Group Belief Dynamics under Iterated Upgrades

1. Belief Convergence (or endless cycles?) by Iterated Learning
2. Belief Merge (or endless disagreements?) by Iterated Sharing (and Persuasion)

Alexandru Baltag, Oxford University

Based on joint work with Sonja Smets, University of Groningen

Iterated Revision with Doxastic Information

GENERAL PROBLEM: investigate the **long-term behavior of iterated "learning" or "sharing"** of **higher-level doxastic information**.

"**Learning**": belief revision with new **true** information.

"**Sharing**": revision induced by **sincere** *communication* by either of the agents (the "*speaker*").

Sincerity: *(s)he already accepts that information* (before sharing it).

Higher-level (doxastic) information: may refer to the agents' own *beliefs*, or even to their *belief-revision plans*.

Where does all this go?

Long-term behavior:

What happens in the long run?

Where does all this go?

CONVERGENCE Pb: Does "Learning" Ever End?

SPECIAL PROBLEM 1: Does the "learning" process necessarily **come to an end, by stabilizing** the agents' doxastic structure into a **fixed-point**; or can it *keep changing forever*, in **infinite cycles**?

If the second is the case, then what are the conditions for reaching a fixed point?

What are the conditions for *stabilizing on the truth* (=converging to **true beliefs**)?

Obvious connections to Learning Theory.

MERGE Pb: Does "Sharing" Ever End?

SPECIAL PROBLEM 2: does the "sharing" process necessarily **come to an end, by merging** the agents' doxastic structure?

Obvious connections to the problem of "preference aggregation" in Social Choice Theory. "*Aggregating beliefs*" (or rather, *belief structures*).

What types of merge can be dynamically realized by what type of "sharing"?

Do the **communication agenda** and the **group's hierarchy** make any difference?

Contrast with Classical Theory

Classical Belief Revision theory deals only with learning of **propositional** information by a **single agent**.

In that context, the process of learning new, true information **always converges** (in finitely many steps if the initial model was finite): *the most* one can learn by iterated revisions is *the correct valuation* (all the true).

Moreover, *in that context* **it is useless to repeatedly revise** with the **same** information: after learning a *propositional* sentence φ once, learning it again would be **superfluous** (leaving the doxastic state *unchanged*).

Preference Models for Information

Interpret the accessibility relation R_i of a multi-modal Kripke model as a "doxastic preference" , a **plausibility relation**, meant to represent *"soft" information*: in this reading, sR_it means that **world** t **is at least as plausible for agent** a **as world** s. For this interpretation, it is customary to use the notation $s \geq_i t$ for the plausibility relation R_i (and $\leq_i$ for its converse).

SEMANTICS: Epistemic Plausibility Structures

A **(finite) plausibility frame**:

$$\mathcal{M} = < I, W, (\leq_i)_{i \in I} >$$

- I a finite set of **agents**
- W a **finite** (and non-empty) set of **states** ("worlds")
- $\leq_i$ **"locally connected" preorders** on W ("i**'s plausibility order**"), one for each agent $i \in I$

Read $s \leq_i t$ as "s **is at least as plausible for** i **as** t".

Preorder: *reflexive* and *transitive*.

Locally connected:

$$s \leq_i t \wedge s \leq_i w \;\Rightarrow\; t \leq_i w \vee w \leq_i t,$$

$$t \leq_i s \wedge w \leq_i s \;\Rightarrow\; t \leq_i w \vee w \leq_i t.$$

As a consequence, the **comparability relation** $\overset{i}{\sim}$, given by

$$s \overset{i}{\sim} t \text{ iff either } s \leq_i t \text{ or } t \leq_i s,$$

is an *equivalence relation*, called i**'s epistemic indistinguishability**. This induces an **information partition** for each agent i.

Strict Order and Equi-plausibility

We also consider the *"strict" plausibility* relation:

$$s <_i t \text{ iff: } s \leq_i t \text{ but } t \not\leq_i s$$

$$s \sim_i t \text{ iff either } s \leq_i t \text{ or } t \leq_i s.$$

Equi-plausibility is the equivalence relation $\cong_a$ induced by the preorder $\leq_i$:

$$s \simeq_i t \text{ iff: both } s \leq_i t \text{ and } t \leq_i s$$

When using the R_i notation for the relation $\leq_i$, the correspond strict version, indistinguishability and equi-plausibility relations are denoted by $R_i^{<}$, $R_i^{\sim}$, $R_i^{\simeq}$.

Plausibility Models

A **(finite, pointed) plausibility model** :

- a finite plausibility frame $\mathcal{M} =< I, W, (\leq_i)_{i\in I} >$,
- a *designated world* $s_0 \in W$, called the "**real world**",
- a **valuation map**, assigning to each atomic sentence p (in a given set At of atomic sentences) some set $\|p\| \subseteq W$.

The valuation tells us which "ontic" (non-epistemic, objective) "facts" hold at each world.

Knowledge

A sentence φ is **known by agent** i at state s if it is **true at all states in the same information cell as** s; i.e. in all the worlds in the set

$$\{t \in S : t \overset{i}{\sim} s\}.$$

(Conditional) Belief

A sentence φ is **believed by agent** i at state s if φ **is true in all the "most plausible" worlds in s's information cell**; i.e. in all "minimal" states in the set

$$\{t \in S : t \leq_i w \text{ for all } w \overset{i}{\sim} s\}.$$

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 in s's information cell**; i.e. in all the states in the set

$$\{t \in P : t \leq_i w \text{ for all } w \in P \text{ such that } w \overset{i}{\sim} s\}.$$

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 φ was the case, I will believe that ψ was the case.

Strong Belief

A sentence φ is **strongly believed by agent** i at state s if the following two conditions hold

1. φ **is consistent with the agent's knowledge** at s:

$$\exists w \overset{i}{\sim} s \text{ such that } w \models \varphi$$

2. **within each information cell, all** φ**-worlds are strictly more plausible than all non-**φ**-worlds**:

$$t <_i t' \text{ for all } t \overset{i}{\sim} t' \text{ such that } t \models \varphi \text{ and } t' \not\models \varphi.$$

It is easy to see that **strong belief implies belief**.

Strong Belief is Believed Until Proven Wrong

Actually, strong belief is so strong that **it will never be given up except when one learns information that contradicts it**!

More precisely:

φ is **strongly believed** iff φ **is believed and is also conditionally believed given any new evidence (truthful or not) EXCEPT if the new information is known to contradict** φ; i.e. if:

1. $B\varphi$ holds, and

2. $B^{\theta}\varphi$ holds for every θ such that $\neg K(\theta \Rightarrow \neg\varphi)$.

Example 0: Prof Winestein

Professor Albert Winestein feels that he is a genius. He **knows** that there are only two possible explanations for this feeling: either he *is* a genius or he's drunk. He doesn't feel drunk, so **he believes that he is a sober genius**.

However, **if** he realized that he's drunk, he'd think that his genius feeling was just the effect of the drink; i.e. **after learning he is drunk** he'd come to **believe that he was just a drunk non-genius.**

In reality though, he is **both drunk and a genius**.

A Model for Example 0

The **actual** world is (D, G). Albert considers $(D, \neg G)$ as being **more plausible** than (D, G), and $(\neg D, G)$ as **more plausible** than $(D, \neg G)$. But he can distinguish all these worlds from $(\neg D, \neg G)$, since (in the real world) he **knows** (K) he's either drunk or a genius.

$$\boxed{\neg D, \neg G} \qquad \boxed{D, G} \xrightarrow{a} \boxed{D, \neg G} \xrightarrow{a} \boxed{\neg D, G}$$

Drawing Convention: We use *labeled arrows for converse plausibility relations* $\geq_a$, going from less plausible to more plausible worlds, but *we skip loops and composed arrows* (since $\geq_a$ are reflexive and transitive).

True Belief is not Knowledge

At the real world (D, G), we can check that **Albert believes he's a genius**

$$(D, G) \models B_a G,$$

but **he doesn't "know" he's a genius**:

$$(D, G) \models \neg K_a G.$$

However, Albert knows that he's either drunk or a genius:

$$(D, G) \models K_a (D \vee G)$$

Mary Curry

Albert Winestein's best friend is Prof. Mary Curry (not to be confused with Marie Curie!).

She's **pretty sure that Albert is drunk**: she can see this with her very own eyes. All the usual signs are there!

She's **completely indifferent with respect to Albert's genius**: she considers the possibility of genius and the one of non-genius as equally plausible.

However, having a philosophical mind, Mary Curry **is aware of the possibility that the testimony of her eyes may in principle be wrong**: it is in principle possible that Albert is not drunk, despite the presence of the usual symptoms.

$$
\begin{tikzcd}
\boxed{\neg D, \neg G} \arrow[r, leftrightarrow, "m"] & \boxed{\neg D, G} \arrow[r, "m"] & \boxed{D, G} \arrow[r, leftrightarrow, "m"] & \boxed{D, \neg G}
\end{tikzcd}
$$

Marry "knows" though she doesn't Know

In the *real world* (D, G), Marry **truthfully believes that Albert is a drunk genius**:

$$(D, G) \models B_m D \wedge B_m G$$

But *she doesn't know these things*:

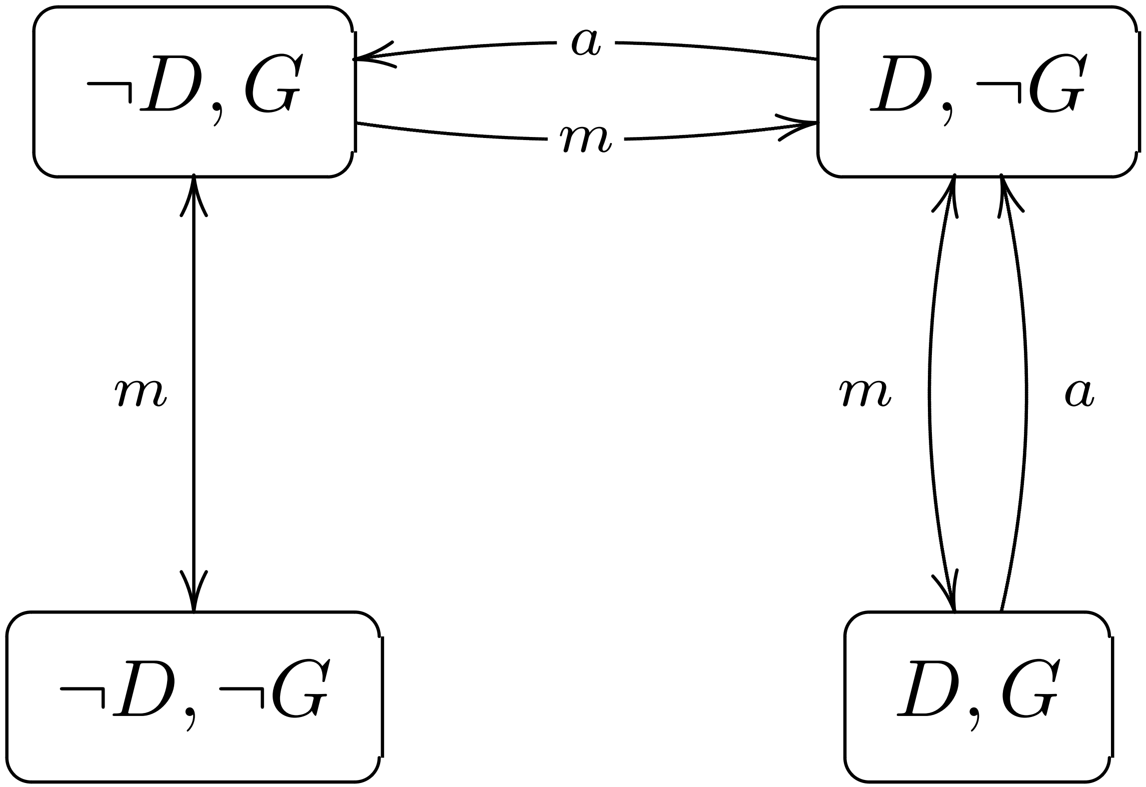
$$(D, G) \models \neg K_m D \wedge \neg K_m G.$$

However, *she strongly believes them*:

$$(D, G) \models Sb_m D \wedge Sb_m G.$$

A Multi-Agent Model S

Putting together Marry's order with Albert's order, we obtain a multi-agent plausibility model **S** for the whole epistemic situation:

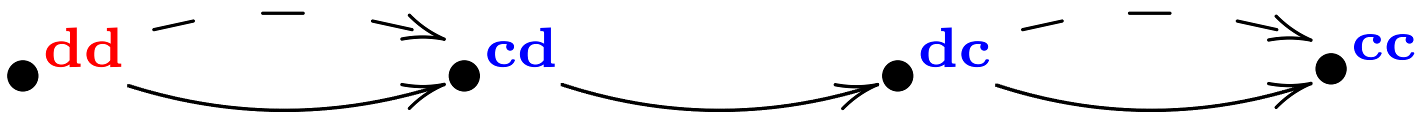


Example 1

Two children played with mud, and they **both have mud in their hair**. They **stand in line**, with child 1 looking at the back of child 2. So 1 *can see if* 2*'s hair is dirty or not, but not the other way around.* (And no child can see himself.)

Let's assume that (it is common knowledge that) each of them thinks that *it is more plausible that he is clean than that he is dirty.* Also, (it is common knowledge that) child 2 thinks that *it is more plausible that he himself (child 2) is clean than that child* 1 *is clean.*

Plausibility Model



Arrows: **converse plausibility** relations $\geq_i$ (going from less plausible to more plausible), but **we skip all the loops and composed arrows**.

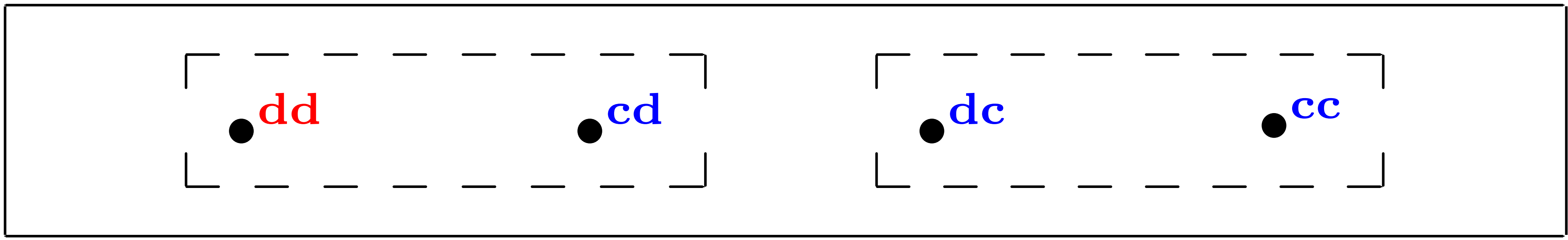
Dotted arrows: child 1.

Continuous arrows: child 2.

RED: the real world.

Information Partitions

From this, we can extract the information partitions:

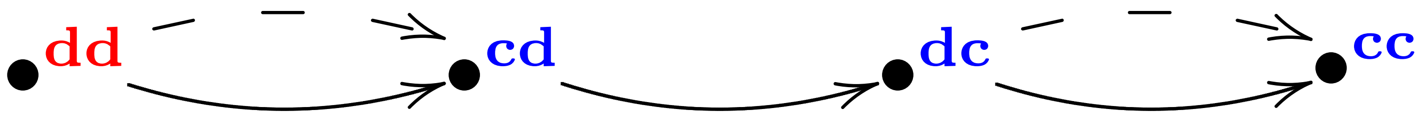


Squares: children's information cells.

Dotted squares: child 1. *Continuous* squares: child 2.

Examples

In the real state (d, d) of our model



child 2 (continuous line) has a *strong belief that he's clean*.

He also *believes that child 1 is clean*. But *this is NOT a strong belief*.

Modelling 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).*

An **upgrade** is a **model-changing operation** α, taking *any model* $\mathcal{M} = (I, W\ \sim_i, \leq_i, \|\cdot\|, s_0)$, and returning a *new model* $\alpha(\mathcal{M}') = (I, W', \sim_i', \leq_i', \|\cdot\| \cap W', s_0)$, with:

- set of states: some **subset** $W' \subseteq W$,
- valuation: the **restriction** to W' of the old valuation,
- **the same real world** s_0 as the old model (but **possibly different relations**).

Examples of Upgrades

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

(2) Lexicographic upgrade $\Uparrow\varphi$:
all φ-worlds become "better" (more plausible) than all $\neg\varphi$-worlds in the same cell, and *within the two zones, the old relations are kept.*

(3) Conservative upgrade $\uparrow\varphi$:
the "best" φ-worlds become better than all other worlds in the same cell; *all else stays the same.*

Joint Upgrades and Single Upgrades

These operations can be applied *simultaneously to all the relations*, obtaining **joint upgrades**, or can be applied only to a single agent's relations (keeping the others unchanged), obtaining **single upgrades**.

EXPLANATION: The three types of upgrades correspond to **three different attitudes** of the learners towards **the reliability of the source** (of the new information):

Explanation continued

- **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 if at any later moment some contradiction forces some belief revision, the first thing to go is the trust in this source!

Comparison with probabilistic conditioning

Updates= the qualitative, multi-agent analogue of Bayesian conditioning.

Lexicographic Upgrade = the qualitative, multi-agent analogue of Jeffrey conditioning (with a binary partition $\{\varphi, \neg\varphi\}$) .

Conservative Upgrade = the "minimal" revision of the old doxastic structure that is compatible with the new information.

The last one is the favorite choice of many people in Belief Revision.

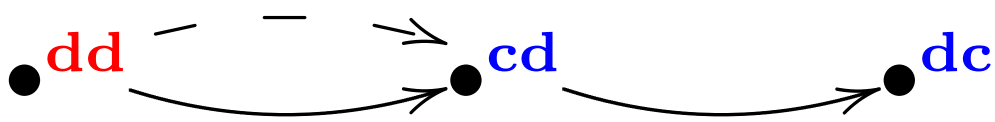
EXAMPLE 2: A Joint Update

The Father announces:

"*At least one of you is dirty*".

We take the Father to be an **infallible** source.

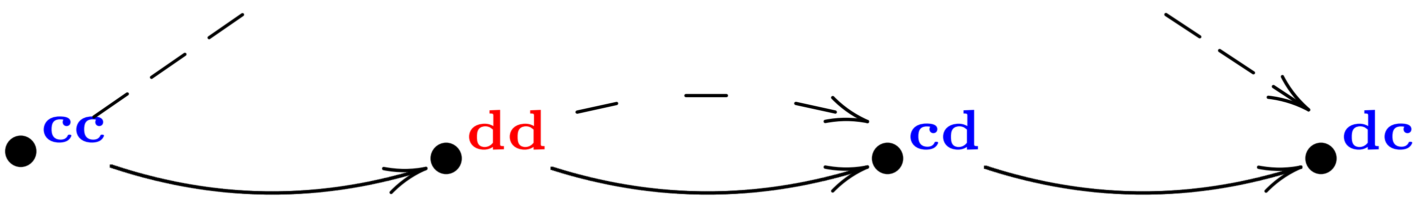
So this is an **update** $!(d_1 \vee d_2)$, yielding the updated model:



EXAMPLE 3: Joint Lexicographic Update

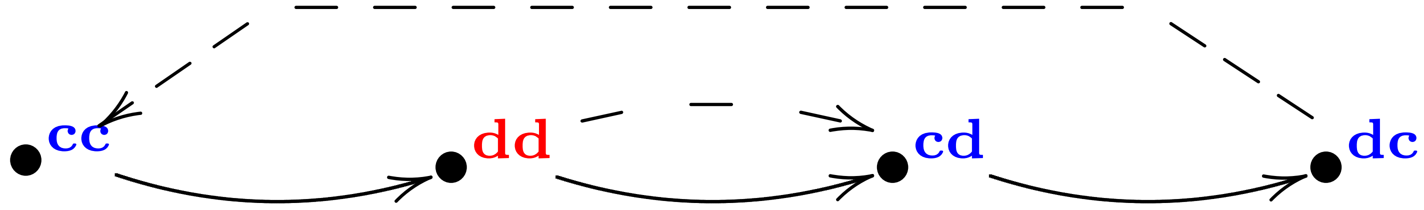
Alternatively, an older sister announces: "*At least one of you is dirty*". She is a **highly trusted source, though not infallible**: once in a while, she may just make up "dirty" stories.

This lexicographic upgrade yields:



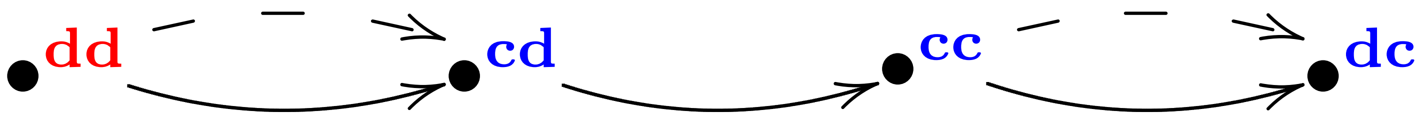
EXAMPLE 4: Single Lexicographic Update

Alternatively, suppose that it is *common knowledge* that **only child 2 highly trusts the sister**; but that **child 1 always disregards her announcements**, assuming they are just made-up stories. So sister's announcement will induce a *single upgrade* by child 2:



EXAMPLE 5: Joint Conservative Upgrade

Alternatively, children hear a **rumor** that at least one of them is dirty. It is **barely believable**, so they perform a *joint conservative upgrade*:



Redundancy and Fixed Points

A joint upgrade $\dagger\varphi$ is **redundant on a pointed model S with respect to a group of agents** G iff the upgraded model is G-*bisimilar* to the original one:

$$\dagger\varphi(\mathbf{S}) \simeq_G \mathbf{S}.$$

This means that all the doxastic attitudes (knowledge, conditional beliefs, strong beliefs, common knowlege etc) *stay the same within group* G after the upgrade.

A model $\mathbf{S}$ is a **fixed point** of $\dagger\varphi$ if $\dagger\varphi$ is *redundant on* $\mathbf{S}$ *with respect to the group* $\mathcal{A}$ *of all agents*: $\mathbf{S} \simeq \dagger\varphi(\mathbf{S})$.

(Non-)Informativity and Sincerity

An upgrade $\dagger\varphi$ is **informative** *(on* **S***) to group* G if it is not redundant with respect to G.

Intuitively, an announcement is "sincere" when it agrees with the speaker's prior epistemic state: *accepting the announcement should not change the speaker's own state.*

DEFINITION: A **public announcement** $\dagger\varphi$ **by agent** a is said to be **sincere** if *it doesn't change agent* a*'s own plausibility structure*; i.e. if it is *redundant (=non-informative) to the agent* a (=redundant with respect to the singleton group $\{a\}$).

Logical Characterizations of Sincerity

1. **An infallible public announcement** (inducing a joint update) $!\varphi$ **by an agent** a is **sincere** iff a **knows** φ.

2. **A (fallible but) strongly persuasive announcement** (inducing a joint radical upgrade) $\Uparrow \varphi$ **by an agent** a is **sincere** iff a **strongly believes** φ.

3. **A "barely trusted" announcement** (inducing a joint conservative upgrade) $\uparrow \varphi$ **by an agent** a is **sincere** iff a **(simply) believes** φ.

Iterating Upgrades

To study iterated belief revision, consider a **finite model** $\mathcal{M}_0$, and an **(infinite) sequence of (joint or single) upgrades**

$$\alpha_0, \alpha_1, \ldots, \alpha_n, \ldots$$

In particular, these can be updates

$!\varphi_0, !\varphi_1, \ldots, !\varphi_n, \ldots$

or conservative upgrades

$\uparrow \varphi_0, \uparrow \varphi_1, \ldots, \uparrow \varphi_n, \ldots$

or lexicographic upgrades

$\Uparrow \varphi_0, \Uparrow \varphi_1, \ldots, \Uparrow \varphi_n, \ldots.$

The iteration leads to **an infinite succession of upgraded models**

$$\mathcal{M}_0, \mathcal{M}_1, \ldots, \mathcal{M}_n, \ldots$$

defined by:

$$\mathcal{M}_{n+1} = \alpha_n(\mathcal{M}_n).$$

PROBLEM 1:

CONVERGENCE (OR CYCLES?) OF ITERATED “LEARNING”

Iterated Updates Always Stabilize

OBSERVATION: For every initial finite model $\mathcal{M}_0$, every infinite sequence of updates

$$!\varphi_0, \ldots, !\varphi_n, \ldots$$

stabilizes the model after finitely many steps.

I.e. there exists n such that

$$\mathcal{M}_n = \mathcal{M}_m \text{ for all } m \geq n.$$

The reason is this is a *deflationary* process: the model keeps contracting until it eventually must reach a fixed point.

Iterated Upgrades Do Not Necessarily Stabilize!

This is **NOT** the case for arbitrary upgrades.

First, it is obvious that, if we allow for **false** upgrades, the revision may oscilate forever: the sequence

$$\Uparrow p, \Uparrow \neg p, \Uparrow p, \Uparrow \neg p, \ldots$$

will forever **keep reverting back and forth the order between the p-worlds and the non-p -worlds.**

"Learning": Tracking the Truth

This is to be expected: such an "undirected" revision with mutually inconsistent pieces of "information" is not real learning.

As Nozick put it, **"knowledge" and "learning" have to do with tracking the truth** (in the real world).

But, surprisingly enough, **we may still get into an infinite belief-revision cycle, even if the revision is "directed" towards the real world**: i.e. even if we allow only upgrades that are always **truthful**!

Iterated Learning can produce Doxastic Cycles

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

In Example 2, in the situation after the Father's announcement

$$\begin{tikzcd}[column sep=large] \bullet^{\mathbf{dd}} \arrow[r, bend left, dashed] \arrow[r, bend right] & \bullet^{\mathbf{cd}} \arrow[r, bend right] & \bullet^{\mathbf{dc}} \end{tikzcd}$$

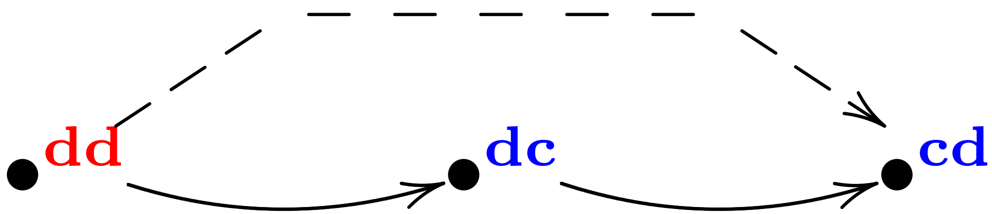
, child 1 announces the following (true) sentence φ:

$$\mathbf{c_1} \vee \mathbf{c_2} \Rightarrow (\mathbf{B_2 c_2} \Leftrightarrow \neg \mathbf{c_2})$$

"If at least one of us is clean, then whatever you believe about whether you're clean or not is wrong."

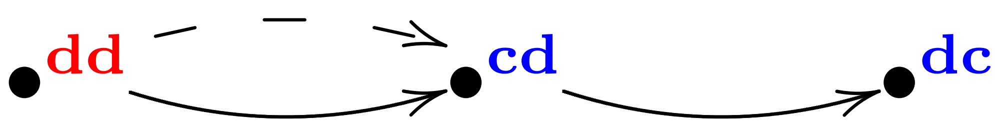
Infinite Oscillations by Truthful Upgrades

Sentence φ is *true in the real world dd*, as well as in *cd*, but not in *dc*. Let's suppose that child 2 only *barely trusts* child 1: so this is a **truthful (single) conservative upgrade** by child 2, resulting in



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

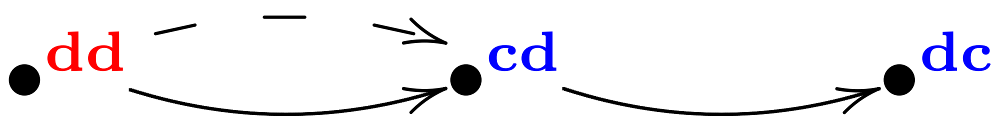
However, if child 1 announces φ again and child 2 **conservatively upgrades again** with this true information φ, he will promote dc again as his most plausible state, **reverting to the original model**:



Moreover, **not only the whole model (the plausibility order) keeps changing forever**, but child 2's (simple, un-conditional) **beliefs keep oscillating forever** (between cd and dc)!

Iterating Truthful Lexicographic Upgrades

Consider again the model in Example 2, after Father's announcement:



What happens if child 2 **strongly trusts** child 1, so that whatever he says induces a **lexicographic upgrade**? Can child 1 still induce infinite oscillations by truthful announcements?

Iterating Truthful Lexicographic Upgrades

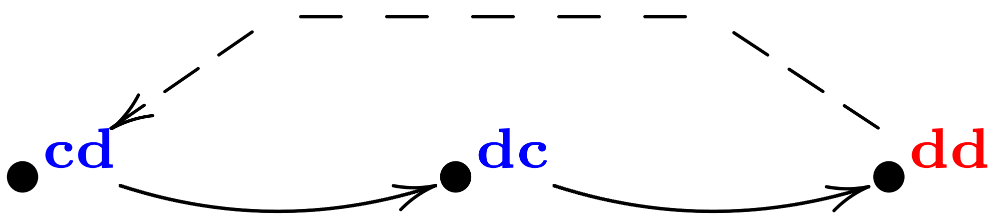
Let child 1 announce to child 2:

"*If the Father announces that at least one of us is clean, then after that your belief about whether you're clean or not would be wrong.*"

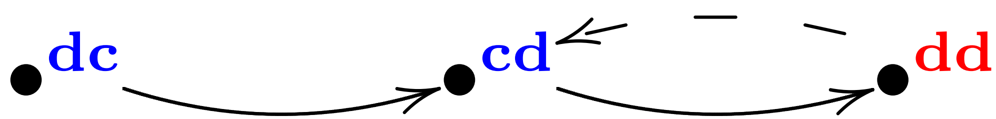
$$(\mathbf{c_1} \vee \mathbf{c_2}) \Rightarrow (\mathbf{B_2^{c_1 \vee c_2} c_2} \Leftrightarrow \neg \mathbf{c_2})$$

φ' is true in the real world *dd* and in *cd* but not in *dc*, so a **truthful lexicographic upgrade** will give us:

The same sentence is again true in (the real world) dd and in dc, 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 the previous model (above)!

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 point!

However, *unlike in the conservative upgrade example*, **in this example the simple (unconditional) beliefs eventually stabilize**: *from some moment onwards, the children correctly believe that the real world is dd, and they will never lose this belief again*!

This is a *symptom* of a more general phenomenon:

Beliefs Stabilize in Iterated Lexicographic Upgrades

THEOREM:

In any infinite sequence of truthful lexicographic upgrades $\{\Uparrow \varphi_i\}_i$ on an initial (finite) model $\mathcal{M}_0$, 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 and knowledge 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*.

Extension and Proof

This Theorem is our main “positive” result. It can be generalized to a larger class of upgrades (that can be thought of as multi-agent analogues of Jeffrey conditioning with arbitrary, not just binary, partitions).

The proof is not long, but not completely trivial. It uses a combinatorial property underlying an old “magic card” trick.

Conclusions

Iterated revision with truthful higher-level information can be highly non-trivial.

The long-term behavior **depends both on the type of sentence that is learnt**, and on **the specific way in which the "learning" takes place** (in particular, the *reliability of the source*).

Conclusions – continued

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

But, when iterating truthful LEXICOGRAPHIC upgrades, simple (non-conditional) beliefs DO converge to some stable belief.

Truthful CONSERVATIVE upgrades do NOT have this last property: simple beliefs may oscillate forever during iterated conservative upgrades with true information.

PROBLEM 2:

REALIZING DOXASTIC MERGE BY ITERATED "SHARING"

Sharing=Sincere Communication

By **"sharing" information within a group** G we mean **sincere public announcements** $\dagger\varphi$ (of any of the three types $!\varphi$, $\Uparrow\varphi$ or $\uparrow\varphi$) **by any of the members of the group** G.

NOTE: We no longer require truthfulness (but only sincerity).

Preference Merge and Information Merge

In Social Choice Theory, the main issue is how to *merge* the agent's individual preferences. A **merge operation for a group** G is a function $\bigodot$, taking preference relations $\{R_i\}_{i \in G}$ into a *"group preference" relation* $\bigodot_{i \in I} R_i$ (on the same state space).

So the problem is to find a *"natural" merge operation* (subject to various *fairness conditions*), for merging the agents' preference relations. Depending on the conditions, one can obtain either an **Impossibility Theorem** (*Arrow* 1950) or a **classification of the possible types of merge** operations (*Andreka, Ryan & Schobbens* 2002).

Belief Merge and Information Merge

If we want to *merge the agents'* **beliefs** B_i, so that we get a notion of "group belief", then it is enough to merge the belief relations $\rightarrow_i$.

If we want to merge the agents' **knowledge ("hard information")** K_i, then it is enough to merge the epistemic indistinguishability relations $\stackrel{i}{\sim}$.

If we want to merge the agents' **soft information (="strong beliefs" or "conditional beliefs")** Sb_i (or, equivalently, to merge all their *conditional beliefs*), then we have to merge the plausibility relations $\leq_i$.

Merge by Intersection

The so-called **parallel merge** (or "**merge by intersection**") simply takes the merged relation to be

$$\bigcap_{i \in G} R_i.$$

In the case of two agents, it takes:

$$R_a \bigodot R_B := R_a \cap R_b$$

This could be thought of as a *"democratic" form of preference merge.*

Distributed Knowledge is Parallel Merge

This form of merge is particularly suited for "knowledge" K: since this type of knowledge is absolutely certain, there is no danger of inconsistency. The agents can pool their information in a *completely symmetric manner*, *accepting the other's bits without reservations*.

In fact, **parallel merge of the agents' irrevocable knowledge gives us the standard concept of "distributed knowledge"** DK:

$$DK_G P = [\bigcap_{i \in G} \overset{i}{\sim}] P.$$

Lexicographic Merge

In lexicographic merge, a "priority order" is given on agents, to model the group's hierarchy. For two agents a, b, the "**lexicographic merge**" $R_{a/b}$ **gives priority to agent** a **over** b:

The strict preference of a **is adopted by the group**; **if** a **is indifferent, then** b**'s preference** (or lack of preference) **is adopted**; finally, a**-incomparability gives group incomparability**. Formally:

$$R_{a/b} := R_a^{>} \cup (R_a^{\simeq} \cap R_b) = R_a^{>} \cup (R_a \cap R_b) = R_a \cap (R_a^{>} \cup R_b).$$

Lexicographic merge of soft information

This form of merge is particularly suited for "soft information", given by either *strong beliefs* Sb or *conditional beliefs* B, **in the absence of any hard information**: since soft information is not fully reliable, some "screening" must be applied (and so some hierarchy must be enforced) to ensure consistency of merge.

Relative Priority Merge

Note that, **in lexicographic merge, the first agent's priority is "absolute"**. But **in the presence of hard information**, the lexicographic merge of soft information must be **modified**, by **first pooling together all the hard information and then using it to restrict the lexicographic merge of soft information**. This leads us to a "more democratic" **combination of Merge by Intersection and Lexicographic Merge** , called "**(relative) priority merge**" $R_{a\otimes b}$:

$$R_{a\otimes b} := (R_a \cap R_b^{\sim}) \cup (R_a^{\simeq} \cap R_b) = R_a \cap R_b^{\sim} \cap (R_a^{>} \cup R_b).$$

Essentially, this means that **both agents have a "veto" with respect to group incomparability**: the group can only compare options that **both** agents can compare; **and whenever the group can compare two options, everything goes on as in the lexicographic merge**: agent a's strong preferences are adopted, while b's preferences are adopted only when a is indifferent.

Example: merging Marry's beliefs with Albert's

If we give priority to Marry (the more sober of the two!), the relative priority merge $R_{m\otimes a}$ of Marry's and Albert's original plausibility orders

$$\begin{tikzcd} \boxed{\neg D, \neg G} \arrow[r, leftrightarrow, "m"] & \boxed{\neg D, G} \arrow[r, "m"] & \boxed{D, G} \arrow[r, leftrightarrow, "m"] & \boxed{D, \neg G} \end{tikzcd}$$

$$\begin{tikzcd} \boxed{\neg D, \neg G} & \boxed{D, G} \arrow[r, "a"] & \boxed{D, \neg G} \arrow[r, "a"] & \boxed{\neg D, G} \end{tikzcd}$$

gives us:

$$\begin{tikzcd} \boxed{\neg D, \neg G} & \boxed{\neg D, G} \arrow[r] & \boxed{D, G} \arrow[r] & \boxed{D, \neg G} \end{tikzcd}$$

Merging Beliefs is Not a Sure Way to the Truth

If instead we give priority to Albert, we simply obtain Albert's order as our "merge":

$$R_{a\otimes m} = R_a .$$

NOTE: in BOTH cases, some of the resulting *joint ("merged") beliefs are wrong*: when giving priority to Marry, both agents end up believing that Albert is not a genius; while if we give priority to Albert, they both end up believing that Albert is sober!

In fact, **no type of hierarchic belief merge is a warranty of veracity**!

4. "Realizing" Preference Merge Dynamically

Intuitively, the purpose of "preference merge" $\bigodot_{i \in G} R_i$ is to achieve a state in which the G-agents' preference relations are "**merged**" accordingly, i.e. *to perform a sequence* π *of upgrades, transforming the initial model* $(S, R_i)_{inG}$ into a model $(S, R'_i)_{i \in G}$ such that

$$R'_j = \bigodot_{i \in G} R_i$$

for all $j \in G$.

Let us call this a **"realization" of the merge operation** $\bigodot$.

Merging by Public Communication

For each of the above types of public communication $(!, \Uparrow, \uparrow)$, we can ask **which merge operations are realizable by iterated sincere announcements of that type**.

The answer **will depend on the constraints** (e.g. *transitivity, connectedness* etc.) assumed on the agents' epistemic, doxastic or plausibility relations. So it matters whether we are looking at merging hard information K, soft information Sb or beliefs B.

Realizing Distributed Knowledge

In the case of **knowledge**, it is easy to design an algorithm to realize it, **as the parallel merge of agents' knowledge**, operation by a *sequence of joint updates*, as follows: **in no particular order, the agents have to publicly and sincerely announce (in an infallible manner) "all that they know"** .

More precisely, for each set of states $P \subseteq S$ such that P is *known to a given agent a*, an update $!P$ is performed. This essentially is the algorithm in van Benthem's paper "One is a Lonely Number".

The Algorithm

Formally, **the algorithm for realizing distributed knowledge within group** G requires the following protocol (=sequence of announcements):

$$\pi := \prod_{i \in G} \prod \{!P : P \subseteq S \text{ such that } s \models K_a P\}$$

(where $\prod$ is sequential composition of a sequence of actions). The order of the agents in the first $\prod_i$ and the order in which the announcements are made by each agent (in the second $\prod$) are arbitrary.

It is easy to see that after this, we indeed obtain:

$$\overset{j}{\sim}{}' = \bigcap_{i \in G} \overset{i}{\sim}$$

for all $j \in G$. So **distributed knowledge is converted into common knowledge**:

$$s \models DK_G P \leftrightarrow [\pi] Ck_G P.$$

Order-independence

As mentioned, the **order** in which the agents make each of their announcements (and even the **order of the speakers**) **doesn't actually matter**.

The announcements may even be **interleaving**: if the initial model is finite, then **any** "public" dialogue, with a announcing some facts she irrevocably knows, b answering, a announcing some new facts she knows etc., will converge to the realization of distributed knowledge, as long as the agents keep announcing *new things* (i.e. that are not already common knowledge).

Realizing Priority Merge

We can **realize the priority merge** $\bigotimes_i \leq_i$ of soft information **by joint radical upgrades**, via an algorithm very similar to the one for distributed knowledge.

Essentially, the agents are asked to **publicly and sincerely announce (via radical upgrades) "all that they strongly believe".**

Order-dependence

The main difference is that **now the speakers' order matters**!

To realize priority merge, the agents that have "priority" in the merge has to be given priority in the protocol.

A lower-priority agent will be permitted to speak ONLY after the higher-priority agents finished announcing "ALL that they strongly believe".

No interruptions, please!

Be Persuasive!

Note that *simply announcing that they believe it, or that they strongly believe it, won't do*: this will not in general be enough to achieve preference merge (or even simple belief merge!).

Being informed of another's beliefs is not enough to convince you of their truth.

What is needed for belief merge is that the agents try **to be persuasive**: *to "convert" the other to their own beliefs* by **persuasively announcing** φ **when they just strongly believe** φ.

The Algorithm

Formally, the protocol π' for realizing priority merge of plausibility relations $\{\leq_i\}_{i\in G}$ is the following:

$$\pi' := \prod_{(i_1,\ldots i_k)\in G} \prod \{\Uparrow P : P \subseteq S \text{ such that } s \models Sb_i P\}.$$

Here, *the order* $(i_1, \ldots, i_k)$ *of the agents in the first* $\prod_i$ *is the priority order in the desired merge* (while the order in which the announcements are made by each agent in the second $\prod$ is still arbitrary).

It is easy to see that after this, we indeed obtain

$$\leq'_j = \bigotimes_{i \in G} \leq_i$$

for all $j \in G$.

The Rules of the Game

The "rules of the game" in the above algorithm are:

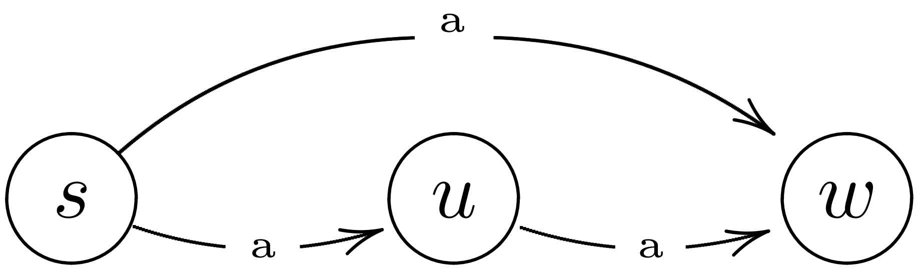
(1) **"sincerity"**: agents announce that they "know" *only things that they believe they "know"*;

(2) **"exhaustiveness"**: the algorithm stops only when the agents have announced ' *'all" they (think they) "know"*;

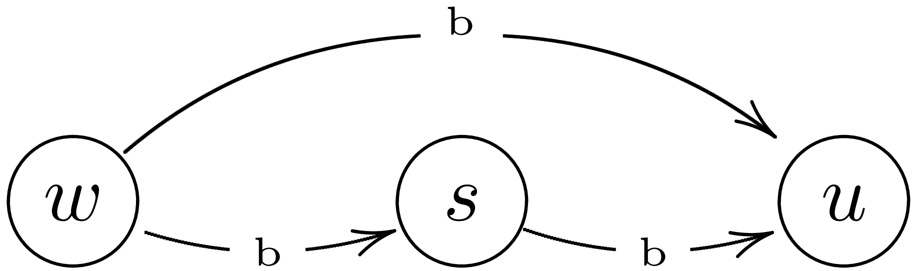
(3) **"priority order"**, strictly enforced: the agents with higher priority have to*finish announcing all they (think they) "know" before* agents with lower priority can speak.

Order-dependence: counterexample

The priority merge of the ordering



with the ordering

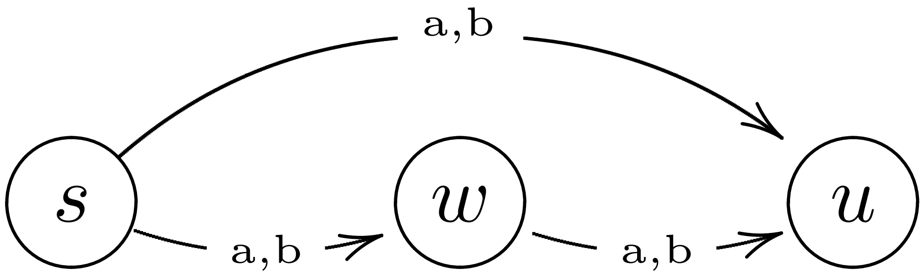


is equal to either of the two orders (depending on which agent has priority). But...

... suppose we have the following public dialogue

$$\Uparrow \Box_b u \cdot \Uparrow \Box_a (u \vee w)$$

This *respects the "sincerity" rule* of our algorithm. It also *respects in a sense the "exhaustiveness" rule*, since the agents only stop when they shared everything. But it *doesn't respect the "order" rule*: b lets a answer before she finishes giving him all the information she has. The resulting order is neither of two priority merges:



Example

Recall the initial Marry & Albert orders:

$$\begin{tikzcd}
\boxed{\neg D, \neg G} \arrow[r, leftrightarrow, "m"] & \boxed{\neg D, G} \arrow[r, "m"] & \boxed{D, G} \arrow[r, leftrightarrow, "m"] & \boxed{D, \neg G}
\end{tikzcd}$$

$$\begin{tikzcd}
\boxed{\neg D, \neg G} & \boxed{D, G} \arrow[r, "a"] & \boxed{D, \neg G} \arrow[r, "a"] & \boxed{\neg D, G}
\end{tikzcd}$$

The algorithm to realize the relative priority merge $R_{m \otimes a}$:

$$\Uparrow K_a(D \vee G); \Uparrow \Box_m D; \Uparrow \Box_a \neg G$$

The first upgrade is of the required form, despite appearances, because of the equivalence:

$$K_a P = \Box_a K_a P$$

Expertise-guided Priority

But recall that in this case **priority merge does NOT lead to entirely correct beliefs**. The only way to recover "the Truth" is *to give each of the agents its due*, **by considering each of them as "expert" in one of the two issues** (scientific genius and drunkness): *let Albert* (as a Professor of Physics) *decide the issue of "genius"*, and *let Marry* (as a Professor of Cooking) *decide the issue of drunkness*. In addition, *let Albert speak first* (and of course let him convey his hard information as well!). The ensuing algorithm is:

$$\Uparrow K_a(D \vee G); \Uparrow \Box_a G; \Uparrow \Box_m D$$

The Way to the Truth

This results in the merged order:

$$\boxed{\neg D, \neg G} \qquad \boxed{D, \neg G} \longrightarrow \boxed{\neg D, G} \longrightarrow \boxed{D, G}$$

So now *the resulting joint beliefs are all correct*!

The lesson is that, **by giving each of the agents relative priority only with respect to the issues over which they have relevant expertise, the group MAY be able to recover (or at least approach) the Truth!**

But... the Order still Matters!

The order still matters: if we assign the same expertise-based priorities, but we allow Marry speak first, we obtain instead the **same merged order as the lexicographic merge** $R_{m\otimes a}$, i.e.

$$\boxed{\neg D, \neg G} \qquad \boxed{\neg D, G} \longrightarrow \boxed{D, G} \longrightarrow \boxed{D, \neg G}$$

leading to the incorrect belief in non-genius!

The reason, again, is that Albert's "expert" opinion on genius is easy to manipulate, because it is an *unsafe belief*.

The Power of Agendas

Things get even worse if we **mix up the relevant expertise**, by letting Albert decide on drunkness and Marry decide on genius!

All this illustrates the **important role of the person who "sets the agenda"**: the "Judge" who assigns **priorities to witnesses' stands** and determines the **witnesses' relevant field of expertise**. Or the "Speaker of the House", who determines the **order of the speakers** as well as the **the issues** to be discussed and **the relative priority of each issue**.

Open Problem

So, depending on the "agenda", soft announcements can realize **a whole plethora of merge operations**. Nevertheless, **NOT everything goes**: the requirements imposed on the plausibility relations generally pose restrictions to which kinds of merge are realizable.

OPEN QUESTION: **characterize the class of merge operations realizable by lexicographic upgrades**.