ are eliminated: $\ast \varphi$ cannot be executed in them!
Validities

\[ \tau \Rightarrow K \tau \]

\[ ! \Rightarrow ([* \varphi] B \psi \iff (\varphi \Rightarrow B^{\varphi}[* \varphi] \psi)) \]
(Conditional) Obligations

We now add a hierarchy of obligations to our models, in the form of a total preorder $\preceq$ on states in $S$, called deontic (pre)order.

Obligations and conditional obligations are defined similarly to beliefs and conditional beliefs, but by taking the maximal states (in the whole state space) with respect to $\preceq$:

$$ s \models O\psi \text{ iff } \text{Max}_{\preceq}S \subseteq \|\psi\|, $$

and

$$ s \models O(\psi|\varphi) \text{ iff } \text{Max}_{\preceq}\|\varphi\| \subseteq \|\psi\|. $$
We can now capture dynamic-doxastic norms, as well as their revision, via (conditional) obligations concerning the agent’s doxastic attitudes towards the source of information.
Example: the maximal-trust hierarchy of norms

Suppose that, in the Student example, the source is required to trust as much as possible the source of information (the Teacher), in the following sense: he should consider the source infallible, unless this is inconsistent; otherwise he should have strong trust (performing radical upgrades), unless this is inconsistent; otherwise, he should consider conservative upgrade, unless this is inconsistent; etc.

Even more, he should obey FAGM, unless this is inconsistent; otherwise, he should obey MAGM.
Modeling a norm hierarchy

For this, we set up a relation $\preceq$, such that: all $\preceq$-maximal worlds satisfy $!$; the next best ones satisfy $\uparrow$; etc. Moreover, all the worlds satisfying FAGM are better than the others; finally, among the remaining ones, the MAGM are better than the others.

In such a model, all states satisfy:

$$O(!) \land O(\uparrow | \neg !) \land O(\uparrow | \neg ! \land \neg \uparrow) \land \ldots$$

and

$$O(FAGM) \land O(MAGM | \neg FAGM).$$
Validities

\[ O(!) \implies O([\ast \varphi] B\psi \iff (\varphi \implies B^{\varphi} [\ast \varphi] \psi)) \]

\[ O(!) \implies ([\ast \varphi] O\psi \iff (\varphi \implies O^{\varphi} [\ast \varphi] \psi)) \]
We could now consider *norm-changing actions* (similar to the belief-changing actions).

E.g. the *radical norm change* $\uparrow_0 \varphi$ will change the relation $\preceq$ in such a way that all $\varphi$-worlds become $\preceq$-better than all non-$\varphi$-worlds (and within the two zones, the old $\preceq$-order is maintained).

But, even more interestingly, even *simple belief-revision actions induced by announcements* can change norms!
Example

\[ O(!) \implies [(p \land \neg Bp)][(p \land \neg Bp)] \neg O(!) \]

The “infallibility” norm is shattered after the second announcement of the (same) Moore sentence!
EXAMPLE: 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 student will still not be sure that the exam is tomorrow.
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 student 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!
The natural interpretation of Teacher’s announcement is self-referential:

“(EVEN after I’m telling you this) the exam’s date WILL be a surprise: the evening before the exam day, you will not believe that the exam is tomorrow”

Gerbrandy thinks that this version gives rise to a genuine (Liar-like) paradox: i.e. according to Gerbrandy, there is simply no solution to this “self-referential” version of the puzzle!
The sentence saying "the evening before the exam day, you will not believe that the exam is tomorrow" can be formalized as:

\[
\text{surprise} = \bigwedge_{1 \leq i \leq 5} \left( i \rightarrow \left[ \neg \left( \bigwedge_{1 \leq j < i} \neg j \right) \right] \neg Bi \right).
\]
Equivalent Formula

By the Reduction Laws for beliefs after updates, this reading of \textit{surprise} is equivalent to:

\[
surprise = \bigwedge_{1 \leq i \leq 5} \left( i \rightarrow \neg B(i \mid \bigwedge_{1 \leq j < i} \neg j) \right).
\]
But the announced sentence is NOT “surprise”.

It is a self-referential sentence, saying “even after I am announcing you this, the sentence surprise will (still) hold”.

In other words, this is an upgrade

\(*(NEXT \text{surprise})*\).
Assume Maximal-Trust Hierarchy

Let us assume the norm hierarchy described above, encoding the duty of the student to *give as much credit to the teacher as it is consistent to give*:

\[ O(!) \land O(\uparrow | \neg!) \land O(\uparrow | \neg! \land \neg \uparrow) \land \ldots \]

and

\[ O(FAGM) \land O(MAGM | \neg FAGM). \]
Updates are Ruled Out

But, as in the case of the double announcement of a Moore sentence, updates can be easily ruled out!

Indeed, an update !(NEXT surprise) leads immediately to paradox!

The proof is as follows:
The Impossibility Proof

The validity

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

together with the validity

\[ BEFORE (NEXT \text{surprise}) \Leftrightarrow \text{surprise}, \]

gives us that

\[ ![\text{(NEXT surprise)}] K\text{surprise}. \]
BUT it is easy to see that, given our assumption that the student knows there’ll be an exam, the sentence $K \text{surprise}$ is inconsistent: indeed, we have the validity

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

Using the persistence under updates of our above assumption (that $K(\bigvee_{1 \leq i \leq 5} i)$), we conclude that such an update is impossible:

$$K(\bigvee_{1 \leq i \leq 5} i) \Rightarrow \neg K \text{surprise} \Rightarrow \neg [(\neg \text{NEXT surprise})]FALSE.$$
 Radical and Conservative Updates are Ruled Out

In fact, \((\text{NEXT surprise})\) CANNOT be a radical NOR a conservative upgrade: both 
\(\uparrow (\text{NEXT surprise})\) and \(\uparrow (\text{NEXT surprise})\) lead to paradox. They are “impossible events”, similarly to \(! (\text{NEXT surprise})\). The proof is similar to the above one.

What this means is that \((\text{NEXT surprise})\) is not executable in any world satisfying \(!, \uparrow \) or \(\uparrow\). These worlds do not survive \((\text{NEXT surprise})\), hence the corresponding obligations are overruled!
Any FAGM is ruled out

More generally, any doxastic attitude satisfying the FAGM principle is ruled out!

Using the equivalence

\[ \text{BEFORE}(\text{NEXT surprise}) \Leftrightarrow \text{surprise} \]

we can see that

\[ \text{FAGM} \Rightarrow (K\neg\text{surprise} \lor [*(\text{NEXT surprise})]B\text{surprise}). \]
But it is easy to see that $B\text{surprise}$ is inconsistent, given our assumption that the student knows there’ll be an exam: we have the validity

$$K\left( \bigvee_{1 \leq i \leq 5} i \right) \implies \neg B\text{surprise}.$$  

Hence we have

$$FAGM \implies K\neg \text{surprise}.$$  

But the sentence $K\neg \text{surprise}$ is false in the original model!
The Only Remaining Obligation: MAGM

So we obtain that

\[ \phi \not\rightarrow O(FAGM) \]

By the norm hierarchy that we postulated, we must have:

\[ \phi \not\rightarrow O(MAGM) \]

The student’s only remaining obligation towards the teacher is to try to respect the Minimal AGM principle. Is this consistent, or we have indeed a genuine paradox?
Existence and Uniqueness of the Solution

THERE DOES exist a soft upgrade (satisfying MAGM) with the sentence $NEXT \ surprise_B$.

So the paradox has a solution!

Moreover, there is ONLY ONE such upgrade: the solution is unique!
Proof

Let $T$ be such a soft upgrade with $NEXT\ surprise_B$. Then, using the equivalence

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

and FAGM principle, we obtain that:

$$[T]\neg K\neg surprise \implies [T]Bsurprise.$$
But, as we saw already, \( B_{\text{surprise}} \) is \textbf{inconsistent}, given our assumption that the student \textit{knew} there’ll be an exam: we have the validity

\[
K\left( \bigvee _{1 \leq i \leq 5} i \right) \Rightarrow \neg B_{\text{surprise}}.
\]

Hence we have

\[
[T]\neg K\neg \text{surprise} \Rightarrow [T]\text{FALSE}.
\]

This means that such an upgrade \( T \) is \textbf{possible (executable)} \textbf{ONLY} if \( \neg[T]\neg K\neg \text{surprise} \) holds in the original model, i.e. if \( K\neg \text{surprise} \) holds in the new model after the upgrade.
Finally, it is easy to see that $K$–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

```
1 ← 2 ← 3 ← 4 ← 5
```
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!

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.
Irreversibility of the belief (and norm) change

Starting with a *dynamic-doxastic norm hierarchy* that requires **maximizing student’s trust** in the teacher (as much as it is logically consistent), the student arrived at the *opposite of trust*: he came to know that the teacher *lied*.

However, the logical path for showing this involved a process of belief *revision* (inducing a process of norm revision) by the student. This revision process, though induced by a “duty of trust”, is *irreversible*: it can no longer be undone in the end, despite the loss of trust.