Dynamic Logic for Interactive Belief Revision

Sonja Smets, ILLC, Amsterdam

Financial Support Acknowledgement:

European Research Council erc

Netherlands Organisation for Scientific Research

PLAN OF THIS COURSE

1. **Standard ("Hard") Dynamic-Epistemic Logics**: Logics of knowledge and belief. Epistemic models. Public and private announcements. Event models. The Product Update.

2. **Belief Revision**: Plausibility Models. Conditional belief. Belief Upgrades. Doxastic event models and the Action-Priority Rule.

3. **Further Topics**: Iterated Belief Revision. Belief Merge. Collective Learning. Informational Cascades.

Relevant Textbooks and Surveys

- A. Baltag, H. P. van Ditmarsch and L.S. Moss, "Epistemic logic and information update", in *Handbook of Philosophy of Information*, Elsevier, 2008.
- A. Baltag and S. Smets, "A Qualitative Theory of Dynamic Interactive Belief Revision", in G. Bonanno, W. van der Hoek, M. Wooldridge (eds.), *Texts in Logic and Games*, Vol 3, pp.9-58, Amsterdam Univ Press, 2008.
- J. van Benthem, **Modal Logic for Open Minds**, CSLI Publications, Stanford, 2011.
- J. van Benthem, **Logical Dynamics of Information and Interaction**, Cambridge Univ Press, 2011.
- H. P. van Ditmarsch, W. van der Hoek and B. Kooi, **Dynamic Epistemic Logic**, Springer, 2007.
- R. Fagin, J.Y. Halpern, Y. Moses and M.Y. Vardi, **Reasoning about Knowledge**, MIT Press, Cambridge MA 1995.

- Research papers of van Benthem, Baltag and Smets (see their personal websites)

1.1 Epistemic Puzzles: Muddy Children

Suppose there are 4 children, all of them being good logicians, exactly 3 of them having dirty faces. *Each can see the faces of the others, but doesn't see his/her own face.*

The father publicly announces:

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

Then the father does another paradoxical thing: *starts repeating over and over the same question* "**Do you know if you are dirty or not, and if so, which of the two?**"

After each question, the children have to *answer publicly, sincerely and simultaneously, based only on their knowledge, without taking any guesses*. No other communication is allowed and nobody can lie.

One can show that, after 2 rounds of questions and answers, **all the dirty children will come to know they are dirty**! So they give this answer in the 3rd round, after which **the clean child also comes to knows she's clean**, giving the correct answer at the 4th round.

Muddy Children Puzzle continued

First Question: *What's the point of the father's first announcement ("At least one of you is dirty")?*

Apparently, this message is not informative to any of the children: the statement was already known to everybody! But the puzzle wouldn't work without it: in fact this announcement adds information to the system! The children implicitly learn some new fact, namely the fact that what each of them used to know *in private* is now *public knowledge*.

Second Question: *What's the point of the father's repeated questions?*

If the father knows that his children are good logicians, then at each step the father knows already the answer to his question,

before even asking it! However, the puzzle wouldn't work without these questions. In a way, it seems the father's questions are *"abnormal"*, in that they don't actually aim at filling a gap in father's knowledge; but instead they are part of *a Socratic strategy of teaching-through-questions*.

Third Question: *How can the children's statements of ignorance lead them to knowledge?*

Puzzle no 2: Sneaky Children

Let us modify the last example a bit.

Suppose the children are somehow rewarded for answering as quickly as possible, but they are punished for incorrect answers; thus they are interested in getting to the correct conclusion as fast as possible.

Suppose also that, **after the second round of questions, two of the dirty children "cheat"** on the others by **secretly announcing each other that they're dirty**, while none of the others suspects this can happen.

Honest Children Always Suffer

One can easily see that the **third dirty child will be totally deceived, coming to the "logical" conclusion that... she is clean**!

So, after giving the wrong answer, she ends up by being punished for her credulity, despite her impeccable logic.

Clean Children Always Go Crazy

What happens to the clean child?

Well, **assuming she doesn't suspect any cheating, she is facing a contradiction**: two of the dirty children answered too quickly, coming to know they're dirty before they were supposed to know!

If the third child simply updates her knowledge monotonically with this new information (and uses classical logic), then she ends up believing everything: **she goes crazy**!

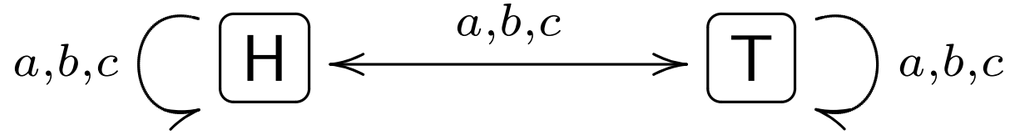
1.2. Epistemic-Doxastic Models and Logics

Epistemic Logic was first formalized by Hintikka (1962), who also sketched the first steps in formalizing doxastic logic.

They were further developed and studied by both philosophers (Parikh, Stalnaker etc.), economists (Aumann) and computer-scientists (Halpern, Vardi, Fagin etc.)

Scenario 1: the concealed coin

Two players a, b and a referee c play a game. In front of everybody, the referee throws a fair coin, catching it in his palm and fully covering it, before anybody (including himself) can see on which side the coin has landed.



Kripke Models

For a set Φ of *facts* and a finite set $\mathcal{A}$ of *agents*, a **Φ-Kripke model** is a triple

$$\mathbf{S} \quad = \quad (S, \xrightarrow{\mathcal{A}}, \|.\|)$$

consisting of

1. a set S of "worlds"
2. a family of binary accessibility relations $\xrightarrow{a} \subseteq S \times S$, one for each agent $a \in \mathcal{A}$
3. and a *valuation* $\|.\| : \Phi \to \mathcal{P}(S)$, assigning to each $p \in \Phi$ a set $\|p\|_{\mathbf{S}}$ of states

The valuation is also called a *truth map*. It is meant to express the *factual content* of a given world, while the arrows $\xrightarrow{A}$ express the agents' uncertainty between various worlds.

A Kripke model is called a **state model** whenever we think of its "worlds" as *possible states*. In this case, the elements $p \in \Phi$ are called *atomic sentences*, being meant to represent **basic "ontic" (non-epistemic) facts**, which may hold or not at a given state.

Satisfaction Relation

Write $s \models_{\mathbf{S}} \varphi$ for the **satisfaction relation**: φ is true at world s in model $\mathbf{S}$. This is defined inductively:

$$s \models_{\mathbf{S}} p \quad \text{iff} \quad s \in \|p\|_{\mathbf{S}}$$

$$s \models_{\mathbf{S}} \neg\varphi \quad \text{iff} \quad s \not\models_{\mathbf{S}} \varphi$$

$$s \models_{\mathbf{S}} \varphi \wedge \psi \quad \text{iff} \quad s \models_{\mathbf{S}} \varphi \text{ and } s \models_{\mathbf{S}} \psi$$

Extending the Truth Map

Equivalently, this allows us to *extend the truth map* $\|\varphi\|_{\mathbf{S}}$ to *all* propositional formulas, by putting:

$$\|\varphi\|_{\mathbf{S}} := \{s \in S : s \models_{\mathbf{S}} \varphi\}.$$

Obviously, this has the property that

$$\|\neg\varphi\|_{\mathbf{S}} = S \setminus \|\varphi\|_{\mathbf{S}},$$

$$\|\varphi \wedge \psi\|_{\mathbf{S}} = \|\varphi\|_{\mathbf{S}} \cap \|\psi\|_{\mathbf{S}},$$

$$\|\varphi \vee \psi\|_{\mathbf{S}} = \|\varphi\|_{\mathbf{S}} \cup \|\psi\|_{\mathbf{S}}.$$

Any *new* propositional operator $A(\varphi_1, \ldots, \varphi_n)$ is *"defined" by extending the truth map* to define $\|A(\varphi_1, \ldots, \varphi_n)\|_{\mathbf{S}}$, i.e. by *giving a defining inductive clause for satisfaction* $s \models A(\varphi_1, \ldots, \varphi_n)$.

Modalities

For every sentence φ, we can define a sentence $\Box\varphi$ by (universally) quantifying over accessible worlds:

$$s \models_{\mathbf{S}} \Box_a \varphi \;\text{ iff }\; t \models_{\mathbf{S}} \varphi \text{ for all } t \text{ such that } s \xrightarrow{a} t.$$

$\Box\varphi$ may be interpreted as **knowledge** (in which case we use the notation $K_a\varphi$ instead) or **belief** (in which case we use $B_a\varphi$ instead), depending on the context.

Its *existential dual*

$$\Diamond_a \varphi := \neg\Box_a\neg\varphi$$

denotes a sense of **"epistemic/doxastic possibility"**.

"Common" Modalities

The sentence $C\Box\varphi$ is obtained by quantifying over all worlds that are accessible by any concatenations of arrows:

$$s \models_{\mathbf{S}} C\Box\varphi \text{ iff } t \models_{\mathbf{S}} \varphi \text{ for every t and every a finite chain}$$

$$\text{(of length } n \geq 0) \text{ of the form } s = s_0 \stackrel{a_1}{\rightarrow} s_1 \stackrel{a_2}{\rightarrow} s_2 \cdots \stackrel{a_n}{\rightarrow} s_n = t.$$

$C\Box\varphi$ may be interpreted as **common knowledge** (in which case we use the notation $Ck\varphi$ instead) or **common belief** (in which case we use $Cb\varphi$ instead), depending on the context.

The Problem of Common Knowledge

Note that common knowledge *cannot be expressed* in basic epistemic logic:

$$\varphi \quad ::= \quad p \quad | \quad \neg\varphi \quad | \quad \varphi \wedge \varphi \quad | \quad \Box_a \varphi$$

This means: there is no sentence in this syntax to fully capture the meaning of "sentence φ is common knowledge".

The obvious way to do this would be via the infinite sentence:

$$\varphi \wedge \bigwedge_{a \in \mathcal{A}} \Box_a \varphi \wedge \bigwedge_{a,b \in \mathcal{A}} \Box_a \Box_b \varphi \wedge \ldots$$

But this is *not* a correct sentence in this language: being infinite, it cannot be constructed in finitely many steps using the logical operators $\neg$, $\wedge$ and $\Box$. This is true even if the set $\mathcal{A}$ of all agents is finite!

So, to capture common knowledge, we have to extend our language to *"full" epistemic logic*:

$$\varphi \quad ::= \quad p \quad | \quad \neg\varphi \quad | \quad \varphi \wedge \varphi \quad | \quad \Box_a \varphi \quad | \quad C\Box\varphi$$

Doxastic Models

A **doxastic model** (or $KD45$**-model**) is a Φ-Kripke model satisfying the following properties:

- **(D) Seriality**: for every s there exists some t such that $s \xrightarrow{a}$ t;
- **(4) Transitivity**: If $s \xrightarrow{a} t$ and $t \xrightarrow{a} w$ then $s \xrightarrow{a} w$
- **(5) Euclideaness** : If $s \xrightarrow{a} t$ and $s \xrightarrow{a} w$ then $t \xrightarrow{a} w$

In a doxastic model, $\Box_a$ is interpreted as **belief**, and denoted by B_a.

EXERCISE

The following are valid in every doxastic model:

1. **Consistency of Beliefs**:

$$\neg B_a(\varphi \wedge \neg\varphi)$$

2. **Positive Introspection of Beliefs:**

$$B_a\varphi \Rightarrow B_a B_a\varphi$$

3. **Negative Introspection of Beliefs:**

$$\neg B_a\varphi \Rightarrow B_a \neg B_a\varphi$$

Epistemic ($S5$-) Models

An **epistemic model** (or $S5$**-model**) is a Kripke model in which all the accessibility relations are **equivalence relations**, i.e. **reflexive**, **transitive** and **symmetric** (or equivalently: **reflexive, transitive and Euclidean**).

In an epistemic model, $\Box_a$ is interpreted as **knowledge**, and denoted by K_a.

EXERCISE

The following are valid in every epistemic model:

1. **Veracity of Knowledge:**

$$K_a\varphi \Rightarrow \varphi$$

2. **Positive Introspection of Knowledge:**

$$K_a\varphi \Rightarrow K_aK_a\varphi$$

3. **Negative Introspection of Knowledge:**

$$\neg K_a\varphi \Rightarrow K_a\neg K_a\varphi$$

S4 Models for weak types of knowledge

Many philosophers deny that knowledge is introspective, and in particular **deny that it is negatively introspective**. Both common usage and Platonic dialogues suggest that people **may believe they know things that they don't actually know**.

An $S4$**-model for knowledge** is a Kripke model satisfying only *reflexivity and transitivity* (but not necessarily symmetry or Euclideaness). This gives a model to a **weaker notion of "knowledge"**, one that is truthful and positively introspective, but *not necessarily negatively introspective*.

1.3. Logics of public and private announcements

PAL (the logic of public announcements) was first formalized (including Reduction Laws) by Plaza (1989) and independently by Gerbrandy and Groeneveld (1997).

The problem of **completely axiomatizing PAL in the presence of the common knowledge operator** was first solved by Baltag, Moss and Solecki (1998).

A **logic for "secret (fully private) announcements"** was first proposed by Gerbrandy (1999).

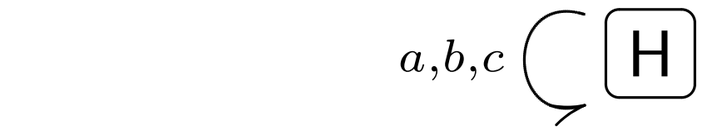
A **logic for "private, but legal, announcements"** (what we will call "*fair-game announcements*") was developed by H. van Ditmarsch (2000).

Scenario 2: The coin revealed

The referee c opens his palm and shows the face of the coin to everybody (to the public, composed of a and b, but also to himself): they **all see** it's Heads up, and **they all see that the others see it** etc.

So this is a **"public announcement" that the coin lies Heads up**.

We denote this event by !H. Intuitively, after the announcement, we have common knowledge of H, so the model of the new situation is:



Public Announcements are (Joint) Updates!

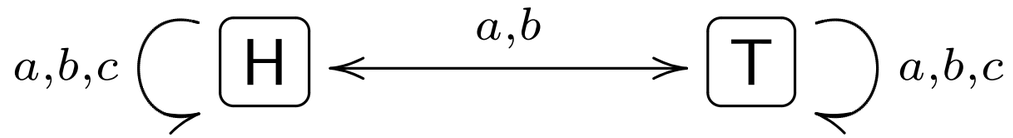
But this is just the result of **updating with** H: deleting all the non-H-worlds.

So, in the multi-agent case, **updating captures public announcements**.

From now on, we denote by $!\varphi$ the operation of deleting the non-φ worlds, and call it **public announcement with** φ, or **joint update with** φ.

Scenario 3: 'Legal' Private Viewing

Instead of Scenario 2: in front of everybody, the referee (c) uncovers the coin, so that (they all see that) **he, and only he, can see the upper face**. This changes the initial model to



Now, c **knows** the real state. E.g. if it's Heads, he knows it, and disregards the possibility of Tails. a and b don't know the real state, but *they know that c knows it*. c's viewing of the coin is a "legal", non-deceitful action, although a private one.

Fair-Game Announcements

Equivalently: in front of everybody, an announcement of the upper face of the coin is made, but in such a way that (it is common knowledge that) only c hears it.

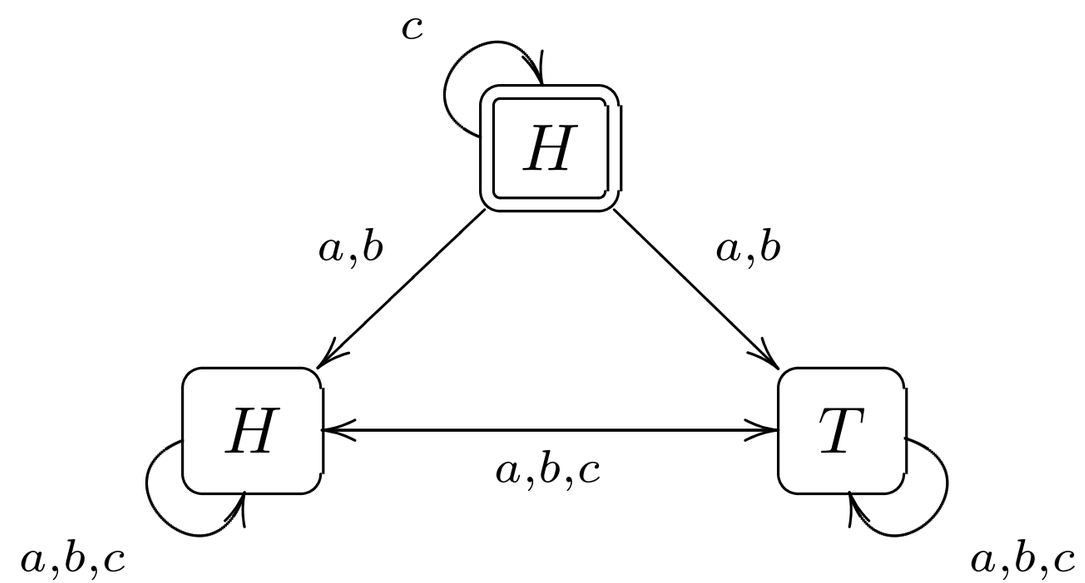
Such announcements (first modeled by H. van Ditmarsch) are called **fair-game announcements**, they can be thought of as "legal moves" in a fair game: nobody is cheating, all players are aware of the possibility of this move, but only some of the players (usually the one who makes the move) can see the actual move. The others know the range of possible moves at that moment, and they know that the "insider" knows his move, but they don't necessarily know the move.

Scenario 4: Cheating

Suppose that, after Scenario 1, the referee c has **taken a peek at the coin**, before covering it. **Nobody has noticed this**. Indeed, let's assume that c **knows that** a **and** b **did not suspect anything**.

This is an instance of **cheating**: a private viewing which is "illegal", in the sense that it is deceitful for a and b. Now, a and b think that nobody knows on which side the coin is lying. But they are wrong!

The Model after Cheating



We indicated the *real world* here. In the actual world (above), a and b think that the only possibilities are the worlds below. That is, they *do not even consider the "real" world as a possibility.*

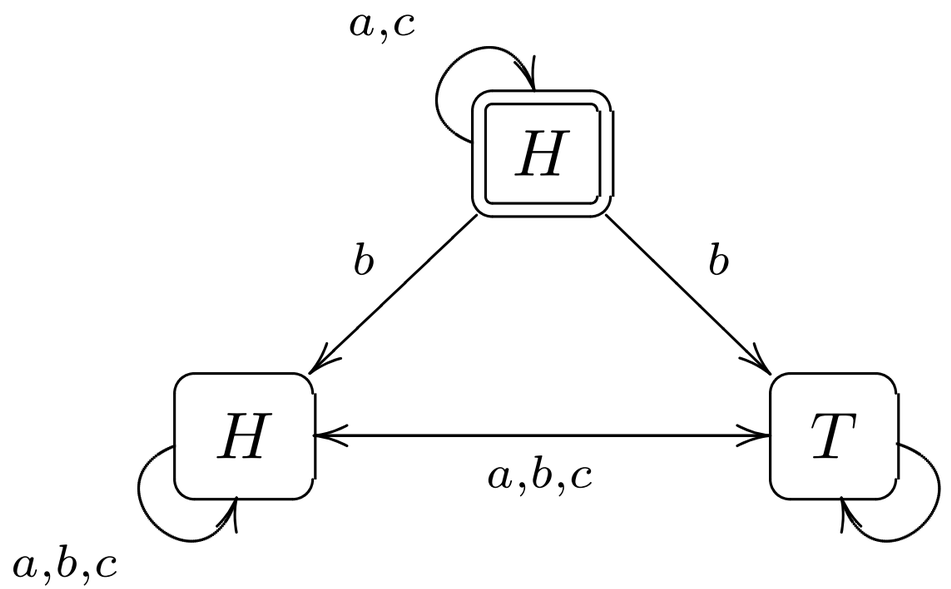
Such models in which we indicate the *real world* are called **pointed models.**

Scenario 5: Secret Communication

After cheating (Scenario 4), c engages in another "illegal" action: **he secretely sends an email to his friend a, informing her that the coin is Heads up**. Suppose the delivery and the secrecy of the message are guaranteed: so a and c have common knowledge that H, and that b doesn't know they know this.

Indeed, b is completely fooled: he doesn't suspect that c could have taken a peek, nor that he could have been engaged in secret communication.

The model is



Private Announcements

Both of the above actions were examples of completely **private announcements**

$$!_G\varphi$$

of a sentence φ to a group G of agents: in the first case $G = \{c\}$, in the second case $G = \{a, c\}$.

The "insiders" (in G) know what's going on, the "outsiders" don't suspect anything.

Scenario 5': Wiretapping?

In Scenario 5′, everything goes on as in Scenario 5, except that in the meantime b is **secretely breaking** into c's email account (or **wiretapping** his phone) and reading c's secret message. Nobody suspects this illegal attack on c's privacy. So both c and a think their secret communication is really secret and unsuspected by b: **the deceivers are deceived**.

What is the model of the situation after this action?! Things are getting rather complicated!

Scenario 6

This starts right after Scenario 2, when it was common knowledge that c knew the face. c attempts to send a secret message to a announcing that H is the case. c is convinced the communication channel is fully secure and reliable; moreover, he thinks that b doesn't even suspect this secret communication is going on. But, in fact, unknown and unsuspected by c, the message is *intercepted, stopped and read* by b. As a result, *it never makes it to* a, and in fact a never knows or suspects any of this. As for b, he *knows* all of the above: not only now he knows the message, but he knows that he "fooled" everybody, in the way described above.

The Update Problem

We need to find a *general method* to solve all the above problems, i.e. to compute all these different kinds of updates.

1.4. "Standard DEL"

- studies the **multi-agent information flow of "hard information"** (irrevocable, absolutely certain, fully introspective "knowledge") as well as "soft", but essentially un-revisable, information ("beliefs" that change monotonically, but are never overturned);

- gives an answer to the Update Problem, based on the BMS (Baltag, Moss and Solecki 98) setting: **logics of epistemic actions**;

- it arose from generalizing previous work on logics for public/private announcements.

- this dynamics is **essentially monotonic** (no belief revision!), though *it can model very complex forms of communication.*

Models for 'Events'

Until now, our Kripke models capture only *epistemic situations*, i.e. they only contain *static* information: they all are *state models*. We can thus represent the *result* of each of our Scenarios, but not what is actually going on.

Our scenarios involve various *types of changes* that may affect agents' beliefs or state of knowledge: a public announcement, a 'legal' (non-deceitful) act of private learning, 'illegal' (unsuspected) private learning etc.

We want to use now Kripke models to represent such types of *epistemic events*, in a way that is similar to the representations we have for epistemic states.

Event Models

An **event model** (or "*action model*")

$$\boldsymbol{\Sigma} \quad = \quad (\Sigma, \xrightarrow{\mathcal{A}}, pre)$$

is just like an Kripke model, except that its elements are now called **actions** (or "*simple events*") and instead of the valuation we have a **precondition map** pre, associating a sentence pre_σ to each action σ.

Epistemic/Doxastic Event Models

An event model is **epistemic**, or respectively a **doxastic**, event model if it satisfies the S5, or respectively the KD45, conditions.

Interpretation

We call of the simple events $\sigma \in \Sigma$ as *deterministic* actions of a particularly simple kind: they do not change the "facts" of the world, but the agents' beliefs. In other words, they are **"purely epistemic" actions**.

For $\sigma \in \Sigma$, we interpret pre_σ as giving the **precondition** of the action σ: this is a sentence that is true in a world iff σ can be performed. In a sense, pre_σ gives the implicit information carried by σ.

Finally, the accessibility relations express the agents' **knowledge/beliefs about the current action taking place**.

The Product Update

Given a state model $\mathbf{S} = (S, \xrightarrow{\mathcal{A}}, \|.\|)$ and an action model $\mathbf{\Sigma} = (\Sigma, \xrightarrow{\mathcal{A}}, pre)$, we define their *update product*

$$\mathbf{S} \otimes \mathbf{\Sigma} \quad = \quad (S \otimes \Sigma, \xrightarrow{\mathcal{A}}, \|.\|)$$

to be a new state model, given by:

1. $S \otimes \Sigma$ is
$$\{(s, \sigma) \in S \times \Sigma \ : s \models_{\mathbf{S}} pre_\sigma) \}.$$
2. $(s, \sigma) \xrightarrow{A} (s', \sigma')$ iff $s \xrightarrow{A} s'$ and $\sigma \xrightarrow{A} \sigma'$.
3. $\|p\|_{\mathbf{S} \otimes \mathbf{\Sigma}} = \{(s, \sigma) \in S \otimes \Sigma : s \in \|p\|_{\mathbf{S}}\}$.

Product of Pointed Models

As before, we can consider **pointed event models**, if we want to specify the **actual event** taking place.

Naturally, if initially the actual state was s and then the actual event is σ, then the actual output-state is (s, σ).

Interpretation

The product arrows encode the idea that: **two output-states are indistinguishable iff they are the result of indistinguishable actions performed on indistinguishable input-states**.

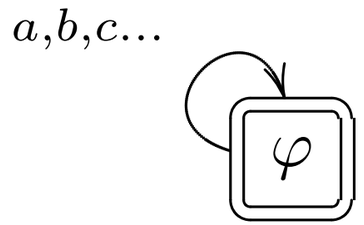
This comprises two intuitions:

1. "*No Miracles*": knowledge can only gained from (the epistemic appearance of) actions;
2. "*Perfect Recall*": once gained, knowledge is never lost.

The fact that the valuation is the same as on the input-state tells us that these actions are **purely epistemic**.

Examples: Public Announcement

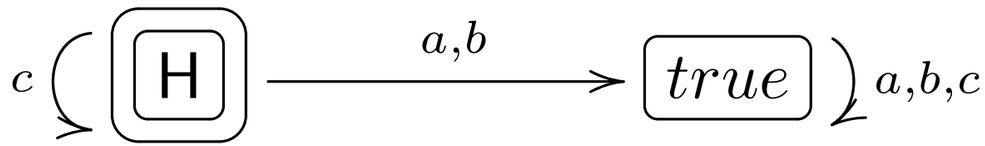
The event model $\mathbf{\Sigma}_{!\varphi}$ for public announcement $!\varphi$ consists of a single action, with precondition φ and reflexive arrows:



EXERCISE: Check that, for every state model $\mathbf{S}$, $\mathbf{S} \otimes \mathbf{\Sigma}_{!\varphi}$ is indeed the result of deleting all non-φ worlds from $\mathbf{S}$.

More Examples: Taking a Peek

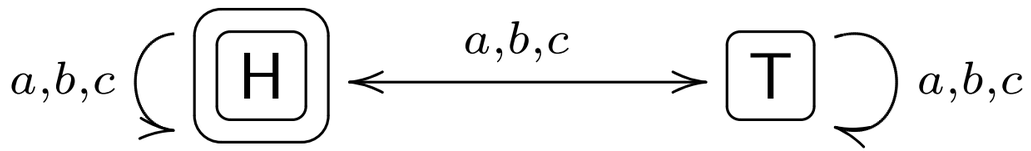
The action in Scenario 4: c takes a peek at the coin and sees the Head is up, without anybody noticing.



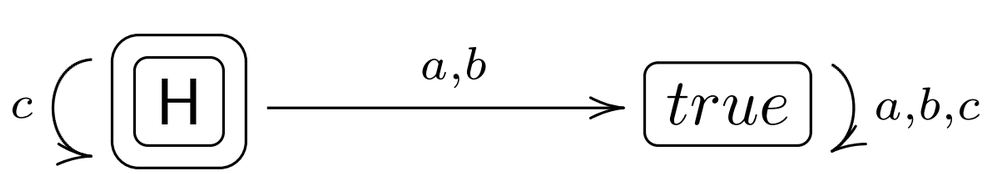
There are two actions in this model: the real event (on the left) is the **cheating action** of c "taking a peek". The action on the right is the apparent action *skip*, having any tautological sentence *true* as its precondition: this is the action in which **nothing happens**. *This is what the outsiders (a and b) think it is going on: nothing, really.*

The Product Update

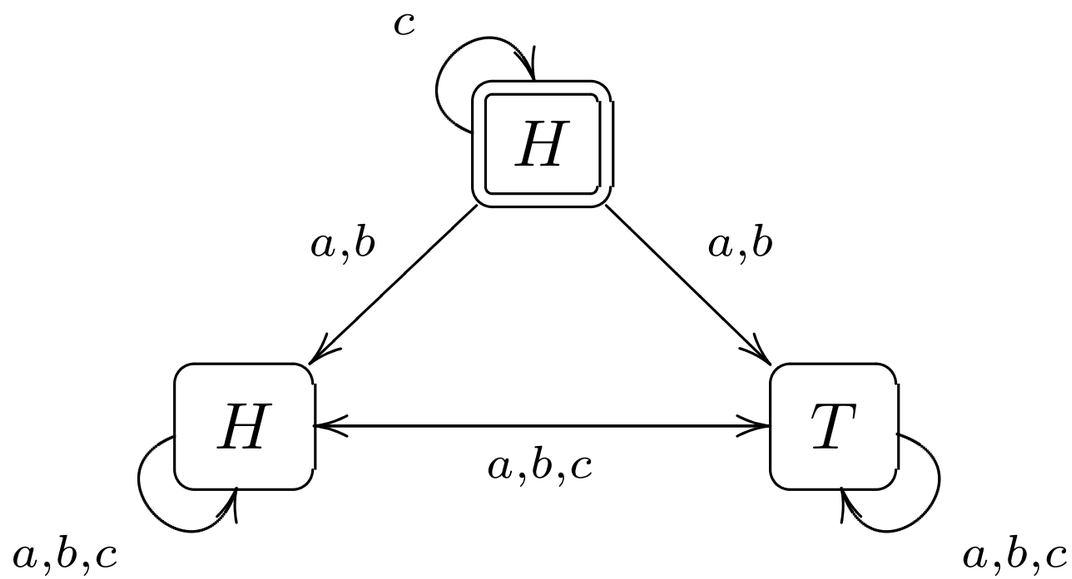
We can now check that the product of



and

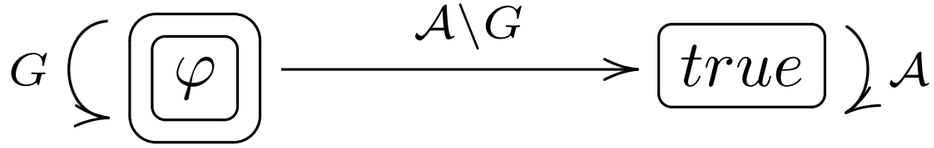


is indeed what intuitively should be:


c
H
a,b
a,b
H
T
a,b,c
a,b,c
a,b,c

Private Announcements

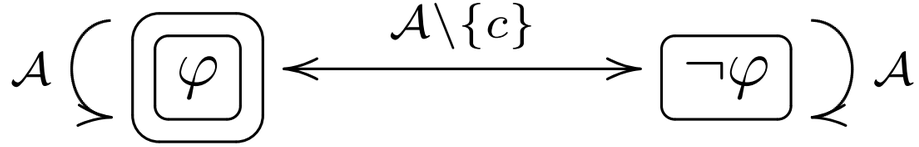
More generally, a fully **private announcement** $!_G\varphi$ of φ to a subgroup G is described by the action on the left in the event model



This subsumes both taking a peak (Example 4) and the secret communication in Example 5.

Fair-Game Announcements

The following event model represents the situation in which *it is common knowledge that an agent c privately learns whether φ or $\neg\varphi$ is the case*:

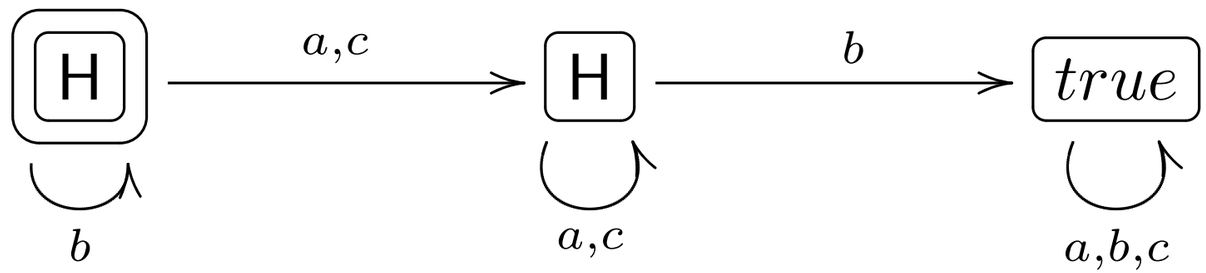


This is a "**fair-game announcement**" $Fair_c\varphi$.

The case $\varphi := \mathsf{H}$ represents the action in Example 3 ("legal viewing" of the coin by c).

Solving Scenario 5': Wiretapping

Recall Scenario 5: the supposedly secret message from c to a is secretly intercepted by b. This is an instance of a *private announcements with (secret) interception by a group of outsiders.*



Dynamic Modalities

For any action $\sigma \in \Sigma$, we can consider the corresponding **dynamic modality** $[\sigma]\varphi$. This is a property *of the original model*, expressing the fact that, if action σ happens, then φ will come to be true after that.

We can easily define the epistemic proposition $[\sigma]\varphi$ by:

$$s \models_{\mathbf{S}} [\sigma]\varphi \text{ iff } (s,\sigma) \in \mathbf{S} \otimes \mathbf{\Sigma} \text{ implies } (s,\sigma) \models_{\mathbf{S}\otimes\mathbf{\Sigma}} \varphi$$

Appearance

For any agent a and any action $\sigma \in \mathbf{\Sigma}$, we define the **appearance of action σ to** a, denoted by σ_a, as:

$$\sigma_a \quad = \quad \{\sigma' \in \Sigma : \sigma \xrightarrow{a} \sigma'\}$$

When σ happens, it appears to a as if either one of the actions $\sigma' \in \sigma_a$ is happening.

Examples

$$(!\varphi)_a \;=\; \{!\varphi\} \;\text{ for all } a \in \mathcal{A},$$

$$(!_G\varphi)_a \;=\; \{!_G\varphi\} \;\text{ for all insiders } a \in G,$$

$$(!_G\varphi)_a \;=\; \{skip\} = \{!(true)\} \;\text{ for all outsiders } a \notin G,$$

$$(Fair_a\varphi)_a \;=\; \{Fair_a\varphi\}$$

$$(Fair_a\varphi)_b \;=\; \{Fair_a\varphi, Fair_a\neg\varphi\} \quad \text{for } b \neq a.$$

Reduction Laws

If $\sigma \in \Sigma$ is a simple epistemic action, then we have the following properties (or "axioms"):

- *Preservation of "Facts"*. For all atomic $p \in \Phi$:

$$[\sigma]p \iff pre_\sigma \Rightarrow p$$

- *Partial Functionality*:

$$[\sigma]\neg\varphi \iff pre_\sigma \Rightarrow \neg[\sigma]\varphi$$

- *Normality*:

$$[\sigma]\left(\varphi \wedge \psi\right) \iff [\sigma]\varphi \wedge [\sigma]\psi$$

Here, $\Box$ can be *either knowledge* K *or belief* B, depending on whether the model is doxastic or epistemic.

- **"Action-Knowledge Axiom"**:

$$[\sigma]\Box_a\varphi \iff pre_\sigma \Rightarrow \bigwedge_{\sigma' \in \sigma_a} \Box_a[\sigma']\varphi$$

This Action-Knowledge Axiom helps us to *compute the state of knowledge/belief* of an agent *after* an event, in terms of the agent's *initial state of knowledge or belief* and of the event's *appearance* to the agent.

Instances of Action-Knowledge Axiom

If $a \in G$, $b \notin G$, $c \neq a$, then:

$$[!\theta]B_a\varphi \iff \theta \Rightarrow B_a[!\theta]\varphi$$

$$[!_G\theta]B_a\varphi \iff \theta \Rightarrow B_a[!_G\theta]\varphi$$

$$[!_G\theta]B_b\varphi \iff \theta \Rightarrow B_b\varphi$$

$$[Fair_a\theta]B_a\varphi \iff \theta \Rightarrow B_a[Fair_a\theta]\varphi$$

$$[Fair_a\theta]B_c\varphi \iff \theta \Rightarrow B_c([[Fair_a\theta]\varphi \wedge [Fair_a\neg\theta]\varphi)$$

EXERCISES

- Solve Scenario 5', by computing the update product of the state model obtained in Scenario 4 with the event model that we saw.
- Solve Scenario 6 using update product.
- Solve the Muddy Children puzzle, using repeated updates. Encode the conclusion of the puzzle in a DEL sentence.

1.5. Cheating and the Failure of Standard DEL

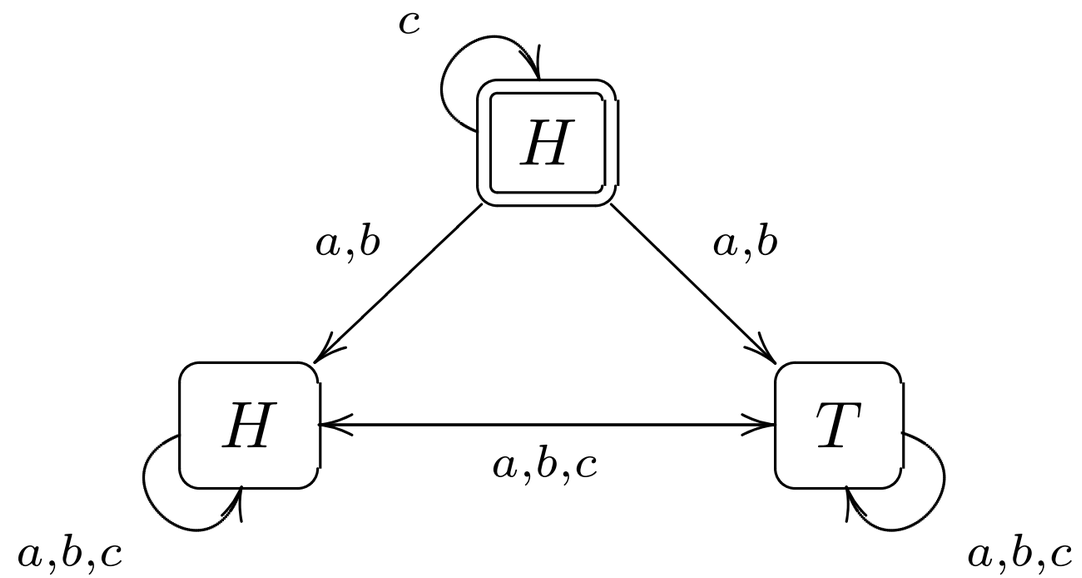
Our update product works very well when dealing with *"knowledge"*, or even with *(possibly false) beliefs*, **as long as these false beliefs are never contradicted by new information**.

However, in the latest case, update product gives unintuitive results: if an agent A is confronted with a contradiction between previous beliefs and new information she starts to believe the contradiction, and so she *starts to believe everything*!

In terms of epistemic models, this means that in the updated model, there are *no A-arrows originating in the real world.*

Counterexample

Recall the state model immediately after taking a peek, i.e. the output of Scenario 4:



So, now, c privately **knows** that the coin lies Heads up.

Counterexample Continued

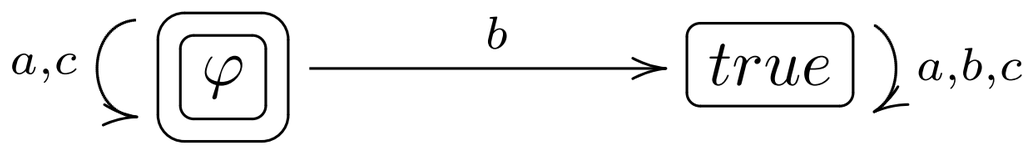
In Scenario 5 (happening after the cheating in Scenario 4), agent c sends a secret announcement to his friend a (**who has not suspected any cheating** till now!), saying:

"I **know** that H ".

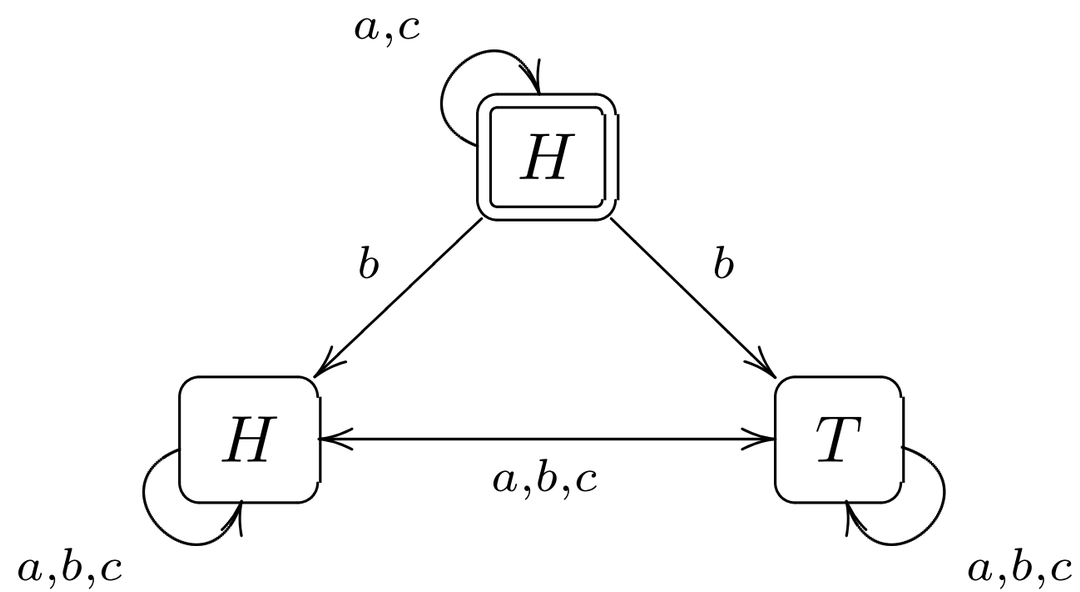
This is a fully **private communication** $!_{a,c}\varphi$ (from c to a) of the sentence

$$\varphi := K_c\mathsf{H},$$

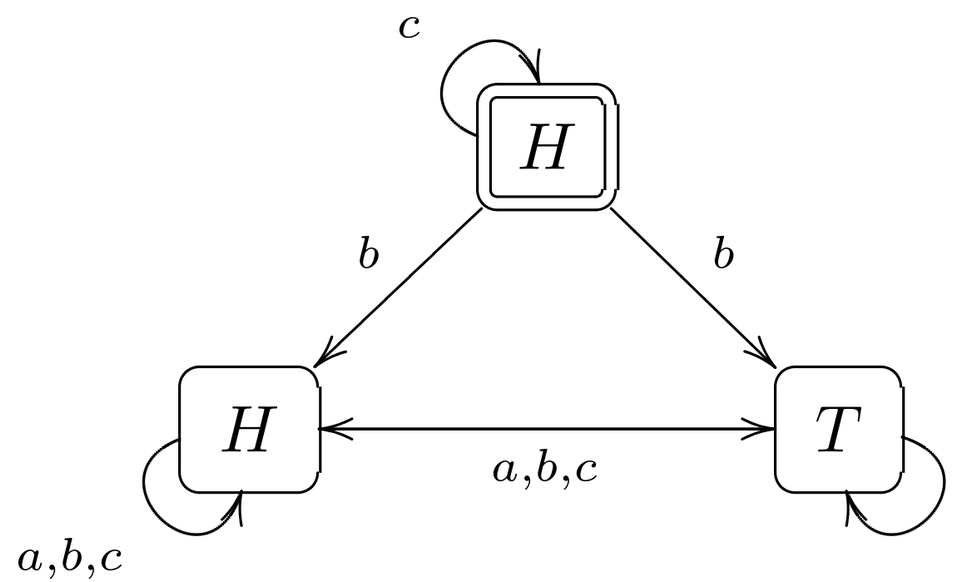
i.e. with event model



Recall that, according to our intuition, the updated model for the situation *after* this private announcement should be:



However, the update product gives us (something bisimilar to):



There are no surviving a-arrows originating in the real world. According to our semantics, *a will believe everything* after this communication: encountering a contradiction, **agent a simply gets crazy**!

Fixing this problem requires modifying update product by incorporating ideas from **Belief Revision Theory**.

2.1. The Problem of Belief Revision

What happens if I learn a new fact φ that goes in contradiction to my old beliefs?

If I accept the fact φ, **I have to give up some of my old beliefs.**

But which of them?

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

Example

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

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

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

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

But the base

$$\{p, q, \neg(p \wedge q)\}$$

is **inconsistent**!

So **I have to do more**!

Obviously, to accommodate the new fact $\neg(p \wedge q)$, **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 : logically closed sets of sentences
- *input: new information* (a formula) φ
- a *revision operator* $*$: a map associating a theory $T * \varphi$ to each pair (T, φ) of a theory and an input

Interpretation

$T * \varphi$ is supposed to represent the *new belief base ("new theory") 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).

AGM Postulates: The "Success" Axiom

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.

EXAMPLE: The 'AGM 'Success" Postulate

$$\varphi \in T * \varphi$$

"*After revising with φ, the agent's (revised) beliefs include (the belief in) φ.*"

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.*

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.*

Conditional Beliefs

Note that this 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.*

Another way to express this is that $T * \varphi$ captures **conditional beliefs** $B^{\varphi}\psi$:

we write $\psi \in T * \varphi$ iff $B^{\varphi}\psi$, i.e. if the agent believes ψ given φ.

We can think of conditional beliefs $B^{\varphi}\psi$ as *"contingency" plans for belief change*: **in case I will find out that φ was the case, I will believe that ψ was the case.**

2.2. Multi-Agent Plausibility Models

A **multi-agent plausibility model**:

$$\mathcal{S} = (S, \leq_a, \sim_a, \|.\|)_{a \in \mathcal{A}}$$

- S a set of **possible "worlds"** ("states")
- $\mathcal{A}$ a (finite) set of **agents**
- $\leq_a$ *preorders* on S "a**'s plausibility**" relation
- $\sim_a$ *equivalence relations* on S: a**'s ("hard") epistemic possibility (indistinguishability)**
- $\|.\| : \Phi \to \mathcal{P}(\mathcal{S})$ a valuation map for a set Φ,

subject to a number of *additional conditions*.

Explanation of terms

Recall:

Preorder means **reflexive** and **transitive**:

$$\forall s \in S \;\; s \leq_a s,$$

$$\forall s, t, w \in S \; (\, s \leq_a t \wedge t \leq_a w \Rightarrow s \leq_a w \,).$$

NOTE: Here, $s <_a t$ means that $s \leq_a t$ but $t \not\leq_a s$.

Reading

We read $s <_a t$ as saying that:
world t is **"better", or "more typical", or "more plausible"** than world s for agent a.

$s \leq_a t$ is the non-strict version:

world t is **"at least as good", "at least as typical", or "at least as plausible"** as world s for agent a.

The Conditions

The conditions are the following:

1. **"plausibility implies possibility"**:

$$s \leq_a t \text{ implies } s \sim_a t.$$

2. **the preorders are "locally connected" within each information cell**, i.e. indistinguishable states are comparable:

$$s \sim_a t \text{ implies either } s \leq_a t \text{ or } t \leq_a s$$

3. We consider S to be **finite** (else we need to require also that $\leq_a$ is **converse well-founded**).

Plausibility encodes Possibility!

Given these conditions, it immediately follows that **two states are indistinguishable for an agent iff they are comparable w.r.t. the corresponding plausibility relation**:

$$s \sim_a t \text{ iff either } s \leq_a t \text{ or } t \leq_a s.$$

But this means that **it is enough to specify the plausibility relations** $\leq_a$. *The "possibility" (indistinguishability) relation can simply be* **defined** *in terms of plausibility*

Simplified Presentation of Plausibility Models

So, from now on, we can **identify** a multi-agent plausibility model with a structure

$$(S, \leq_a, \|.\|)_{a \in \mathcal{A}}\,,$$

satisfying the above conditions, for which we **define** $\sim_a$ as:

$$\sim_a := \leq_a \cup \geq_a$$

In the same way as before, we define the *satisfaction relation* $s \models \varphi$, or equivalently we extend the *truth map* $\|\varphi\|_{\mathbf{S}}$ to all propositional formulas.

Knowledge, Conditional Belief

To define modalities, we need to extend the truth map further.

First the **notion of knowledge** is defined for each agent as follows:

$$s \models K_a\varphi \text{ iff } t \models \varphi \text{ for all } t \text{ such that } s \sim_a t$$

The notion of **(conditional) belief at a world** s is defined as **truth in all the most plausible worlds that are epistemically possible in** s **(and satisfy the given condition** $P \subseteq S$**)**:

$$s \models B_a^P\varphi \text{ iff } t \models \varphi \text{ for all } t \in Max_{\leq_a}\{t \in P : t \sim_a s\}.$$

Example of a Single Agent Model: 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**.

Formalizing the story

Our assumptions can be formalized as:

$$B_a\, genius$$

$$K_a(genius \vee drunk)$$

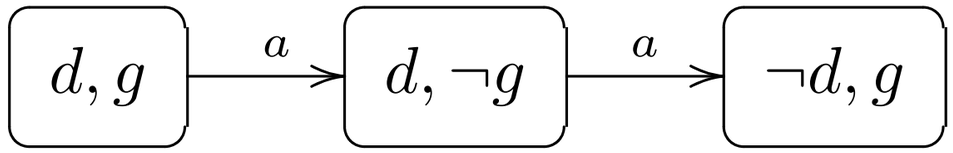
$$B_a \neg drunk$$

$$B_a^{drunk} \neg genius$$

$$drunk \wedge genius$$

The first four assumptions concern Albert's knowledge and (conditional) beliefs, while the fifth concerns reality.

The Model



Here, for precision, I included both positive and negative facts in the description of the worlds. 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 **knows** (K_a) he's drunk or a genius, so we did **NOT** include any world $(\neg d, \neg g)$.

Full Introspection of Knowledge and Beliefs

It is easy to see that our definitions imply that:

$$B_a\varphi \Rightarrow B_a B_a\varphi, \qquad B_a\varphi \Rightarrow K_a B_a\varphi,$$

$$\neg B_a\varphi \Rightarrow B_a \neg B_a\varphi, \qquad \neg B_a\varphi \Rightarrow K_a \neg B_a\varphi.$$

"Ideal" **agents know what they believe and what they don't:** if they believe something, then they believe, and in fact they **know**, **that they believe it**.

Similarly, if they don't believe something, then they believe, in fact they **know, that they don't believe it**.

WARNING: Difference from Kripke semantics

Plausibility models **ARE Kripke models**, but **the semantics of belief** in a plausibility model has **NOT** been given by the standard Kripke semantics. So **"belief" is NOT the Kripke modality for the plausibility relation**.

2.3. The Logic of Knowledge and Conditional Beliefs

Necessitation Rule:

From $\vdash \varphi$ infer $\vdash B_a^{\psi}\varphi$ and $\vdash K_a\varphi$.

Normality: $\vdash B_a^{\theta}(\varphi \Rightarrow \psi) \Rightarrow (B_a^{\theta}\varphi \Rightarrow B_a^{\theta}\psi)$

Truthfulness of Knowledge: $\vdash K_a\varphi \Rightarrow \varphi$

Persistence of Knowledge: $\vdash K_a\varphi \Rightarrow B_a^{\theta}\varphi$

Full Introspection: $\vdash B_a^{\theta}\varphi \Rightarrow KB_a^{\theta}\varphi$

$\vdash \neg B_a^{\theta}\varphi \Rightarrow K_a\neg B_a^{\theta}\varphi$

Hypotheses are (hypothetically) accepted:

$\vdash B_a^{\varphi}\varphi$

Proof System, continued

Consistency of Revision:

$\neg K_a \neg \varphi \Rightarrow \neg B_a^{\varphi} False$

Inclusion:

$\vdash B_a^{\varphi \wedge \psi} \theta \Rightarrow B_a^{\varphi}(\psi \Rightarrow \theta)$

Rational Monotonicity:

$\vdash B_a^{\varphi}(\psi \Rightarrow \theta) \wedge \neg B_a^{\varphi} \neg \psi \Rightarrow B_a^{\varphi \wedge \psi} \theta$

If we add **all the propositional validities** and the **Modus Ponens** rule, we obtain a **complete logic** for plausibility models.

2.4. "Dynamic" Belief Revision

We saw that AGM revision, or (equivalently) conditional beliefs, are in a sense "static":
they capture the agent's new (revised) beliefs about the OLD state of the world (as it was BEFORE the revision).

BUT the important problem is: to compute the agent's new beliefs (after learning some new information φ) **about the NEW state of the world (as it is AFTER the learning)**!

This is the subject of *"Dynamic" Belief Revision* theory.

From a *semantical* point of view, dynamic belief revision is about "revising" the whole relational structure: *changing the plausibility relation (and/or its domain).*

Upgrades (on single-agent models)

A **belief upgrade** **is a** *model transformer* T, that takes *any* plausibility model $\mathbf{S} = (S \leq, \| \cdot \|)$, and returns a *new* model $T(\mathbf{S}) = (S', \leq', \| \cdot \| \cap S')$, having:

- as new set of worlds: some *subset* $S' \subseteq S$,
- as new valuation: *the restriction* $\| \cdot \| \cap S'$ *of the original valuation to* S',
- as new plausibility relation: some converse-well-founded total preorder $\leq'$ on S'.

Hard and Soft Upgrades

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

$$S' = S$$

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

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

Dynamic Operators

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

Examples of Upgrades

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

(2) Radical upgrade $\Uparrow \varphi$ (Lexicographic upgrade with φ):
all φ-worlds become "better" (more plausible) than all $\neg\varphi$-worlds, and *within the two zones, the old ordering remains.*

(3) Conservative upgrade $\uparrow \varphi$ (minimal revision with φ):
the "best" φ-worlds become better than all other worlds, and *in rest the old order remains.*

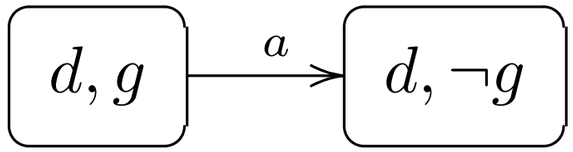
Different attitudes towards the new information

These correspond to *three different possible attitudes* of the agent towards *the reliability* of the source of the new information:

- **Update**: an **infallible** source. The source is *"known" (guaranteed) to be truthful.*
- **Radical (or 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!

Learning that you're drunk

Suppose that Albert **learns that he is definitely drunk** (say, by seeing the result of his blood test). By updating with the sentence d, we obtain:



which correctly reflects Albert's **new belief** that he is **not** a genius.

Exercise

Update Albert's original model with a Moore sentence:

Suppose an **infallible source** (the Pope) tells Albert:

"Albert, you are drunk but you don't believe it!"

$$d \wedge \neg B_a d.$$

Check that after learning the new information, Albert not only believes, but **he knows** that the new information was true before he learnt it.

Updates give you knowledge

After any update $!\varphi$, the agent comes to know that φ was true before the update.

we have the validity

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

"**Updates give you KNOWLEDGE, and not just BELIEF!**"

The reason is that an update $!\varphi$ is performed ONLY when the new information φ is **absolutely certain**: when the source of the information is infallible.

Mary Curry Enters the Story

Suppose that there is no blood test. Instead, he learns that he's drunk from **somebody who is trusted but not infallible**: NOT the Pope, but Albert's good friend Prof Mary Curry (not be confused with the famous Prof Marie Curie).

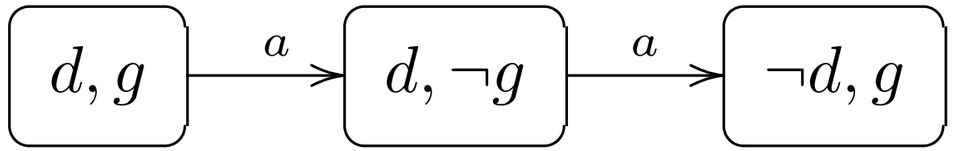
So Mary Curry tells Albert:

"*Man, you're drunk*!"

What to do with Professor Winestein?

Albert **trusts** Mary, so he **believes** she's telling the truth, but he **doesn't know** for sure: maybe she's pulling his leg, or maybe she's simply wrong.

How should we upgrade the model

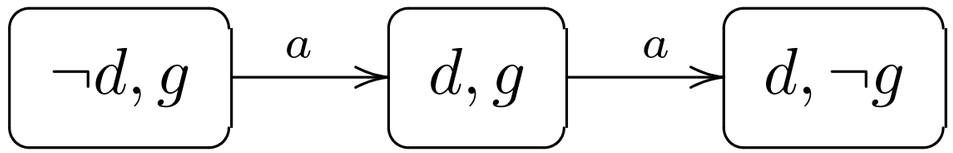


to capture Albert's new beliefs?

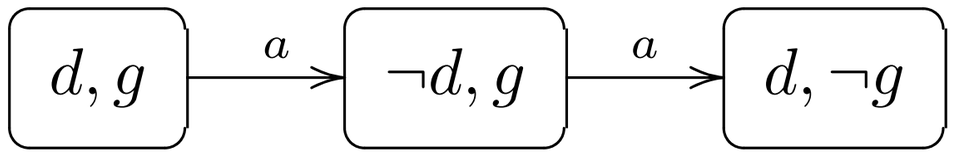
There are two drunk-worlds (d, g) and $(d, \neg g)$. **Which one should we promote ahead of all the others?**

Which is Best?

Maybe we should **promote both** drunk-worlds, making them more plausible than the other world $(\neg d, g)$:



Or maybe we should **promote only the most plausible of the two**:



Which is the best, most natural option??

How Strong is Your Trust

Actually, **they are both natural**, in different contexts and given different assumptions.

It all depends on **how strong is Albert's belief** that Mary tells the truth!

Strong Belief in single-agent models

A sentence φ is **strongly believed** in a single-agent plausibility model $\mathbf{S}$ if the following two conditions hold

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

$$\|\varphi\|_{\mathbf{S}} \neq \emptyset,$$

2. **all φ-worlds are strictly more plausible than all non-φ-worlds**:

$$s > t \text{ for every } s \in \|\varphi\|_{\mathbf{S}} \text{ and every } t \notin \|\varphi\|_{\mathbf{S}}.$$

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**!

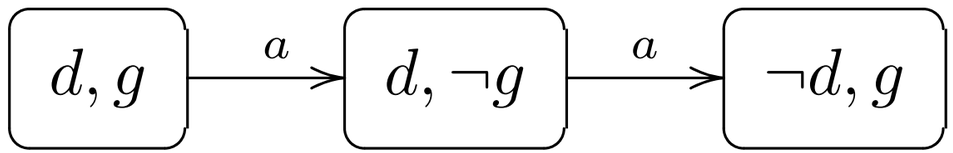
More precisely:
φ is **strongly believed** iff φ **is believed and is also conditionally believed given any new evidence (truthful or not) EXCEPT if the new information is known to contradict** φ; i.e. if:

1. $B_a\varphi$ holds, and
2. $B_a^\theta\varphi$ holds for every θ such that $\neg K_a(\theta \Rightarrow \neg\varphi)$.

Example

The "presumption of innocence" in a trial is a rule that asks the jury to hold a **strong belief in innocence** at the start of the trial.

In our Winestein example



Albert's belief that he is sober ($\neg d$) is a strong belief (although it is a **false** belief).

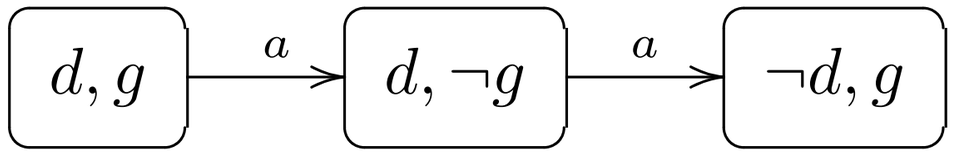
Radical Upgrade

If Albert has a **strong belief that Mary is telling the truth**, he will have to choose the first option: promote both d-worlds (in which Mary's statement is true), making them both more plausible than the other worlds.

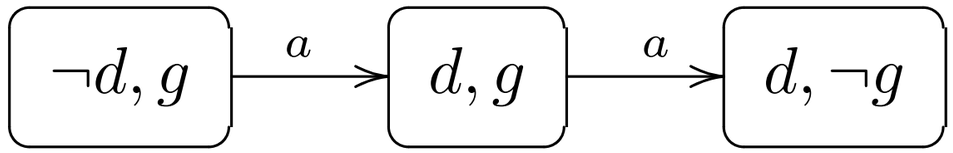
This corresponds to **radical upgrade**: it involves a rather radical revision of the prior beliefs, based on a strong belief in the correctness of the new information.

Example of Radical Upgrade

By performing a radical upgrade $\Uparrow d$ on the original model



we obtain



So we see that **Albert's strong belief that he was sober has been reverted: now he has acquired a strong belief that he is drunk!**

Fragile Trust

What if Albert's trust in Mary is more "fragile"?

Say, **he believes she's telling the truth, but he doesn't strongly believe it**: instead, he **"barely believes" it**.

This means that, after hearing Mary's statement, he acquires a very "weak" belief in it: if **later** some of his beliefs are found to be **wrong** and he will have to revise them, then **the first one to give up** will be his belief in Mary's statement.

Conservative Upgrade

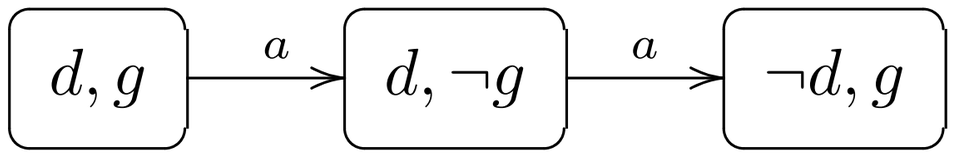
In this case, Albert will have to choose the second option: **promote only the most plausible d-world, leaving the rest the same.**

The change of order in this case is **minimal**: while acquiring a (weak) belief in d, Albert keeps **as much as possible** of his prior plausibility ordering (as much as it is consistent with believing d).

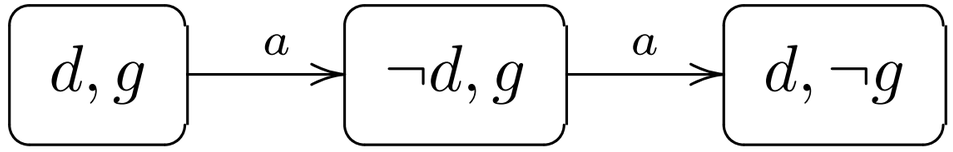
This corresponds to **conservative upgrade**.

Example of Conservative Upgrade

In the original Winestein situation



a conservative upgrade $\uparrow d$ produces the model



In this new model we have: $B_a d \wedge B_a^g \neg d$.

So Albert's new belief that he is drunk is **not strong**, and so is very **fragile**: if later Mary tells him he's a genius, he'll immediately revert to believing that he was sober!

Upgrades induce belief

We already saw that *updates induce knowledge* (in the new information):

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

In contrast, **soft upgrades only induce belief** (in the new information), and even this is only **conditional on consistency with prior knowledge**:

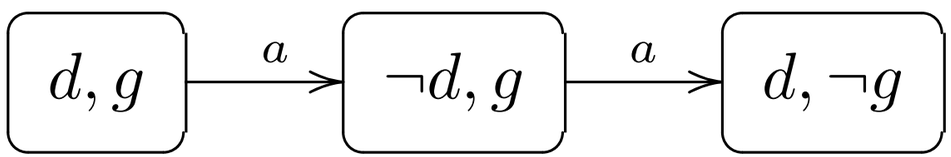
Indeed, after a *conservative or a radical upgrade*, the agent only comes to **believe** that φ (was the case), **UNLESS he already knew** (before the upgrade) that φ was **false**; i.e. we have the validity

$$\neg K_a \neg \varphi \Rightarrow [\uparrow \varphi] B_a(BEFORE\,\varphi)$$

Truthful and Un-truthful Upgrades

An upgrade is **truthful** if the new information φ is **true** (in the real world). The previous upgrades were all truthful.

But one can also upgrade with **false information**: if instead Mary told Albert "*You are not a genius*" and Albert strongly believed her, then the resulting model, obtained by the radical upgrade $\Uparrow \neg g$, would have been



This is an **un-truthful upgrade**: Albert acquires a strong (false) belief that he's not a genius.

Adding Mary Curry to the Winestein story

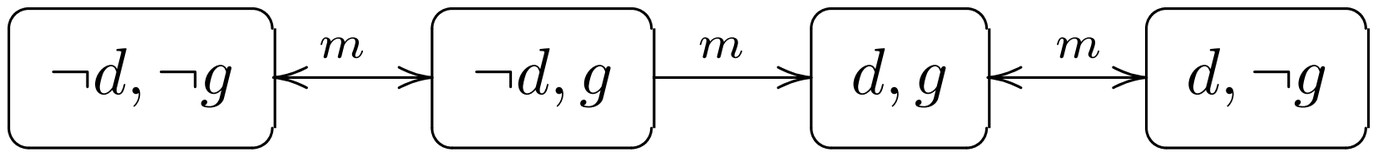
Albert Winestein's best friend is Prof. Mary Curry.

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

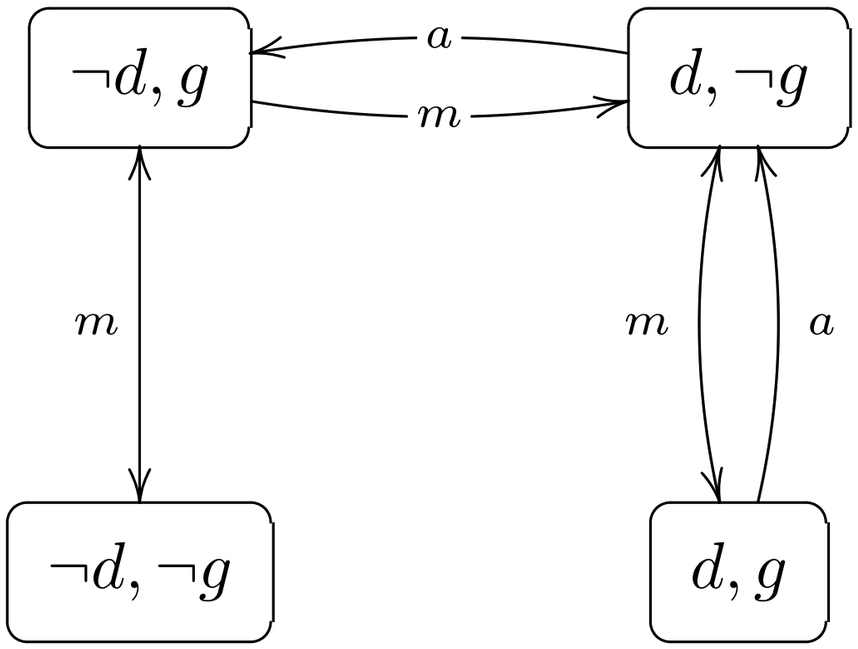
She's **completely indifferent with respect to Albert's genius**: she considers the possibility of genius and the one of non-genius as equally plausible.

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.

The model for Mary alone:



Multi-agent Model for Albert and Mary



Muddy Children Example

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

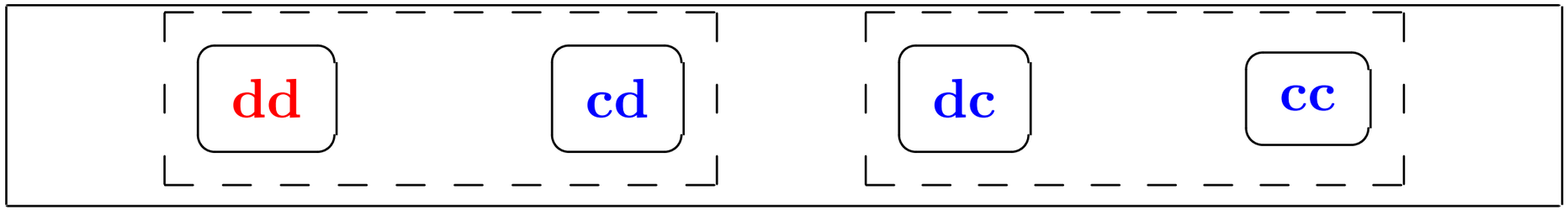
Dotted arrows: child 1's plausibility.

Continuous arrows: child 2's plausibility.

RED: the real world.

Information Partitions

From this, we can extract the information partitions:



Squares around the worlds: children's information cells.

Dotted squares: child 1.

Continuous squares: child 2.

2.5. Joint Upgrades and Updates

We can now apply the update or upgrade operations *simultaneously to all the relations*.

This corresponds to **joint upgrades or joint updates**:

some information φ is publicly announced, and it is common knowledge that all agents have the same attitude towards the announcement: they **upgrade or update with φ in the same way** (all doing an update, or a radical upgrade etc).

"Publicly Announced" Private Upgrades

Or the operation can be applied only to a single agent's relations (keeping the others unchanged), obtaining **"publicly-announced" private upgrades/updates**:

it is common knowledge that a single agent a upgrades/updates with φ, but also that the others do NOT upgrade/update at all with φ.

For instance, imagine a **publicly announces that he is upgrading/updating** with φ. It is *commonly known that he is telling the truth*, but also that *the others (not having direct access to the evidence for φ) are not convinced of the reliability of the information φ.*

Different Attitudes

More generally, we can **allow different agents to have different attitudes** towards the new information, by applying **different kinds of upgrade/update operations to different agents' relations**.

NOTE though that this still assumes **common knowledge of every agent's attitude towards the new information**: *the agents commonly know what kind of upgrade/update is performed by each of them.*

To go beyond that, *we'll need event models*!

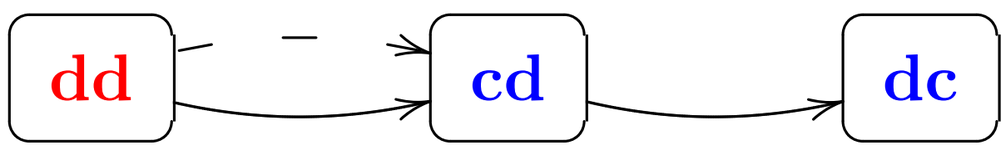
Muddy children example: A Joint Update

The Father announces:

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

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

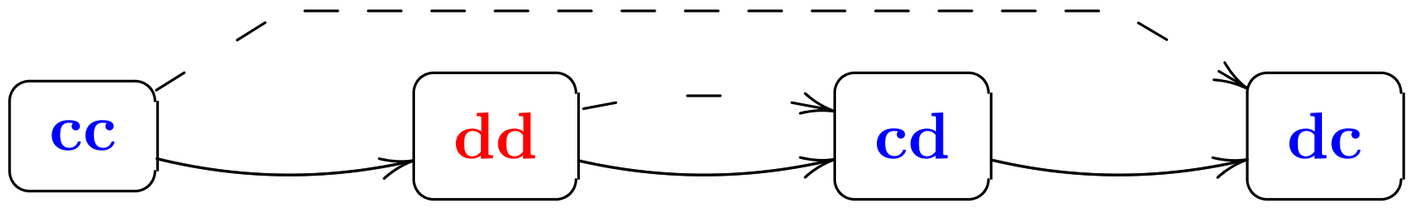
So this is an **update** $!(d_1 \vee d_2)$, yielding the updated model:



Muddy children example : Joint Radical Upgrade

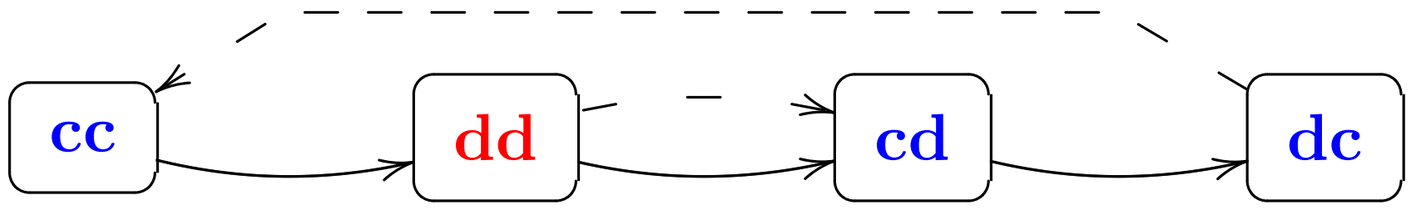
Alternatively, an older sister announces: "*At least one of you is dirty*". She is a **highly trusted source, though not infallible**:

This radical upgrade yields:



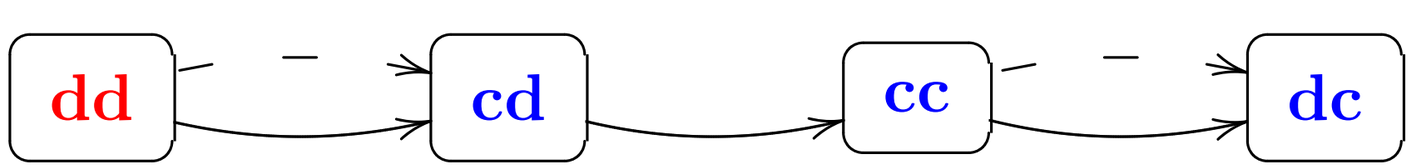
Children example: "Publicly Announced" Private Upgrade

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 *publicly announced private upgrade* by child 2:



Muddy children example: 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*:



2.6. Doxastic Event Models

More general upgrades, will look **very much like the DEL event models**.

There are **some differences** though: first, DEL event models were multi-agent, while the upgrades we saw were single-agent.

BUT... this can be easily fixed:

generalize to multi-agent upgrades, by having plausibility relations $\leq_a$ labeled by agents!

This was done by **G. Aucher** (– though using a different, more "quantitative way", way to encode plausibility relations, in terms of *Spohn ordinals* representing "degrees of belief").

Event Plausibility Models (G. Aucher)

A **multi-agent event plausibility model**

$$\mathbf{\Sigma} \quad = \quad (\Sigma, \leq_a, pre)$$

is just like a multi-agent state plausibility model, except that its elements are now called **events** (or *actions*), and instead of the valuation we have a **precondition map** pre, associating a sentence pre_σ to each action σ.

Now, the preorders $\sigma \leq_a \sigma'$ capture **the agent's plausibility relations on events**: a considers it at least as plausible that σ' is happening than that σ is happening.

Looking for a General Update Rule

We would like to *compose any initial state plausibility model with any event plausibility model* in order to compute the *new state plausibility model* after the event.

We want to *keep the old DEL setting while also doing belief revision*: when **restricted to the "hard" epistemic relations** $\sim_a$, our construction should amount just to taking the **Product Update**

$$(S, \sim_a, \|.\|)_{a\in\mathcal{A}} \otimes (\Sigma, \sim_a, pre)_{a\in\mathcal{A}}$$

But how should we define the new plausibility $\leq_a$ on input-pairs (s, σ) ?

Various Rules

The first such plausibility update rule was proposed by G. Aucher.

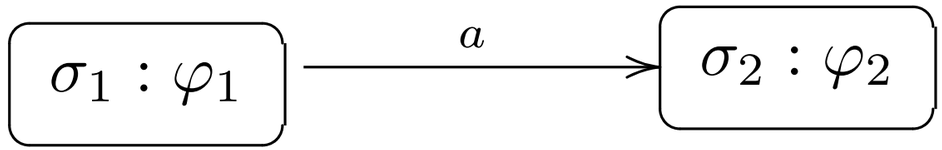
A number of other such rules were proposed and discussed by H. van Ditmarsch.

The one that I present here is the so-called "*Action-Priority* Rule", was proposed in (Baltag and Smets 2006). It has the advantage that it has purely relational, "qualitative" presentation (without the need of performing arithmetic operations on degrees of belief).

To derive the rule, we consider a number of special cases.

First Case

Well, in case that the event models includes a **strict** plausibility order between two events σ_1, σ_2 with precondition φ_1, φ_2



then **we kind of know the answer from the single-agent upgrade**: all the φ_2-worlds (s_2, σ_2) should become **strictly more plausible** than all the φ_1-worlds (s_1, σ_1).

The only problem is that, since we now have also *worlds that are known to be impossible by the agent*, the above rule should *NOT apply to those*:

if the agent can already distinguish between s_1 and s_2, then he knows which of the two is the case, so he doesn't have to compare the outputs (s_1, σ_1) and (s_2, σ_2).

So we get the following conditions:

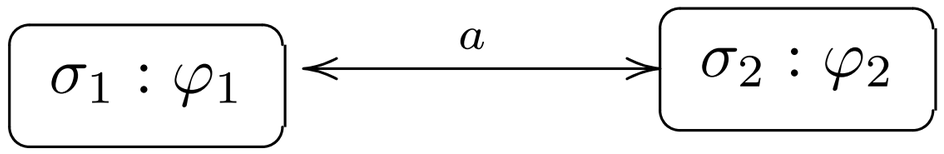
$$s_1 \sim_a s_2 \text{ and } \sigma_1 <_a \sigma_2 \text{ imply } (s_1, \sigma_1) <_a (s_2, \sigma_2),$$

and also

$$s_1 \not\sim_a s_2 \text{ implies } (s_1, \sigma_1) \not\sim_a (s_2, \sigma_2).$$

Second Case

What if the event model includes **two equally plausible events**?



We interpret this as **lack of information**: when the (unknown) event happens, it doesn't bring any information indicating which is more plausible to be currently happening: σ_1 or σ_2. In this case it is natural to expect the agents to *keep unchanged their original beliefs, or knowledge, about which of the two is more plausible.*

Let us denote by $\cong$ the **equi-plausibility relation on events**, given by:

$$\sigma \cong_a \sigma' \text{ iff } \sigma \leq_a \sigma' \leq_a \sigma.$$

Then the last case gives us another condition:

$$s_1 \leq_a s_2 \text{ and } \sigma_1 \cong_a \sigma_2 \text{ implies } (s_1, \sigma_1) \leq_a (s_2, \sigma_2).$$

Third Case

Finally, what if **the two events are epistemically distinguishable**: $\sigma \not\sim_a \sigma'$?

Then, when one of them happens, the agent knows it is not the other one.

By perfect recall, he can then distinguish the outputs of the events, and hence the two outputs are not comparable. So

$$\sigma \not\sim_a \sigma' \text{ implies } (s_1, \sigma_1) \not\leq_a (s_2, \sigma_2).$$

The Action-Priority Rule

Putting all these together, we get the following update rule, called the **Action-Priority Rule**:

$$(s, \sigma) \leq_a (s', \sigma') \text{ iff: either } \sigma <_a \sigma', s \sim_a s' \text{ or } \sigma \cong_a \sigma', s \leq_a s'.$$

This essentially says that we order the product space using the *anti-lexicographic preorder relation on comparable pairs* (s, σ).

The Action-Priority Update

As before, the set of states of the new model $\mathbf{S} \otimes \mathbf{\Sigma}$ is:

$$S \otimes \Sigma := \{(s, \sigma) : s \models_{\mathbf{S}} pre_\sigma\}$$

The valuation is given by the original valuation: $(s, \sigma) \models p$ iff $s \models p$.

The plausibility relation is given by the *Action-Priority Rule*.

Interpretation

The anti-lexicographic preorder gives "priority" to the *action* plausibility relation. This is not an arbitrary choice: it is in the spirit of AGM revision. The action plausibility relation captures the agent's **current beliefs about the current event**: what the agents *really believe is going on at the moment*.

In contrast, the input-state plausibility relations only capture **past beliefs**. **The past beliefs need to be revised by the current beliefs, and NOT the other way around**! *The doxastic action is the one that "changes" the initial doxastic state, and NOT vice-versa.*

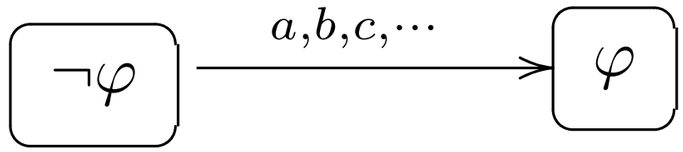
EXAMPLE: joint update

The event model for a joint radical update $!\varphi$ is essentially the same as in standard DEL (the event model for a "public announcement"):

(As usual for plausibility models, we do NOT draw the loops, but they are there.)

EXAMPLE: joint radical upgrade

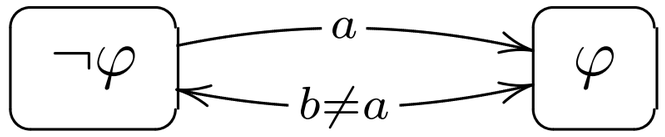
The event model for a joint upgrade $\Uparrow \varphi$ is:



EXERCISE: Check that, for every state model $\mathbf{S}$, $\mathbf{S} \otimes \mathbf{\Sigma}_{!\varphi}$ is indeed (isomorphic to) the result of performing the joint radical upgrade $\Uparrow \varphi$ on $\mathbf{S}$.

EXAMPLE: publicly-announced private upgrade

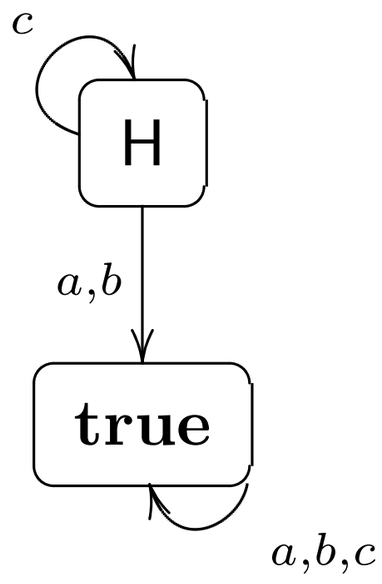
The event model for a publicly-announced private (radical) upgrade with φ is:



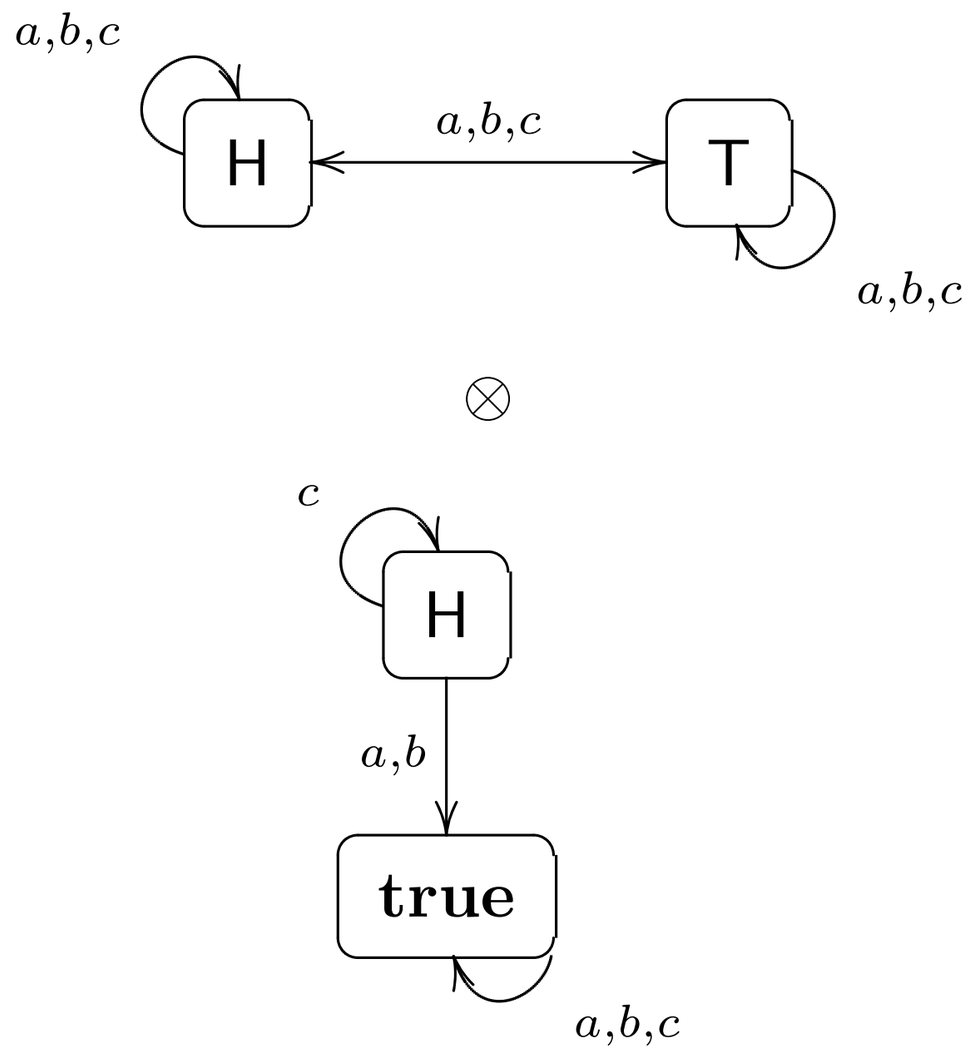
Example: Secret (Fully Private) Announcement

Let us consider again the "cheating" Scenario from the beginning: the referee (Charles, i.e. agent c) takes a peek at the coin and sees it's Heads up, when nobody looks. Alice (a) and Bob (b) don't suspect anything: they *believe that nothing is really happening.*

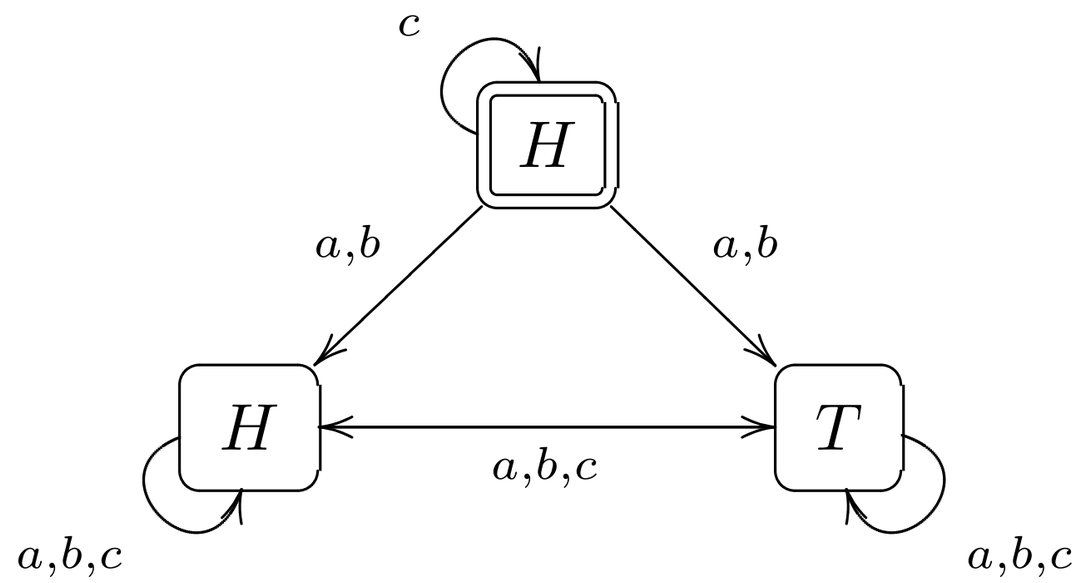
The DEL event model for this action was



By taking update product



of the initial state and this DEL event model, we obtained a state model of the situation after this action:



This correctly reflected the agents' BELIEFS after the cheating action.

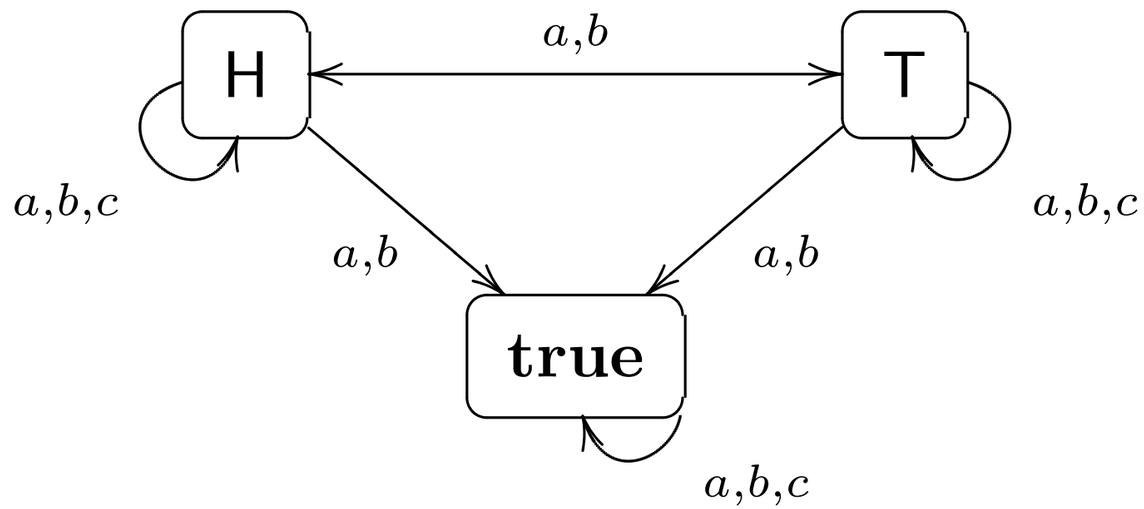
However, this is NOT the correct PLAUSIBILITY model for the new situation: it does NOT correctly reflect the agents' CONDITIONAL beliefs after the cheating.

For instance, the above model (if seen as a plausibility model) would suggest that, if later Charles tells Alice that he took a peek (without telling her what face he saw), she will immediately start to believe that he saw the coin Heads up!

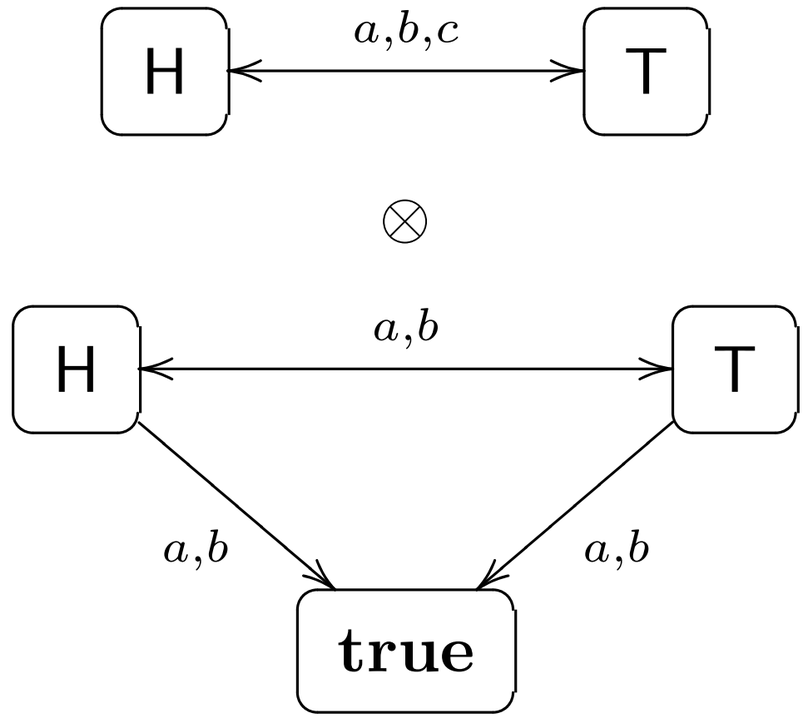
To compute the correct plausibility model, we need first to figure the correct event plausibility for the above action. For this, we still need to ask: *what does this event tell Alice (a) and Bob (b) about the face of the coin in case Charles (c) took a peek*?

In other words, given this event, if Alice or Bob later learn that Charles took a peek, what would they believe as more likely: that he saw H or T?

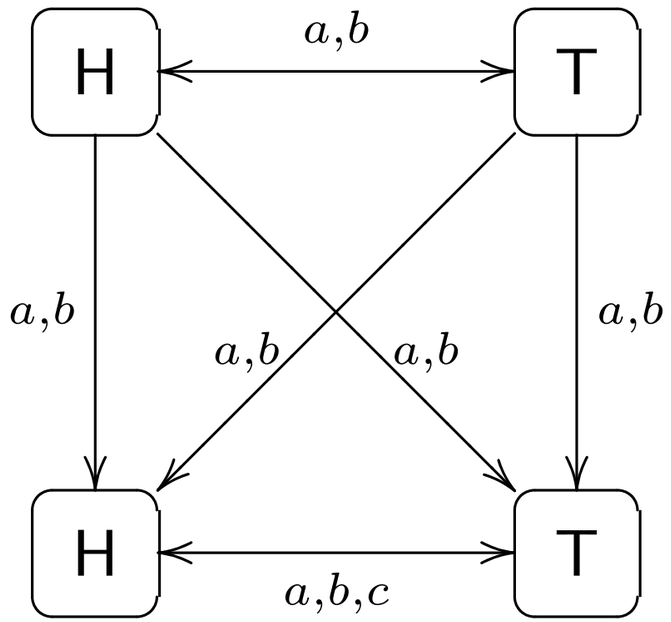
Clearly, this event *doesn't carry ANY new information* for Alice and Bob, so she should stick with whatever she believed before about the coin. Hence, the event model is



The Action-Priority update of the original state (plausibility) model with this event plausibility model (skipping the loops):



gives us:



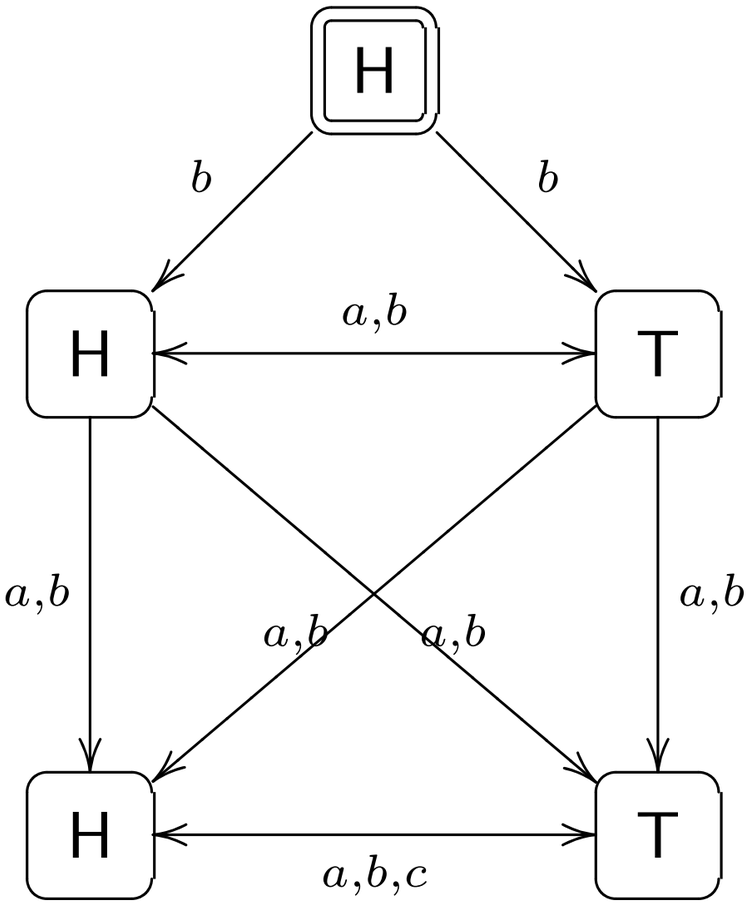
So e.g. a still believes that c doesn't know the face. However, if later she's given the information that he took a peek (without being told what he saw), she'd know that he knows the face; but as for herself, she'd still consider both faces equally plausible.

Solving The Problem from the beginning

What if now Charles **secretely tells** Alice that he knows the face of the coin is Heads up?

With the setting of standard DEL, this drove Alice crazy: she started believing everything!

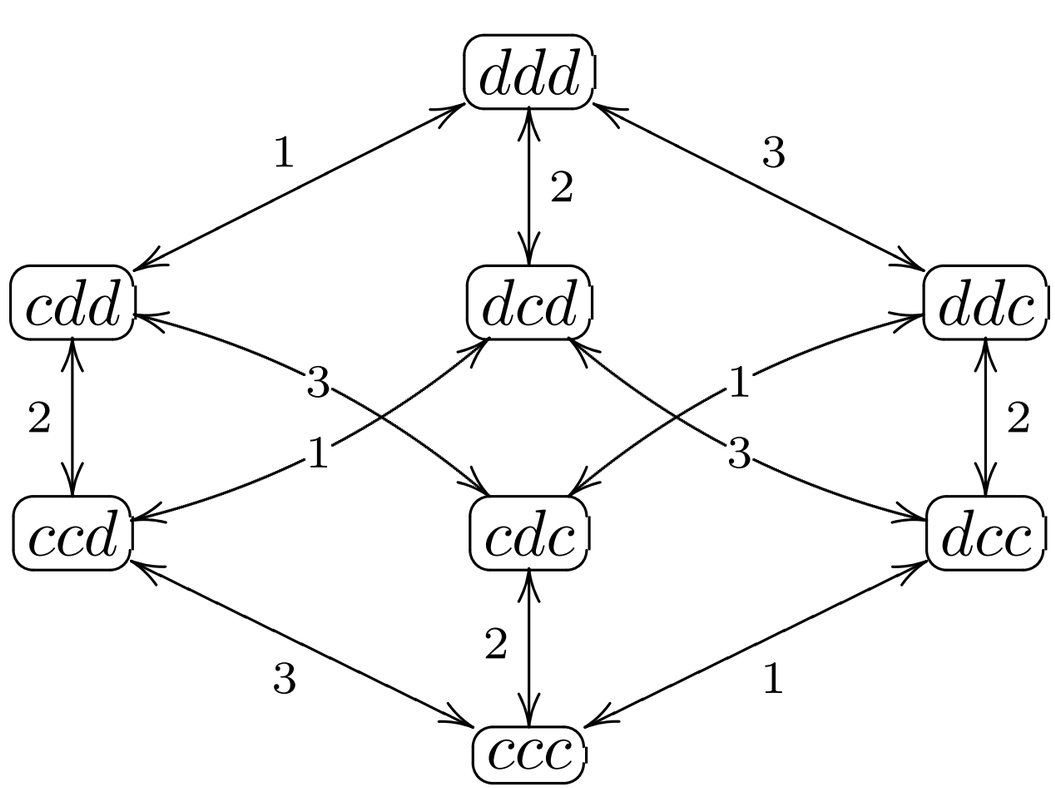
Now, things are better. The real world (in which Charles knows H) is still epistemically possible for Alice. So after the fully private announcement $!_a(K_c H)$, the plausibility model simply becomes:



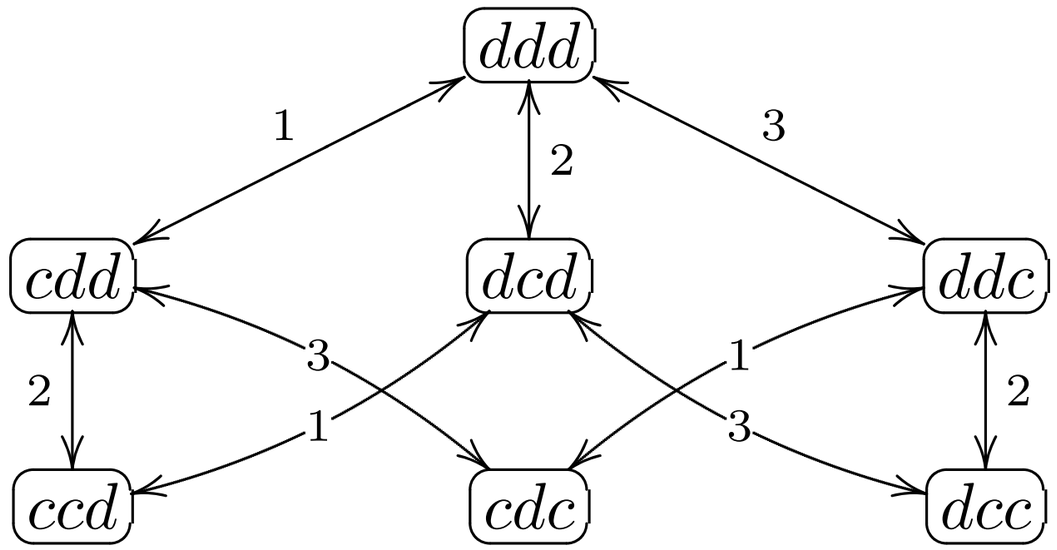
Exercise: what is the event model that gave us this plausibility model?

Solving the standard "Muddy Children"

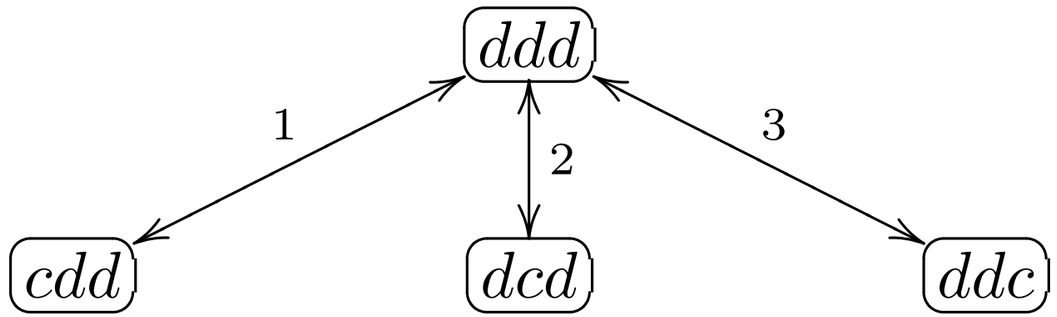
Three children, child 1 and child 2 are dirty. Originally, assume each child considers equally plausible that (s)he's dirty and that (s)he's clean:



Father makes the announcement: "At least one of you is dirty". If he's an infallible source (classical Muddy children), then this is an update $!(d_1 \vee d_2 \vee d_3)$, producing:



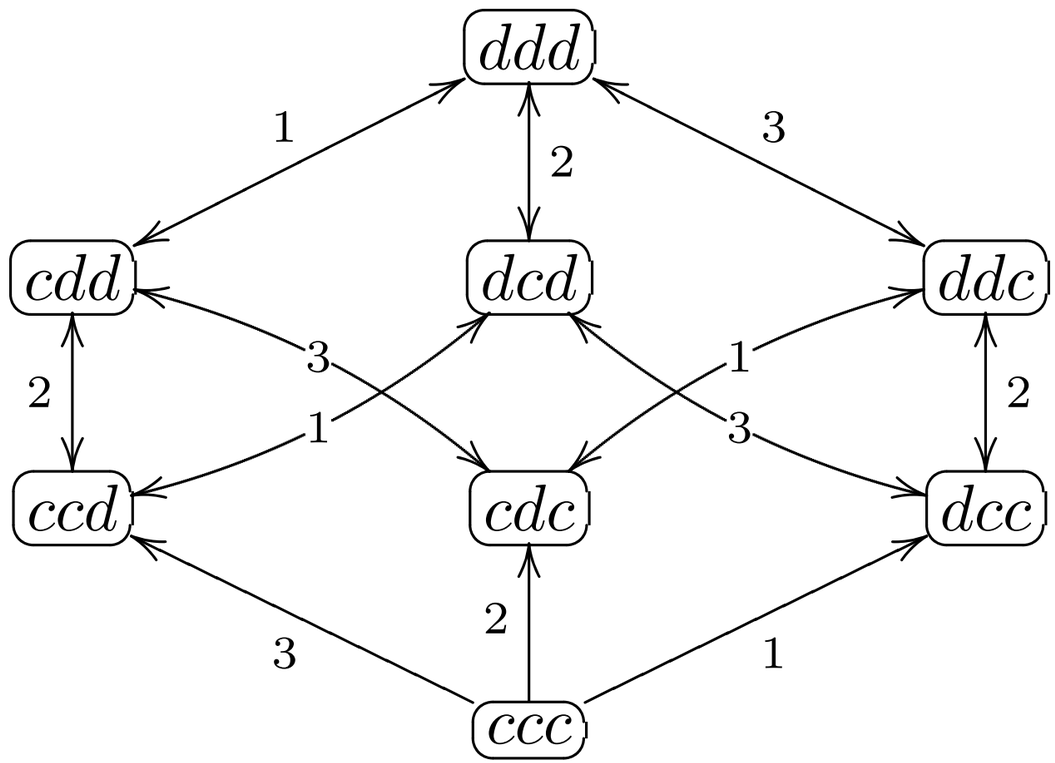
If the children answer "I don't know I am dirty", and they are infallible, then the update $!(\bigwedge_i \neg K_i d_i)$ produces:



Now, in the **real** world (d, d, c), children 1 and 2 **know** they are dirty.

Soft version of the puzzle

What happens if the sources are not infallible? Father's announcement becomes either a *radical upgrade* $\Uparrow(d_1 \vee d_2 \vee d_3)$ or a *conservative* one $\uparrow(d_1 \vee d_2 \vee d_3)$, producing:

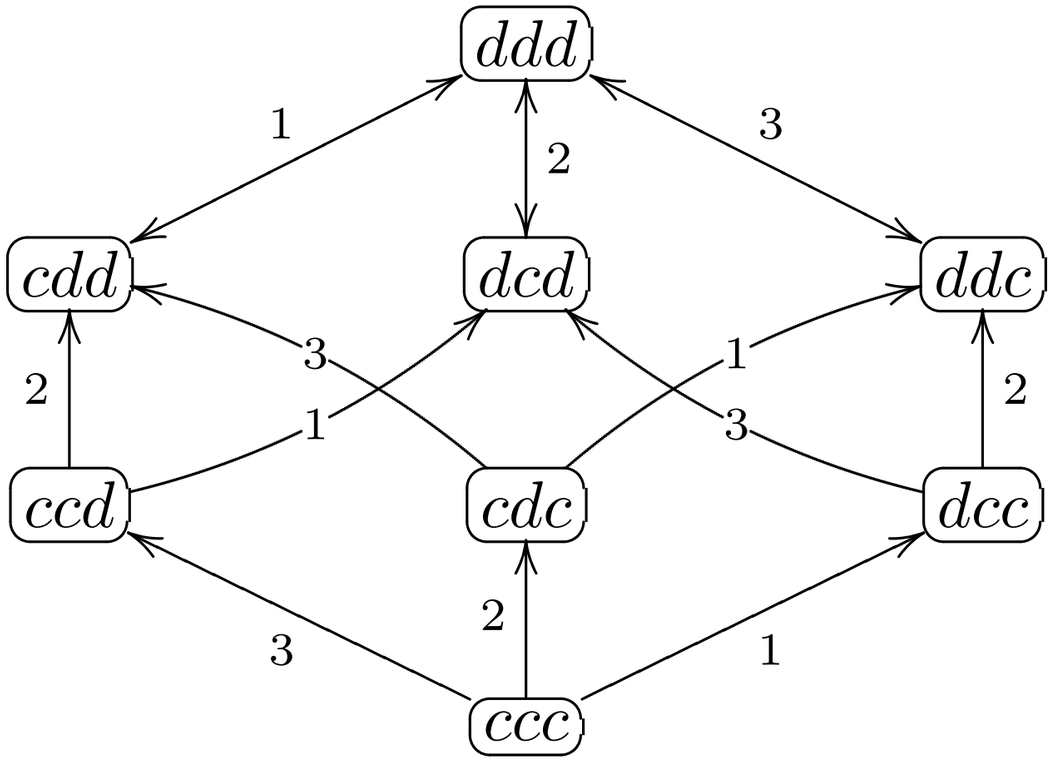


Do you believe you're dirty?

What if next the father only asks them if they **believe** they are dirty?

And what if they are *not infallible* agents either (i.e. don't trust each other, but not completely), so that their answers are also *soft announcements*?

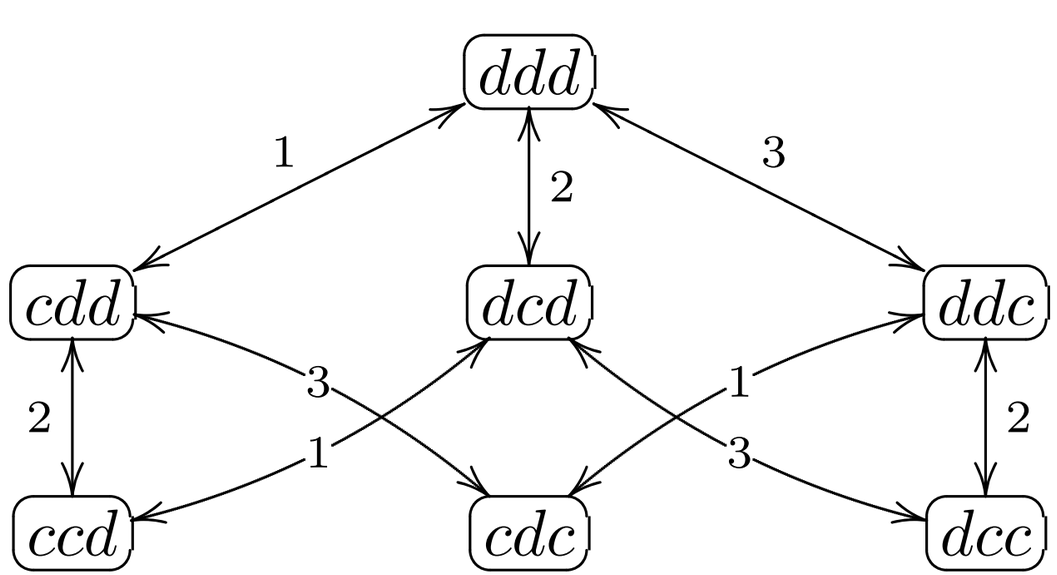
After a (radical or conservative) upgrade with the sentence $\bigwedge_i \neg B_i d_i$, we obtain:



Now (in the real world ddc), children 1 and 2 **believe they are dirty**: so they will answer "yes, I believe I'm dirty".

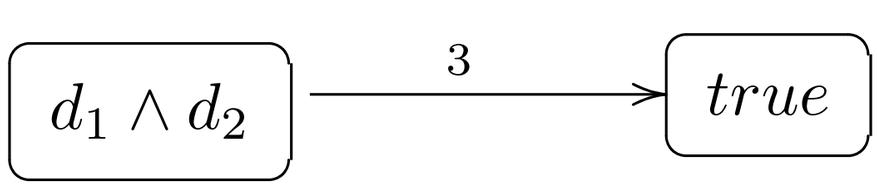
Cheating Muddy Children

Let's get back to the original puzzle: assume again that it is common knowledge that nobody lies, so we have infallible announcement (updates). After Father's announcement, we got



Secret Communication

Suppose now the dirty children cheat, telling each other that they are dirty. This is a *secret communication* between 1 and 2, in which 3 doesn't suspect anything: he thinks nothing happened. So it has the event model:



EXERCISE

Taking the Action-Priority Update of the previous model with this event model.

Then model the next announcement (in which the two children say "I know I'm dirty", while the third says "I don't know") as a joint update $!(K_1 d_1 \wedge K_2 d_2 \wedge K_3 d_3)$.

Note that, after this, child 3 does NOT get crazy: unlike in the standard DEL (with Product update), he simply realizes that the others cheated!

3.1 Iterated Revision and the Learning Problem

Question 1. THE ITERATION PROBLEM: investigate the **long-term behavior of iterated learning** of **higher-level doxastic information**.

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

Long-term behavior: whether the learning process **comes to an end, stabilizing the doxastic structure**, or **keeps changing it forever**. In particular, *do the agent's beliefs stabilize, reaching a fixed point*? Do the *conditional beliefs*?

Question 2. THE LEARNING PROBLEM: *Do the beliefs stabilize on* **truth**, converging to the *real world*?

Iterating Upgrades

To study iterated belief revision, consider a **finite model** $\mathbf{S}_0 = (S, \leq_0, \| \cdot \|_0, s_0)$, and an **(infinite) sequence of 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 radical upgrades

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

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

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

defined by:

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

Iterated Updates Always Stabilize

OBSERVATION: For every initial finite model $\mathbf{S}_0$, 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

$$\mathbf{S}_n = \mathbf{S}_m \text{ for all } m \geq n.$$

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!

Iterated Updates always stabilize, but 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.**

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).

SURPRISE: 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**!

BIGGER SURPRISE: This still holds even if we **revise with the same true sentence every time**:

• Conservative case: $\uparrow \varphi, \uparrow \varphi, \ldots, \uparrow \varphi, \ldots$
Simple beliefs never stabilize.

• Radical case: $\Uparrow \varphi, \Uparrow \varphi, \ldots, \Uparrow \varphi, \ldots$.
simple beliefs stabilize, but conditional beliefs don't.

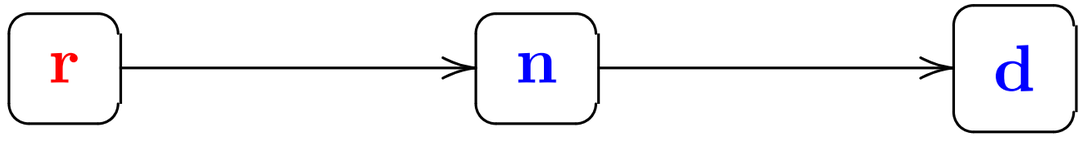
Example : Voting Case

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

He believes she will **vote Democrat**.

But in case this turns out wrong, he'd rather believe that she **won't vote** than accepting that she may vote Republican.

We assume that, **in reality** (unknown to Charles), Marry will **vote Republican**!



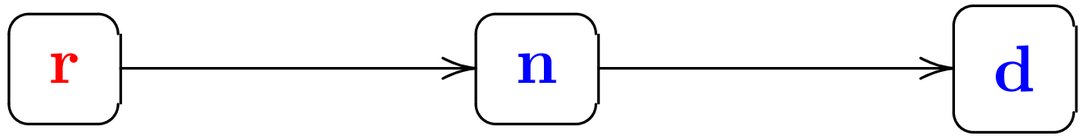
Iterating a Truthful Conservative Upgrade

Suppose a trusted informer tells Charles the following true statement φ:

$$\mathbf{r} \vee (\mathbf{d} \wedge \neg \mathbf{B}\mathbf{d}) \vee (\neg \mathbf{d} \wedge \mathbf{B}\mathbf{d})$$

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

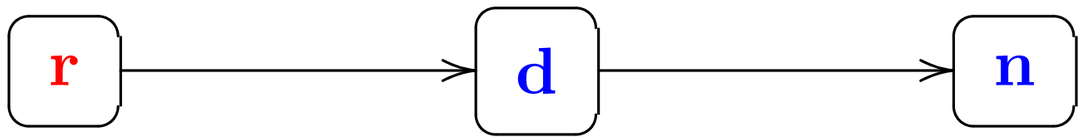
In the original model



the sentence φ is true in worlds r and n, but not in d.

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 n, so this becomes now the most plausible state overall:



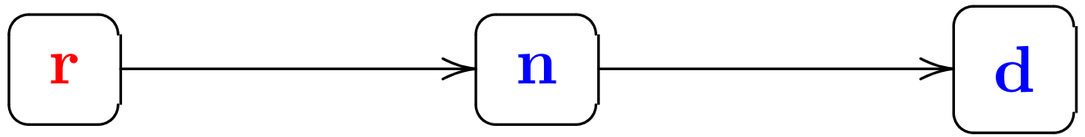
Now φ is *again true at the real world (r)* and in world d. So **this sentence can again be truthfully announced**.

If Charles **conservatively upgrades again** with φ, he will promote d on top, **reverting to the original model**!

Here, **The whole model (the plausibility order) keeps changing**, and Charles' (simple, un-conditional) **beliefs keep oscillating forever** (between d and n)!

Iterating Truthful Radical Upgrades

Consider the same original model:

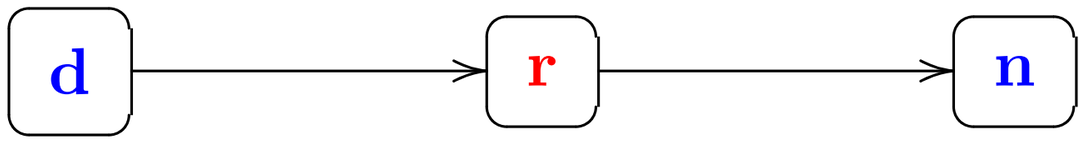


But now consider the sentence φ:

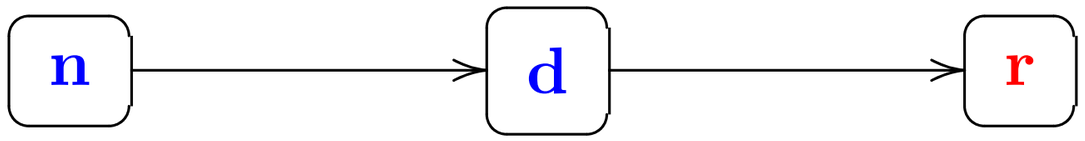
$$\mathbf{r} \vee (\mathbf{d} \wedge \neg \mathbf{B}^{\neg \mathbf{r}} \mathbf{d}) \vee (\neg \mathbf{d} \wedge \mathbf{B}^{\neg \mathbf{r}} \mathbf{d})$$

"*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*".

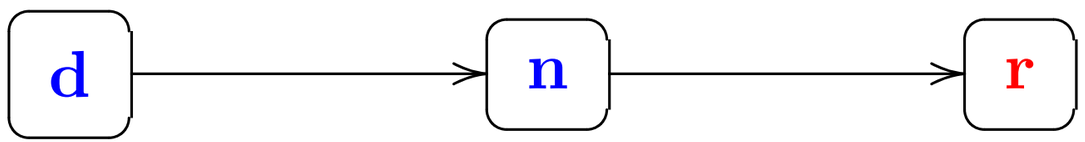
Sentence φ is true in the real world r and in n but not in d, so a **truthful radical upgrade** will give us:



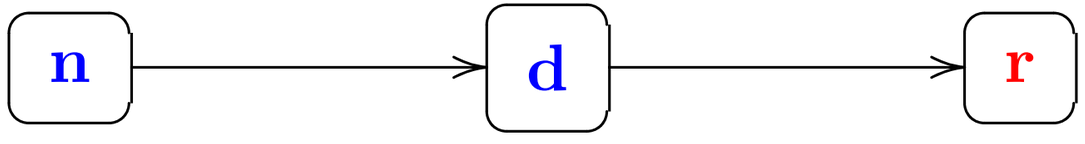
The same φ is again true in (the real world) r and in d, so it can again be truthfully announced, resulting in:



Another truthful upgrade with φ:



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



Stable Beliefs in Oscillating Models

These 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 radical example the simple (unconditional) beliefs eventually stabilize**: from some moment onwards, Charles correctly believes that the real world is r (vote Republican) and he will never lose this belief again!

This is a symptom of a more general phenomenon:

Beliefs Stabilize in Iterated Radical Upgrades

THEOREM:

In any infinite sequence of truthful radical 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.

Intermediate Conclusions

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.

Converging to the Truth?

So simple beliefs stabilize after an infinite series of truthful lexicographic upgrades. **But under what conditions do these beliefs stabilize on the Truth?**

Strongly informative upgrade streams

An upgrade with φ is called "**strongly informative**" on a pointed model $\mathbf{S}$ iff φ is **not already believed** at (the real world of) $\mathbf{S}$. I.e. $\mathbf{S}$ satisfies $\neg B\varphi$.

Now, an upgrade stream $\{\Uparrow \varphi_n\}_n$ is "**strongly informative**" if *each of the upgrades is strongly informative at the time when it is announced:* i.e. in the iteration, we have that

$$\mathbf{S}_n \models \neg B\varphi_n$$

Belief correcting upgrade and streams

Call an upgrade $\Uparrow \varphi$ "**belief-correcting**" on $\mathbf{S}$ iff φ is actually believed to be FALSE at $\mathbf{S}$. I.e.

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

Now, an upgrade stream is called "**belief-correcting**" if each of the upgrades is belief-correcting at the time when it is announced:

$$\mathbf{S}_n \models B\neg\varphi_n.$$

NOTE: "belief correcting" $\Rightarrow$ "strongly informative" (The converse fails.)

Maximal Strongly informative streams

An upgrade stream is a **"maximally" strongly-informative** (OR **"maximally belief-correcting"**), truthful stream if:

- (1) it is strongly-informative (OR belief-correcting) and truthful, and
- (2) it is maximal with respect to property (1): it cannot be properly extended to any stream having property (1).

So a strongly informative truthful stream is "maximal" iff it is **either infinite or** if, in case it is finite (say, of length n) then **there exists NO upgrade** $\Uparrow \varphi_{n+1}$ which would be **truthful and strongly informative** on the last model $\mathbf{S}_n$.

The results

1. **Every maximally belief-correcting lexicographic upgrade stream** $\{\Uparrow \varphi_n\}_n$ (starting on a given finite model $\mathbf{S}$) **is finite and converges to true beliefs**; i.e. in its final model $\mathbf{S}_n$, all the beliefs are true.

2. **Every maximally strongly-informative lexicographic upgrade stream** $\{\Uparrow \varphi_n\}_n$ (starting on a given finite model $\mathbf{S}$) **is finite and stabilizes the beliefs on FULL TRUTH**; i.e. in its final model $\mathbf{S}_n$, all beliefs are true and all true sentences are believed.

Note

But note that the last conclusion is NOT necessarily equivalent to saying that the set of most plausible worlds coincides in the end with only the real world!

The reason is that the language may not be expressive enough to distinguish the real world from some of other ones; and so the conclusion of 2 can still hold if the most plausible worlds are these other ones...

The above results do NOT hold for any other belief-revision methods except lexicographic (and conditioning).

Conclusions

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

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

- If we repeatedly (lexicographically) upgrade with THE SAME sentence in BASIC DOXASTIC logic, then all conditional beliefs eventually stabilize.

- In iterated truthful radical upgrades that are maximal strongly-informative, all believes converge to the truth and all true sentences are believed.

Other types of upgrades do not have these last positive properties.

This is not the full story!

We can extend the above positive result regarding repeated upgrades **beyond** basic doxastic logic, allowing *various forms of "knowledge" operators in the language.*

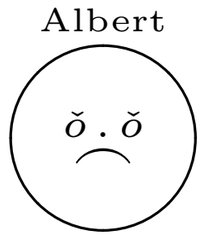
Still, there exist *important conditional-doxastic sentences* lying outside this fragment (e.g. "**Surprise**"-sentence in the Surprise Examination Puzzle) for which repeated lexicographic upgrades nevertheless stabilize the whole model!

3.2 How can "Agreement" be reached by "Sharing"?

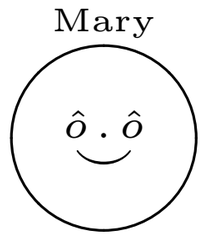
THE PROBLEM: we investigate the issue of **reaching doxastic agreement** among the agents of a group **by "sharing" information or beliefs**.

How can "Agreement" be reached by "Sharing"?

Example of a particular scenario:



Albert knows (D or G), believes G
conditional on D he believes $\neg G$



Mary doesn't know (D or G)
believes D

They share their information

$\Downarrow$

Together they know the same: (D or G) and both believe D and $\neg G$

Main Issues:

• **Agents' goal** = to reach a total doxastic or epistemic agreement ("merge").

• **Different types of agreements** can be reached: agreement only on the things they know, on some simple beliefs, strong beliefs etc.

• Depending on the type of agreement to be reached, what should be the **strategy**? Which communication protocol ? (given that the agents have some limited abilities in the way they communicate)

• We are interested in "**sharing**": joint (group) belief revision induced by **sincere, persuasive public communication** by either of the agents (the "*speaker*").

Talk publicly, sincerely, persuasively!

Rules of the game for our agents:

• **Public** communication: *common knowledge of what* (the content) is announced *and of the fact that all agents adopt* **the same attitude towards the announcement**; i.e. they have the same opinion about the reliability of this information (how plausible it is).

• **Sincerity:** *the communicated information was* **already "accepted"** *by* the speaker (before sharing it).

• **Persuasiveness**: the new information becomes commonly "accepted" by all agents; i.e. *everybody comes to share* ***the same attitude as the speaker*** *towards the communicated information.*

The Goal of Sharing is Total Agreement

After each act of sharing, all agents reach a partial agreement, namely with respect to the piece of information that was communicated.

The natural end of the sharing process is when **total agreement** has been reached: *all the agents' doxastic structures are exactly the same.*

After this, *nothing is left to share*: any further sincere persuasive communication is **redundant** from than on.

Dynamic Merge

When total agreement IS reached in this way, we say that the agents' doxastic structures have been **dynamically "merged" into one**.

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

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

Do the **communication agenda** (order of the items announced, allowing agents to interrupt the speaker) and the **group's hierarchy** make any difference?

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).

Merge Operations

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 $\rightarrow_i$.

• To merge the agents' **knowledge ("hard information")** K_i, it is enough to merge the epistemic indistinguishability relations $\overset{i}{\sim}$.

• To merge the agents' **soft information** (*all their "strong beliefs"* Sb_i, *or equivalently all their "conditional beliefs"* $B_i^P Q$), 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**. 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:

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

Lexicographic merge of soft information

Lexicographic 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**" $R_{a\otimes b}$:

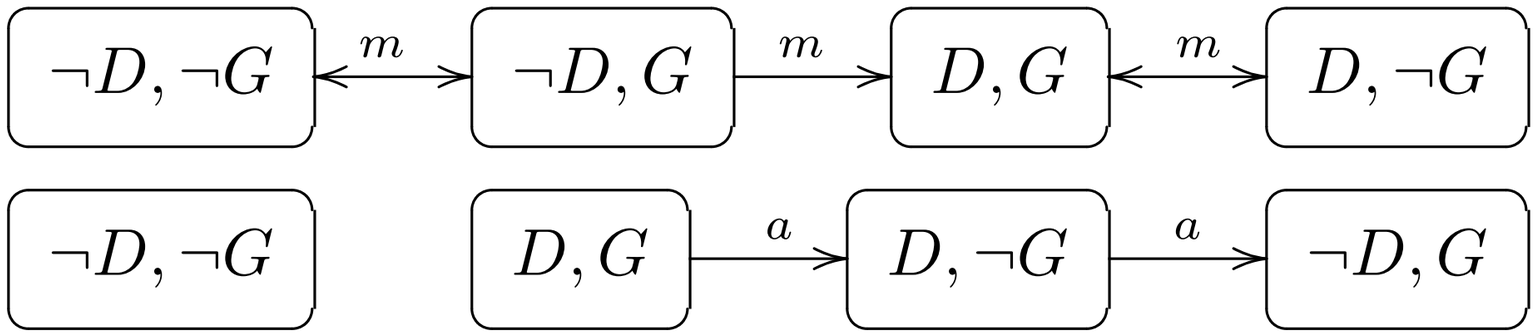
$$R_{a\otimes b} := (R_a^{>} \cap R_b^{\sim}) \cup (R_a^{\cong} \cap R_b) = R_a \cap R_b^{\sim} \cap (R_a^{>} \cup R_b).$$

Essentially, this means that **both agents have a "veto" with respect to group incomparability**:

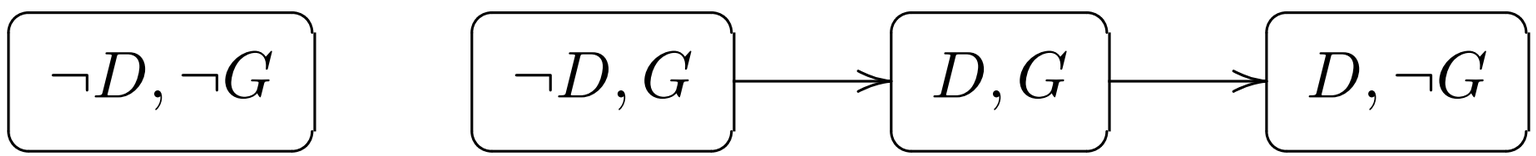
The group can only compare options that **both** agents can compare; **and whenever the group can compare two options, everything goes on as in the lexicographic merge**: agent a's strong preferences are adopted, while b's preferences are adopted only when a is indifferent.

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



gives us:



"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** $\bigodot$.

Realizing Distributed Knowledge

In the case of **knowledge**, it is easy to **design a protocol to realize it**, **as the parallel merge of agents' knowledge**, operation by a *sequence of joint updates*, as follows:

PROTOCOL: 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 Protocol

Formally, **the protocol for realizing distributed knowledge within group G** is:

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

(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.

Order-independence

The announcements may even be **interleaving**:

if the initial model is finite, then **any** "public" dialogue (of agents announcing facts they know) 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 Lexicographic Merge

Assuming we have NO NON-TRIVIAL "HARD" INFORMATION (i.e. that **all knowledge is common knowledge**):

then we can **realize the lexicographic merge** $\leq_{a/b}$ of SOFT INFORMATION **by joint radical upgrades**, via *a protocol very similar to the one for distributed knowledge.*

PROTOCOL: The agents have 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 lexicographic 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".

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 PROTOCOL

Formally, the protocol π' for realizing lexicographic 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).

Realizing Priority Merge

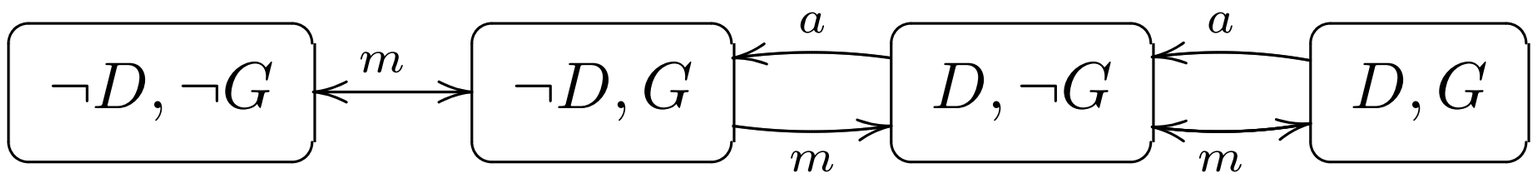
Finally: we can **realize the Priority Merge** $\bigotimes_i \leq_i$ of the whole PLAUSIBILITY ORDERS (encoding BOTH SOFT AND HARD INFORMATION), **by sequentially composing the two protocols above**.

FIRST, *the agents publicly announce "all they know", via joint updates*;

THEN, *respecting the priority order, they take turns announcing "all that they strong believe", via joint radical upgrades.*

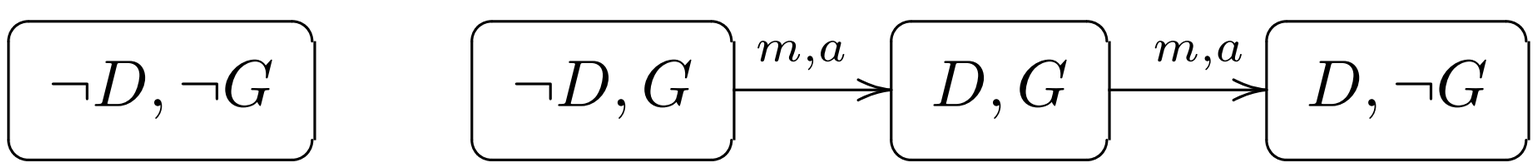
Example

In the situation from Example 1



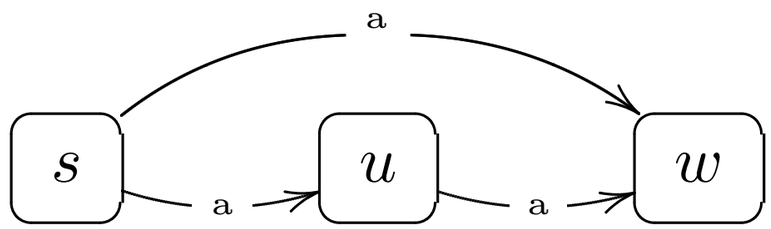
the **protocol to realize the Priority Merge** $R_{m \otimes a}$ consists of: *Albert's infallible* announcement (of *his "hard" knowledge* that $D \vee G$); then *Mary's sincere* announcement (of *her strong belief* D); then *Albert's sincere* announcement (of $\neg G$, which *he strongly believes after Mary's announcement*):

$$!(D \vee G); \Uparrow D; \Uparrow \neg G$$

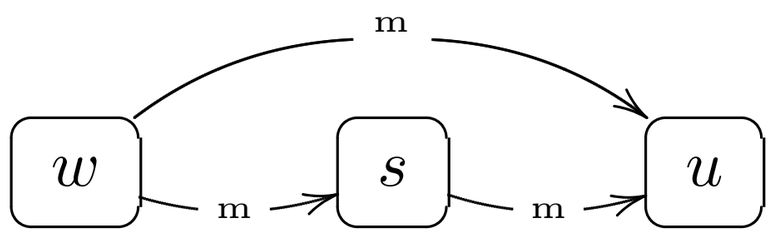


Order-dependence: counterexample

The priority merge of the ordering



with the ordering



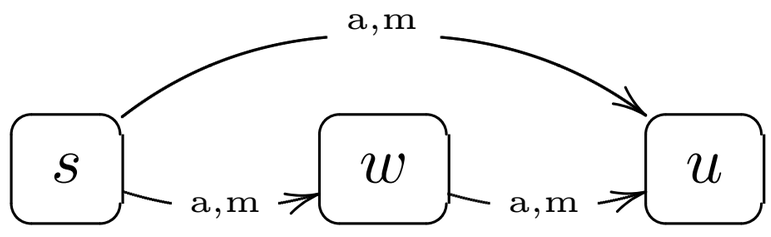
is equal to either of the two orders (depending on which agent has priority). But...

... suppose we have the following public dialogue

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

This *respects the "sincerity" rule* of our protocol, since initially m strongly believes u; then after the first upgrade a strongly believes $u \vee w$.

But this *doesn't respect the "order" rule*: m lets a answer before she finishes all she has to say. The resulting order is neither of two priority merges:



The Power of Agendas

All this illustrates the **important role of the person who "sets the agenda"**:

the "Judge" who assigns **priorities to witnesses' stands**;

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**.