Iterated Dynamic Belief Revision

Sonja Smets, University of Groningen

website:

http://www.vub.ac.be/CLWF/SS

Joint work with Alexandru Baltag, COMLAB, Oxford University

Belief Revision

Standard Belief Revision (BR) theory:

- the AGM (Alchourrón, Gärdenfors, Makinson) axioms;
- the KM (Katsuno-Mendelzon) postulates;
- semantics (Grove sphere models);
- extensions and alternatives (see e.g. H. Rott).

The "Success" Postulate

BR assumes as given:

1. *theories* ("belief sets") T
2. *new information* (a formula) φ
3. a *revision operator* $T * \varphi$

I mention here only one of the AGM axioms, also accepted by KM, but which poses **problems** when revising with **doxastic** sentences:

"Success": $\varphi \in T * \varphi$

Interpretation

$T * \varphi$ is supposed to represent the *new theory after learning* φ: the agent's new set of beliefs, given that the initial set of beliefs was T and that the agent has learned φ (and only φ).

But, beliefs about *what*? Beliefs have to have a *content* and a *referent*: you believe something about something. The content is a *sentence* in some language L. So theories are subsets $T \subseteq L$. The referent is "reality": you believe the sentence to hold in the "real world".

Reality

There always is a *real state* of the world s that our theories are about.

A theory T is *true* in state s iff $s \models T$.

Problems:

Classical belief revision, as well as its more "dynamic" KM version, encounter **a number of problems**:

1. **iterating belief revision**;
2. dealing with **higher-order beliefs** (beliefs about other beliefs);
3. dealing with **multi-agent beliefs**;
4. the **multiplicity of possible belief revision policies**: there is no unique $*$.

In this talk we'll focus more on **the first two problems**.

Higher-Order Beliefs: "No Success"

Take a Moore sentence:

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

After φ is learned, φ obviously becomes *false*!

But the Success Postulate asks us to believe (after learning φ) that φ is true! In other words, it forces us (as a principle of rationality!) to acquire false beliefs!

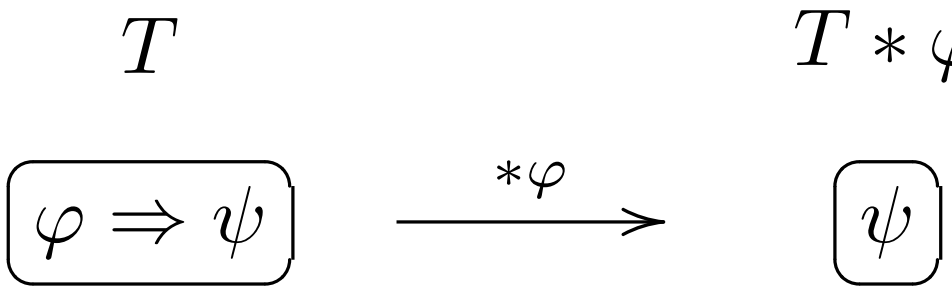
The usual way to deal with this: simply accept that AGM cannot deal with higher-order beliefs, so limit the language L to formulas that express only *"factual", non-doxastic properties of the world.*

But the interpretation we will give to $*$ shows that AGM does not need to be limited in this way: it is perfectly natural and coherent as a belief revision theory about a fixed state of the world (the original state s_0).

The Ramsey Test

The Ramsey Test asks for some notion of "conditional" $\Rightarrow$ such that, for all the possible belief sets T, we have:

$$(\varphi \Rightarrow \psi) \in T \quad \text{iff} \quad \psi \in T * \varphi.$$



Impossibility theorem

A key result in the subject is Gärdenfors' *Impossibility Theorem*: there is no notion of conditional satisfying the Ramsey test with respect to some (non-trivial) operation of revision on belief sets that satisfies the AGM postulates.

Changing beliefs about an unchanging world

The assumption underlying AGM theory is that *the "world" that our beliefs are about is not changed by our changes of belief.*

But the "world" the higher-order beliefs are about includes the beliefs themselves.

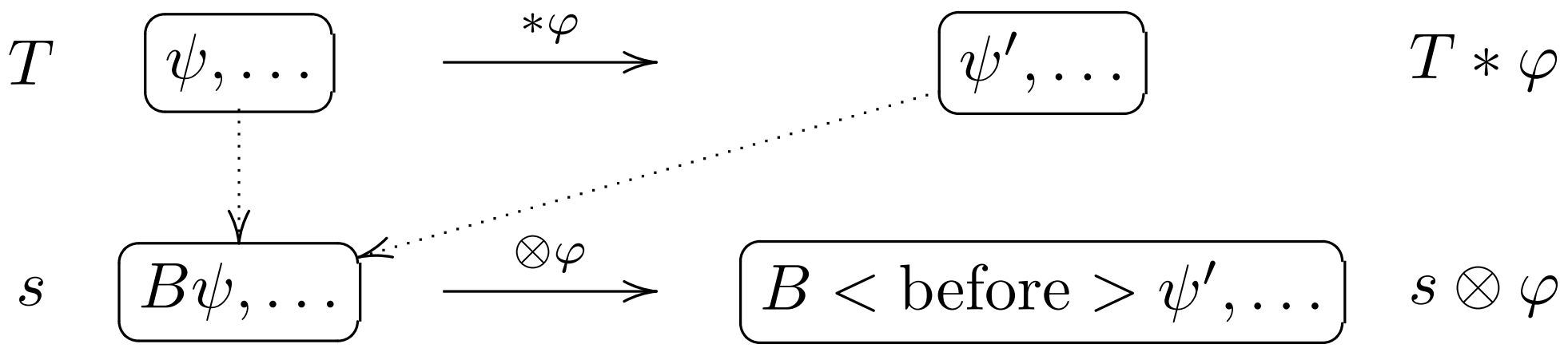
So (as the example of Moore sentences shows) the "world", in this sense, is *always changed by our changes of belief*!

"Saving" AGM

Nevertheless, we can **reinterpret** the AGM postulates to make them applicable to doxastic sentences:

If T is the belief set at a given moment about the real state s at that moment, then $T * \varphi$ should be understood as a belief set about *the same* state s, as it was *before* the learning took place.

In other words, $T * \varphi$ captures *the agent's beliefs after learning φ about what was the case before the learning.*



Here, $s \otimes \varphi$ is the state of the world after the agent learns that φ *was* true in the (initial) state s of the world.

Also, $<$ before $> \theta$ is a past tense operator, saying that θ was true before the last action.

"Static" AGM revision

Let

$$th(s) \;=:\; \{\psi \in L : s \models B\psi\}$$

be the agent's original belief set (theory) in state s about the (original) state s. Then

$$th(s) * \varphi \;\neq\; th(s \otimes \varphi).$$

Instead,

$$th(s) * \varphi \;:=\; \{\psi : < \text{before} > \psi \in th(s \otimes \varphi)\,\}$$

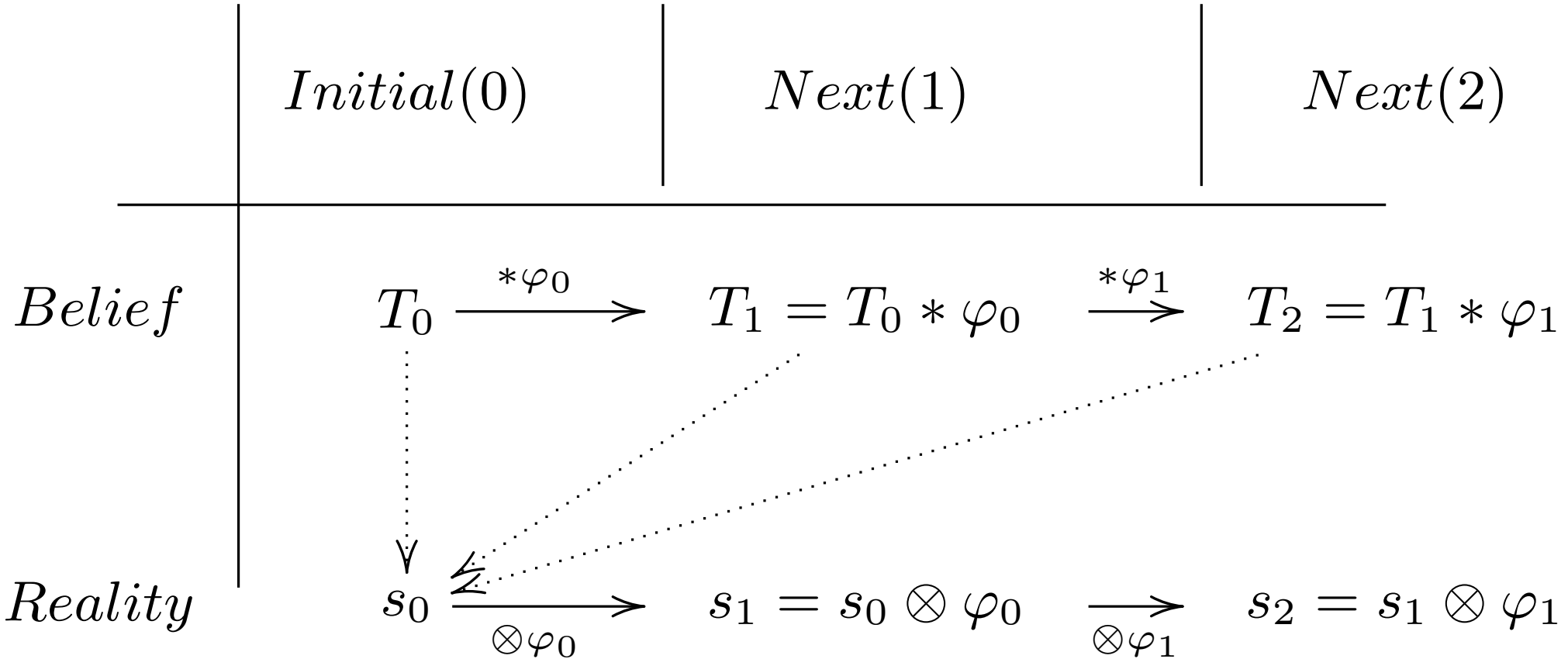
So AGM belief revision is *"static" in an essential sense*!

Iteration

This is exactly what poses problems for *iterating revision*:

what state is $(T * \varphi) * \psi$ about?

Well, the answer is: it is again about *the same state* s that T was about.


Initial(0)
Next(1)
Next(2)
Belief
T_0
$*\varphi_0$
$T_1 = T_0 * \varphi_0$
$*\varphi_1$
$T_2 = T_1 * \varphi_1$
Reality
s_0
$\otimes\varphi_0$
$s_1 = s_0 \otimes \varphi_0$
$\otimes\varphi_1$
$s_2 = s_1 \otimes \varphi_1$

"Static" versus "dynamic" revision

Compare this with a notion of **"dynamic" revision**, defined as

$$th(s) \otimes \varphi := th(s \otimes \varphi)$$

This operator will **NOT satisfy the Success postulate**.

But it does **something better**: *it keeps up with the change of reality.*

Iterated dynamic revision

$$T^0 = th(s_0), T^{i+1} = T^i \otimes \varphi_i = (\cdots(th(s_0) \otimes \varphi_0)\cdots) \otimes \varphi_i$$

	$Initial(0)$	$Next(1)$	$Next(2)$
$Belief$	T^0	$T^1 = T^0 \otimes \varphi_0$	$T^2 = T^1 \otimes \varphi_1$
$Reality$	s_0	$s_1 = s_0 \otimes \varphi_0$	$s_2 = s_1 \otimes \varphi_1$

$T^0 \xrightarrow{\otimes\varphi_0} T^1 = T^0 \otimes \varphi_0 \xrightarrow{\otimes\varphi_1} T^2 = T^1 \otimes \varphi_1$

$T^0 \vdots\!\downarrow\ \ T^1 = T^0 \otimes \varphi_0 \vdots\!\downarrow\ \ T^2 = T^1 \otimes \varphi_1$

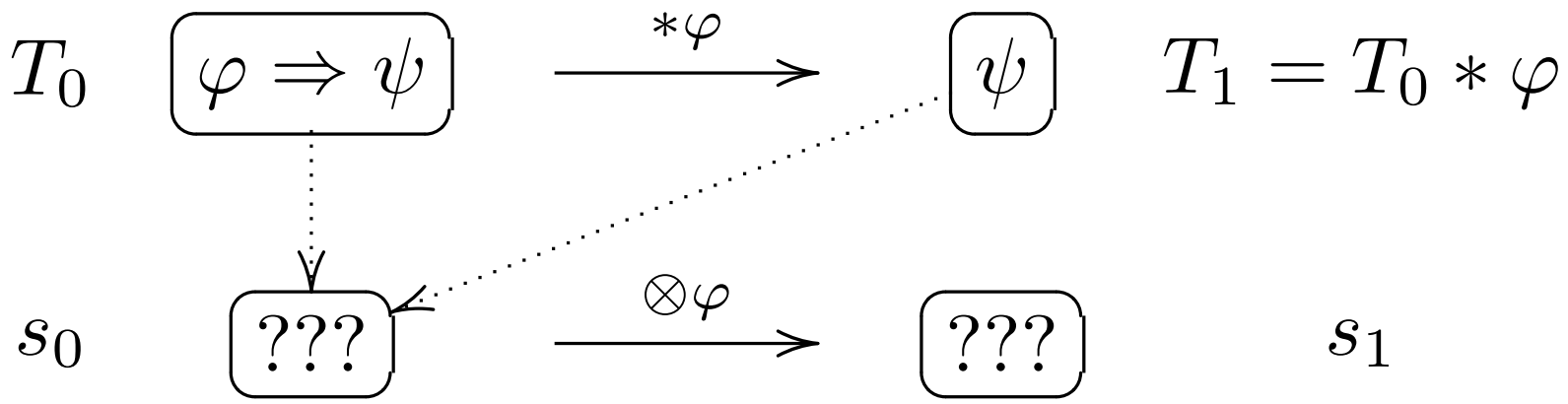
$s_0 \xrightarrow[\otimes\varphi_0]{} s_1 = s_0 \otimes \varphi_0 \xrightarrow[\otimes\varphi_1]{} s_2 = s_1 \otimes \varphi_1$

So what's wrong with the Ramsey test?

Suppose Ramsey test holds for "static" revision $*$, i.e. there exists a conditional $\Rightarrow$ such that

$$(\varphi \Rightarrow \psi) \in T \quad \text{iff} \quad \psi \in T * \varphi.$$

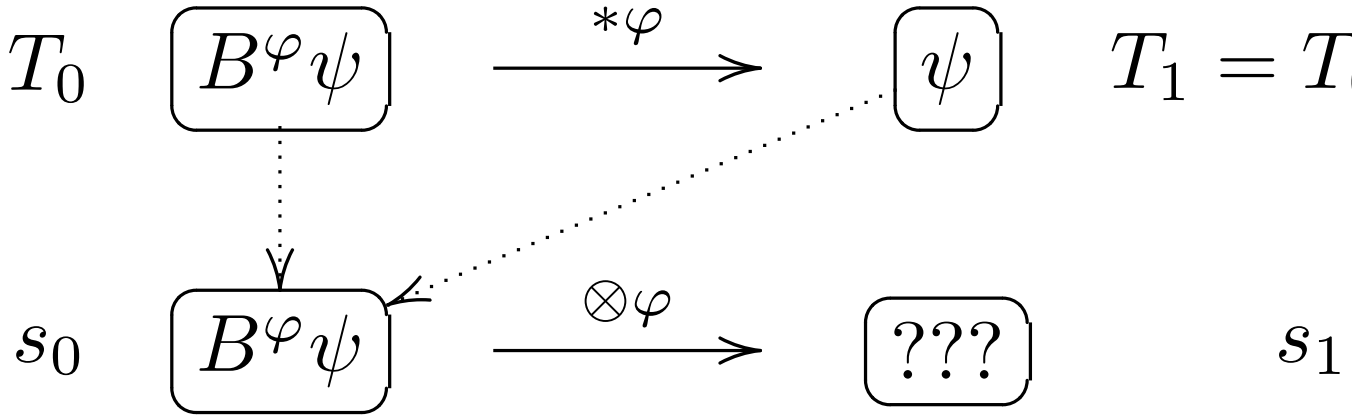
holds for all belief sets T. Then in particular, it must hold for theory $T_0 = th(s_0)$, capturing the agent's original beliefs in the original state of the world:



But then what is the *semantics* of $\Rightarrow$? Is $\varphi \Rightarrow \psi$ true in the original world s_0? Obviously, the above clause implies that $\varphi \Rightarrow \psi$ *cannot be a factual* property of the world. It is a *doxastic* property, that expresses a feature of the agent's *belief revision policy*: if given information φ, the agent would come to believe that ψ was the case. In other words, $\varphi \Rightarrow \psi$ is in fact a *conditional belief* statement $B^{\varphi}\psi$.

Conditional beliefs

But a fully introspective agent *knows* his conditional beliefs (or his belief revision policy)! So, $B^{\varphi}\psi$ is believed at s_0 iff it is true at s_0:



This gives us the desired truth clause for the doxastic "conditional" $B^{\varphi}\psi$:

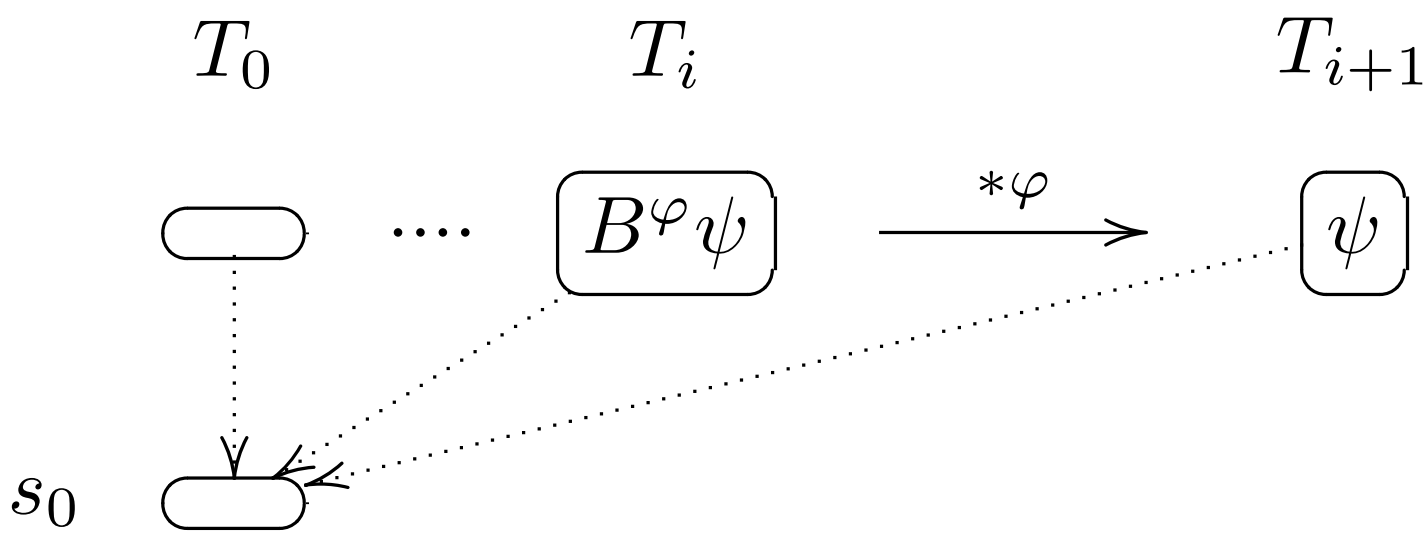
$$s \models B^{\varphi}\psi \text{ iff } \psi \in th(s) * \varphi$$

Weak Ramsey test

We succeeded to *define* a conditional operator that satisfies a weak version of the Ramsey test, stated only for the original theory $th(s)$ at any state s:

$$B^{\varphi}\psi \in th(s) \quad \text{iff} \quad \psi \in th(s) * \varphi$$

Can this possibly satisfy the Ramsey test in its original (strong) version? This would require it to be true at all iterated theories $T_i = (\cdots(th(s_0) * \varphi_0)\cdots) * \varphi_{i-1}$.



$B^\varphi\psi \in T_i$ iff (??) $\psi \in T_i * \varphi$

Ramsey test must fail for iterated theories

But recall these theories are all about the initial state s_0! But, by introspection, the agent already *knows* whether or not $B^{\varphi}\psi$ holds at s_0, and so this belief is *not subject to revision*:

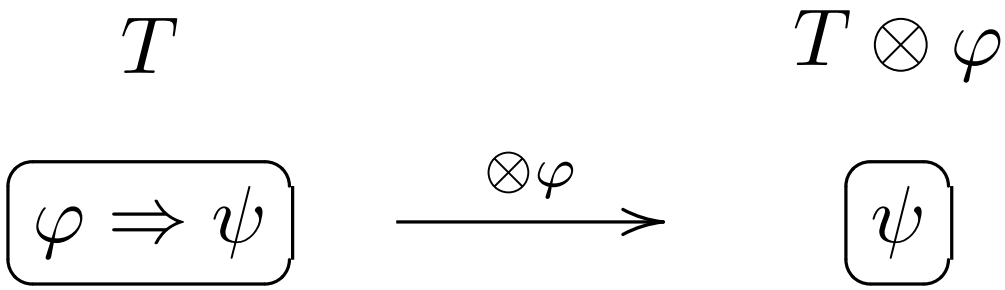
$$B^{\varphi}\psi \in T_i \quad \text{iff} \quad B^{\varphi}\psi \in T_0 \quad \text{iff} \quad s_0 \models B^{\varphi}\psi \quad \text{iff} \quad \psi \in T_0 * \varphi$$

In other words, the presence of $B^{\varphi}\psi$ in any of the iterated revised theories T_i is equivalent to its presence in T_0, and has only to do with the one-step revision $T_0 * \varphi$ of T_0, but *not* with the revision $T_i * \varphi$ of T_i!

Dynamic Ramsey test

However, the Ramsey test can hold for *dynamic* revision: there *does exist* a conditional $\Rightarrow$ such that

$$(\varphi \Rightarrow \psi) \in T \text{ iff } \psi \in T \otimes \varphi.$$



Dynamic Ramsey test - continued

Take the **dynamic "learning" modality**, given by

$$s \models [\varphi]\psi \;\text{ iff }\; s \otimes \varphi \models \psi \text{ whenever } s \models \varphi$$

Then Ramsey test for dynamic revision $\otimes$ holds if we take as our conditional the updated belief operator $[\varphi]B\psi$:

$$[\varphi]B\psi \in T \;\text{ iff }\; \psi \in T \otimes \varphi.$$

"Static" and "Dynamic" Conditionals

The conditional $B^{\varphi}\psi$ is "*static*" w.r.t. the actual state of the system: the state doesn't change, only the agent's belief about it changes. On the other hand, $[\varphi]B\psi$ is a "*dynamic*" conditional: the state changes (the agent learns φ), and after that the agent beliefs ψ.

In fact, we *can* understand conditional beliefs *dynamically*: $B^{\varphi}\psi$ means that, after learning φ, the agent believes that ψ *was* true (in the original state) *before* the learning:

$$B^{\varphi}\psi = [\varphi]B < \text{before} > \psi$$

SEMANTICS: Plausibility (Grove) Models

I only present the **finite-state, single-agent** case. The infinite-model, multiple-agent case is more complex.

A **(finite) plausibility frame** is a *finite set* S of "*states*" (or "*possible worlds*") together with a **connected preorder** $\leq \subseteq S \times S$, called *plausibility relation.*

"**Preorder**": *reflexive* and *transitive.*
"**Connected**" (or "complete"): for all states $s, t \in S$, we have *either* $s \leq t$ *or* $t \leq s$.

This is essentially *the same as a finite Grove sphere model.*

(Conditional) Belief in Plausibility Models

A sentence φ is **believed** in (any state of) a plausibility model $(S \leq)$ if φ **is true in all the "most plausible" worlds**; i.e. in all "minimal" states in the set

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

More generally, a sentence φ is **believed conditional** on P (in which case we write $B^P\varphi$) if φ **is true at all most plausible worlds satisfying P**; i.e. in all the states in the set

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

Example 1

A coin is flipped in front of an agent (let's call her Alice), landing on the table face-up but **being immediately covered**, so that **Alice cannot see the face**.

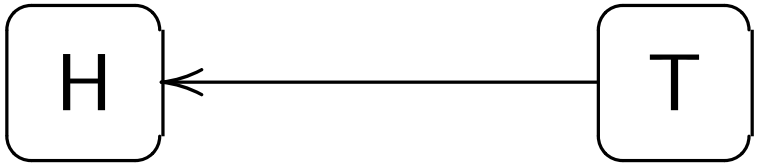
Nevertheless, Alice **believes that the upper face is Heads** (say, because *she had a brief glimpse* at the coin before it was covered, and *thought that she may have seen it Heads up*).

And in fact, **she's right**: *although she doesn't know it* for sure, **the coin is actually lying Heads up**.

A Model for Example 1

Drawing Convention: We use *labeled arrows* to represent **the converse plausibility relation** $\geq$ (going from less plausible to more plausible states), but for convenience **we skip all the loops** (since $\geq$ is reflexive).

A model S for Example 1:



The **actual state** s is the one **on the left.** The agent **truthfully believes Heads is up**; but she **doesn't know** that Heads is up.

Dynamic Belief Revision: upgrades

From a *semantic* point of view, dynamic belief revision is about "revising" the whole relational structure: *changing the plausibility order* (or the models).

An **upgrade** is a **model-changing operation** α, that takes a plausibility model $\mathbf{S} = (S \leq)$ and returns a new model $\alpha(\mathbf{S}) = (S, \leq')$, **having the same set of states**(but **possibly a different order relation**).

An upgrade essentially encodes a belief-revision policy. Any upgrade gives a possible semantics to both the static and the dynamic revision operators ($*$ and $\otimes$).

Examples of Upgrades

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

Explanation

After a *conservative upgrade*, the agent only comes to **believe** that φ (was the case).

The *lexicographic upgrade has a more "radical" effect*: the agent comes to **accept φ with such a conviction** that **she considers all φ-possibilities more plausible than all non-φ ones**.

Dynamic Modalities

The **dynamic "learning" modality** $[\varphi]\psi$ introduced above in an an informal way **can now be defined precisely**, but its meaning (and notation) will **depend on the belief revision policy**:

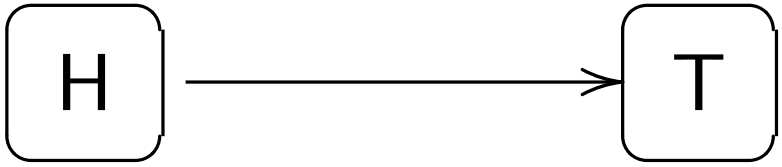
For *Lexicographic Upgrades*: $[\Uparrow \varphi]\psi$

For *Conservative Upgrades*: $[\uparrow \varphi]\psi$

Example 2: An Upgrade

Alice **learns from a trusted informer** that the coin lies **Tails up**, but **without being shown the face**. Alice **believes** him.

But, no matter how trusted is this informer, he is **not infallible**. (In fact, in this case he **is wrong**: the coin was Heads up!) Taking this as a (lexicographic or conservative) upgrade, the **new, upgraded model** is:



Fixed Points

A plausibility model $\mathbf{S} = (S, \leq)$ is a **fixed point** for an upgrade α iff **S is left "unchanged" by** α: i.e. the changed model $\alpha(\mathbf{S})$ is **"the same"** as **S**.

(Technically, this means that $\alpha(\mathbf{S})$ is **bisimilar** to **S**.)

Informally, **S** is a fixed of an upgrade α iff α is **not an "informative" upgrade** of **S**: it is a "redundant" one, leaving unchanged the agent's conditional belief structure.

Truthfulness

If we fix a state $s \in S$, called "the **actual world**", in a plausibility model $(S, \leq)$, then we say that an upgrade ($\Uparrow \varphi$ or $\uparrow \varphi$) is **truthful** if φ is true in the actual world s.

Iterating Upgrades

To study iterated belief revision, consider a **finite model** $\mathbf{S}_0 = (S, \leq_0)$, and an (**(infinite) sequence of upgrades**, either of the conservative kind

$$\uparrow \varphi_1, \ldots, \uparrow \varphi_n, \ldots$$

or the lexicographic kind

$$\Uparrow \varphi_1, \ldots, \Uparrow \varphi_n, \ldots$$

This leads to **an infinite succession of upgraded models**

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

$$\text{defined by:} \quad \mathbf{S}_i = \alpha_i(\mathbf{S}_{i-1}),$$

where $\alpha_i =\uparrow \varphi_i$ in the first case, and $\alpha_i =\Uparrow \varphi_i$ in the second case.

Iterated Upgrades Do Not Necessarily Stabilize!

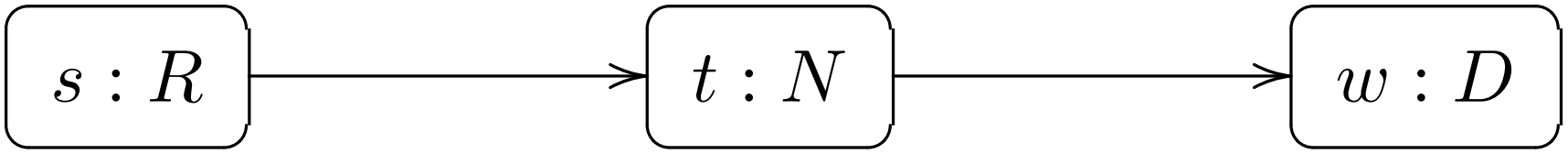
Proposition **For some initial finite models, there exist infinite cycles of truthful upgrades (that never stabilize the model).**

Even worse, this **still holds** if we restrict to iterations of **the same** truthful upgrade (with **one fixed sentence**): no fixed point is reached.

Moreover, when iterating **conservative** upgrades, **even the (simple, unconditional) beliefs may never stabilize, but may keep oscillating forever.**

Iterating a Truthful Conservative Upgrade

Consider a pollster (Charles) with the following beliefs about how a given voter (Mary) will vote:



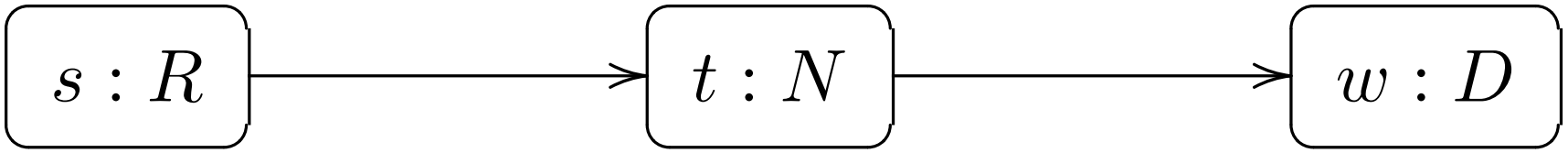
As before, we skip the loops and the arrows that can be obtained by transitivity. We assume the real world is s, so Mary will vote Republican (R)! But Charles believes that she will vote Democrat (D); and in case this turns out wrong, he'd rather believe that she won't vote (N) than accepting that she may vote Republican.

A trusted informer studied Mary's voting behavior and tells Charles the following true statement φ:

$$R \vee (D \wedge \neg BD) \vee (\neg D \wedge BD)$$

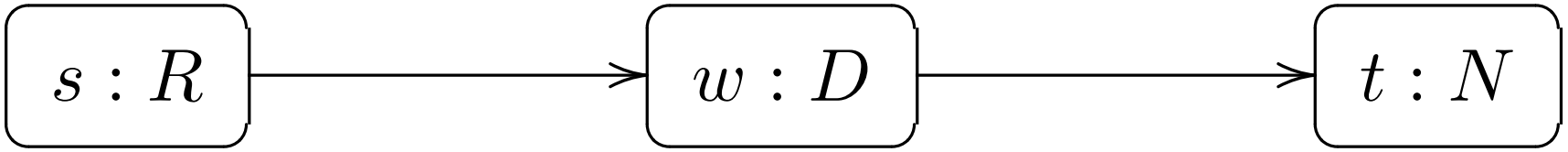
"*Either Mary will vote Republican or else your beliefs about whether or not she votes Democrat are wrong*".

The sentence φ is true in states s and t but not in w:



Infinite Oscillations by Truthful Upgrades

Let's suppose that Charles **conservatively upgrades** his beliefs with this new true information φ. The most plausible state satisfying φ was t, so this becomes now the most plausible state overall:



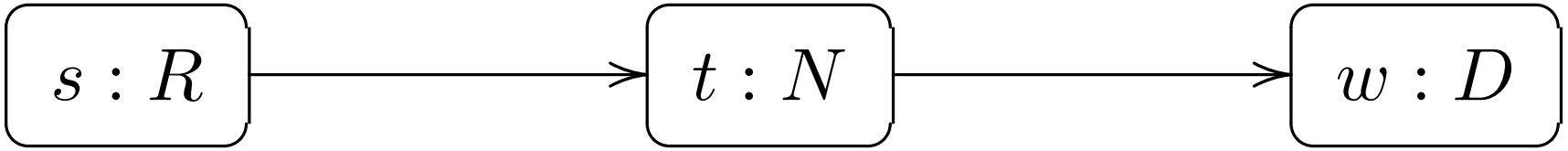
In this new model, the sentence φ is *again true at the real world (s), as well as at the world w*. So **this sentence can again be truthfully announced**.

However, if Charles **conservatively upgrades again** with this new true information φ, he will promote w as the most plausible state, **reverting to the original model**!

Moreover, **not only the whole model (the plausibility order) keeps changing**, but Charles' (simple, un-conditional) **beliefs keep oscillating forever** (between D and N)!

Iterating Truthful Lexicographic Upgrades

Consider the same model with the same real state s:

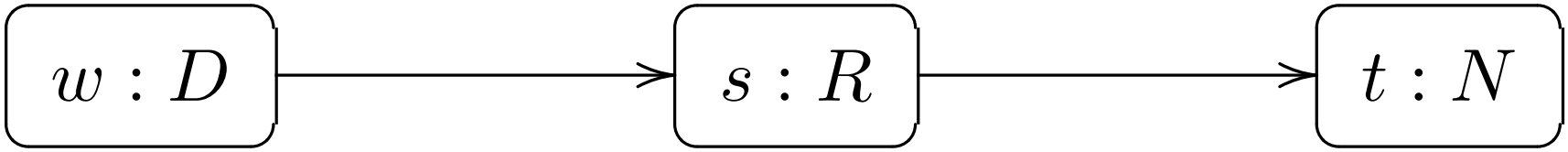


But now consider the sentence

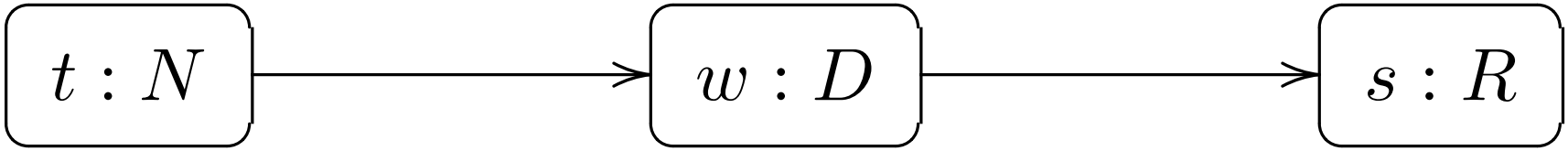
$$R \vee (D \wedge \neg B^{\neg R} D) \vee (\neg D \wedge B^{\neg R} D)$$

"*If you'd truthfully learn that Marry won't vote Republican, then your resulting belief about whether or not she votes Democrat would be wrong*".

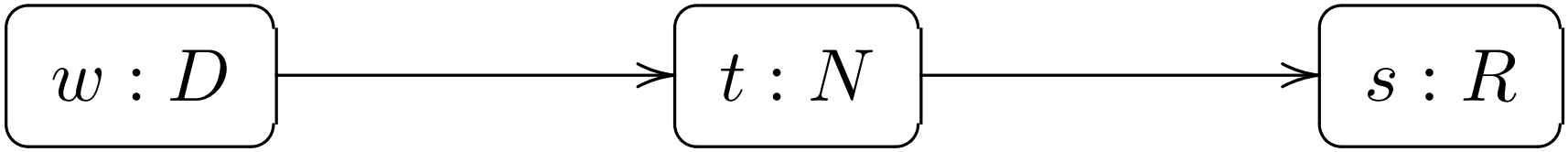
This sentence is true in the real state s and in t but not in w, so a **truthful lexicographic upgrade** will give us:



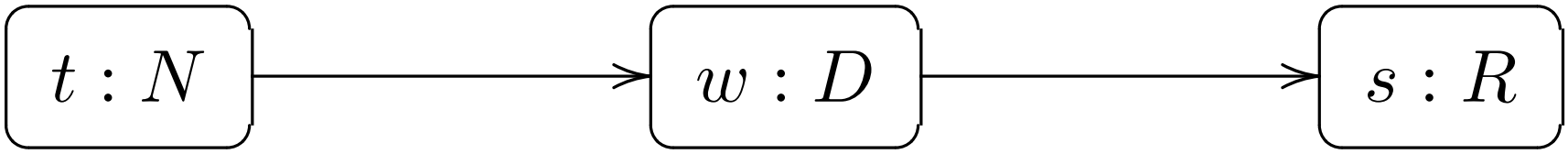
The same sentence is again true in (the real world) s and in w, so it can be again truthfully announced, resulting in:



Another truthful upgrade with this sentence produces



then another truthful upgrade with the same sentence **gets us back** to



Stable Beliefs in Oscillating Models

Clearly from now on the last two models **will keep reappearing, in an endless cycle**: as for conservative upgrades, the process never reaches a fixed!

However, *unlike in the conservative upgrade example*, **in this example the simple (unconditional) beliefs eventually stabilize**: from some moment onwards, Charles correctly believes that the real world is s (vote Republican) and he will never lose this belief again!

This is a symptom of a more general phenomenon:

Beliefs Stabilize in Iterated Lexicographic Upgrades

Proposition:

In any infinite sequence of truthful lexicographic upgrades $\{\Uparrow \varphi_i\}_i$ on an initial (finite) model $\mathbf{S}_0$, the set of most plausible states stabilizes eventually, after finitely many iterations.

From then onwards, the simple (un-conditional) beliefs stay the same (despite the possibly infinite oscillations of the plausibility order).

Upgrades with Un-conditional Doxastic Sentences

Moreover, if the infinite sequence of lexicographic upgrades $\{\Uparrow \varphi_i\}_i$ consists only of sentences belonging to the language of basic doxastic logic (allowing only for simple, un-conditional belief operators) then the model-changing process eventually reaches a fixed point: after finitely many iterations, the model will stay unchanged.

As we saw, this is not true for conservative upgrades.

Conclusions

Iterating static belief revision suffers from the failure of the (strong) Ramsey Test.

"Dynamic" belief revision satisfies the Ramsey Test.

Iterated upgrades may never reach a fixed point: conditional beliefs may remain forever unsettled.

When iterating truthful lexicographic upgrades, simple (non-conditional) beliefs converge to some stable belief.

Truthful conservative upgrades do not have this last property.