Dynamic Epistemic and Doxastic Logics

Sonja Smets, ILLC, Amsterdam

(Slides are based on joint lectures with A. Baltag, ILLC)

Financial Support Acknowledgement:

European Research Council erc

NWO Netherlands Organisation for Scientific Research

PLAN OF THIS COURSE

1. **Puzzles. Logics of knowledge and belief. Epistemic and Doxastic models.**
2. **Core of Standard ("Hard") Dynamic-Epistemic Logic**: Public and Private announcements. Event models. The Product Update Mechanism.
3. **Belief Revision**: Plausibility Models. Conditional belief. Belief Upgrades. Doxastic event models and the Action-Priority Rule.
4. **Further Topics in the last three lectures**: Iterated Belief Revision. Belief Merge. Collective Learning. Informational Cascades. Surprise Examination Paradox etc.

2.5. Cheating and the Failure of Standard DEL

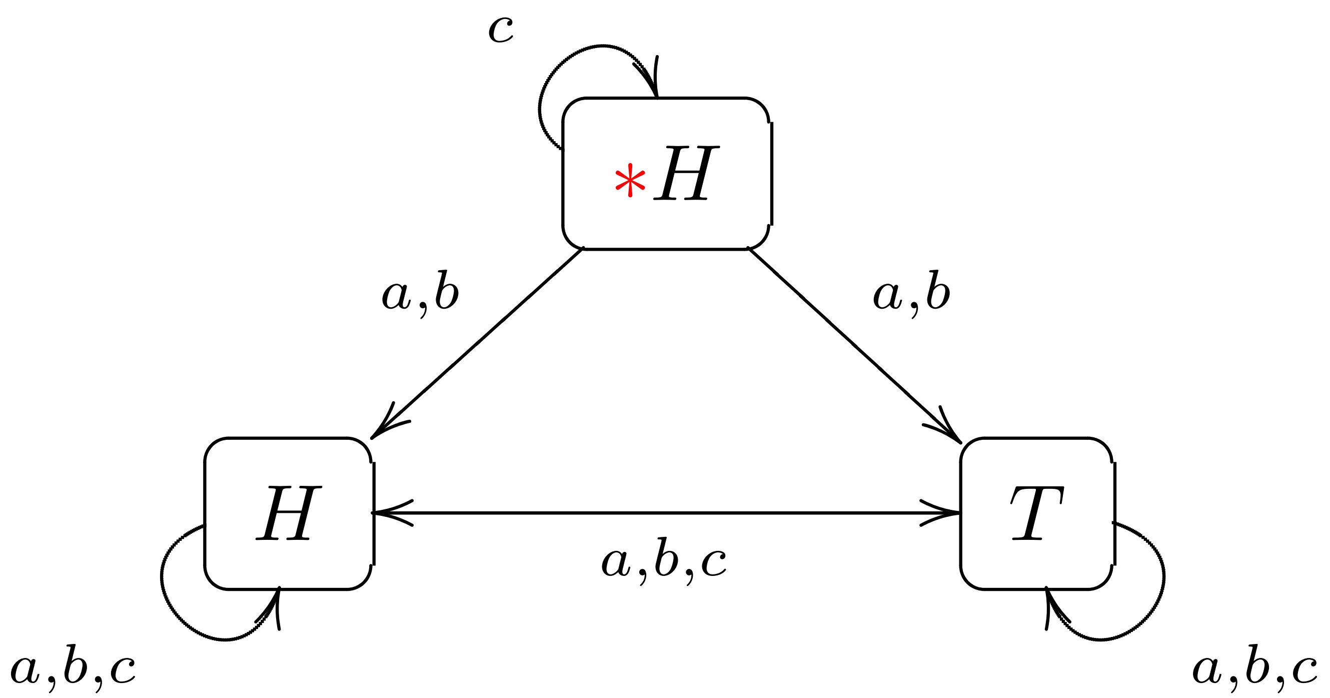
Our update product works very well when dealing with *"knowledge"*, or even with *(possibly false) beliefs*, **as long as these false beliefs are never contradicted by new information**.

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

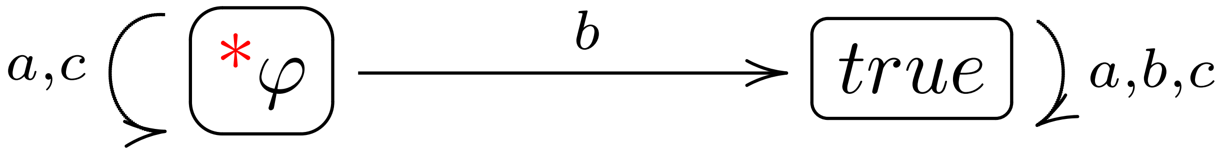
In Scenario 5 (happening after the cheating in Scenario 4), agent c sends a secret announcement to his friend a (**who has not suspected any cheating** till now!), saying:

“I **know** that H ”.

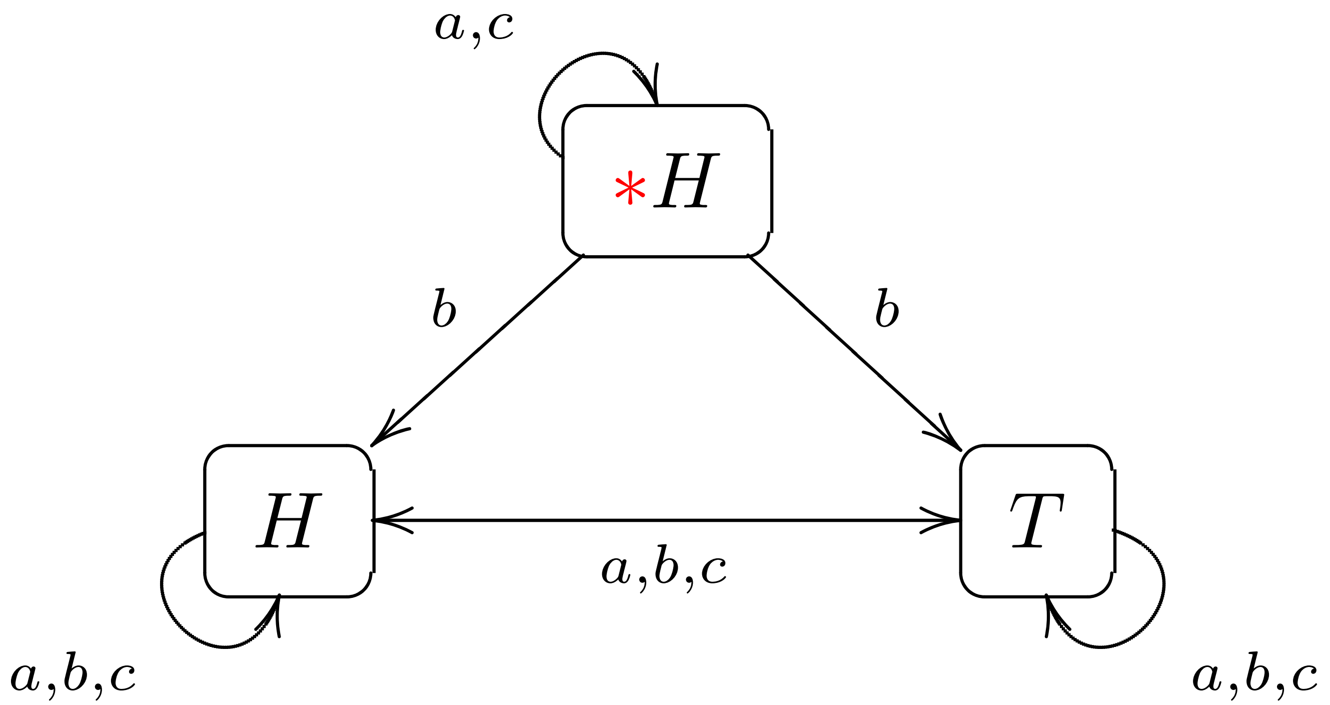
This is a fully **private communication** $!_{a,c}\varphi$ (from c to a) of the sentence

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

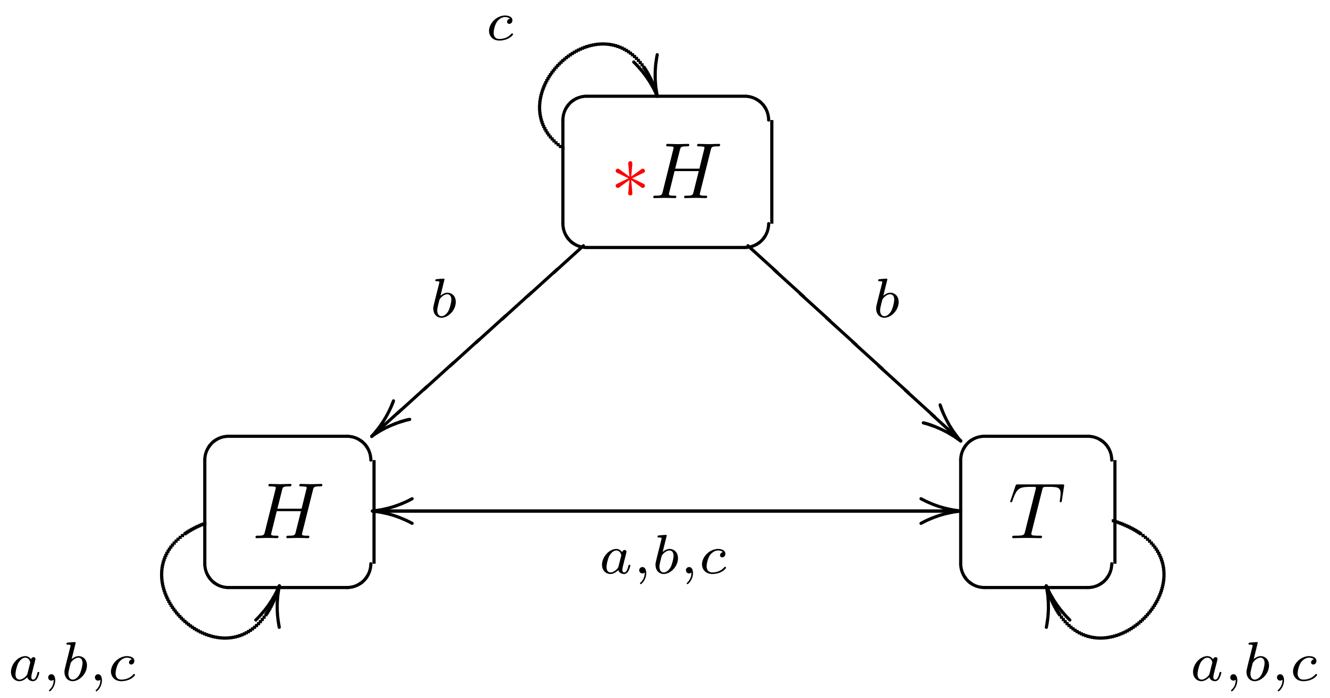
i.e. with event model



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**.

3.1. 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.**

3.2. 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

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

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:

$$B_a\, genius$$

$$K_a(genius \vee drunk)$$

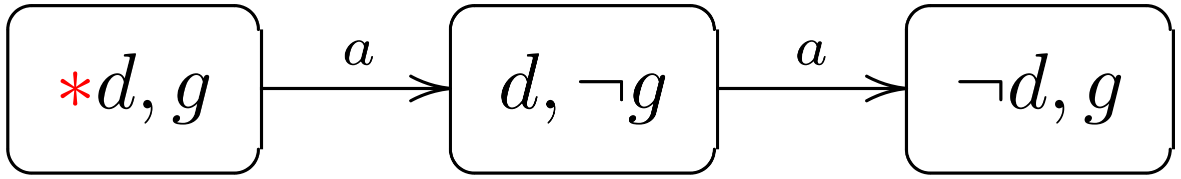
$$B_a \neg drunk$$

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

$$drunk \wedge genius$$

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:

$$B_a\varphi \Rightarrow B_aB_a\varphi, \qquad B_a\varphi \Rightarrow K_aB_a\varphi,$$

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

“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**.

3.3. 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: $\vdash B_a^\theta \varphi \Rightarrow K B_a^\theta \varphi$

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

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.

3.4. “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:

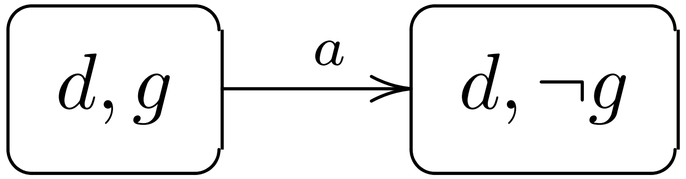
- **Update**: an **infallible** source. The source is *"known" (guaranteed) to be truthful.*

- **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).

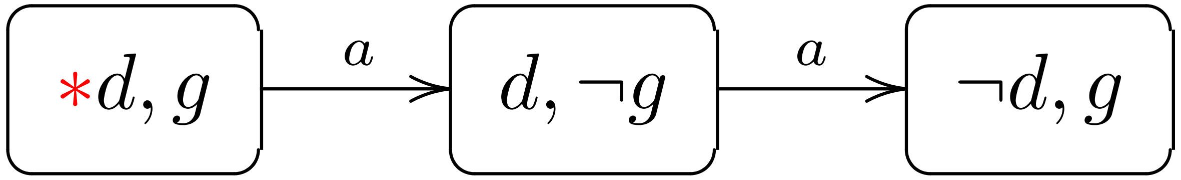
So Mary Curry tells Albert:

"*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

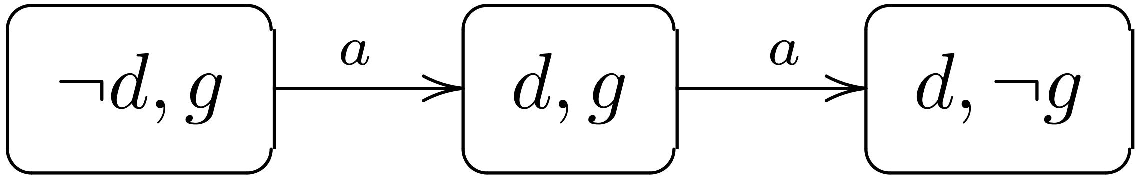


to capture Albert's new beliefs?

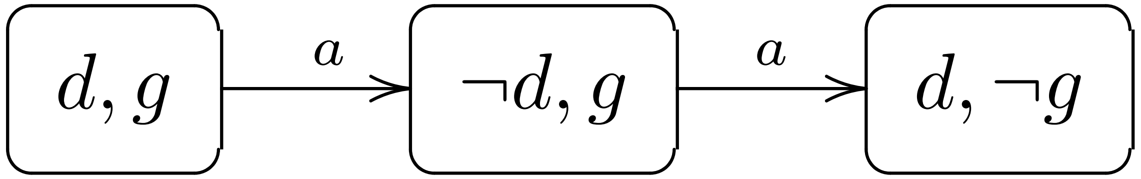
There are two drunk-worlds (d, g) and $(d, \neg g)$. **Which one should we promote ahead of all the others?**

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 **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**!

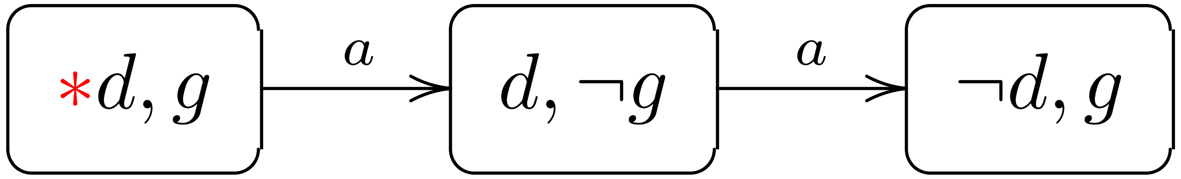
More precisely:
φ is **strongly believed** iff φ **is believed and is also conditionally believed given any new evidence (truthful or not) EXCEPT if the new information is known to contradict** φ; i.e. if:

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).

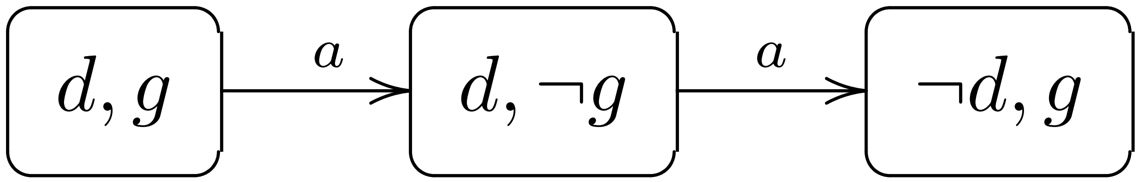
Radical Upgrade

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.

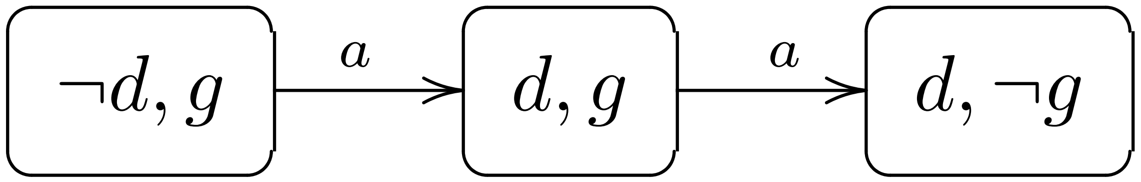
This corresponds to **radical upgrade**: it involves a rather radical revision of the prior beliefs, based on a strong belief in the correctness of the new information.

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

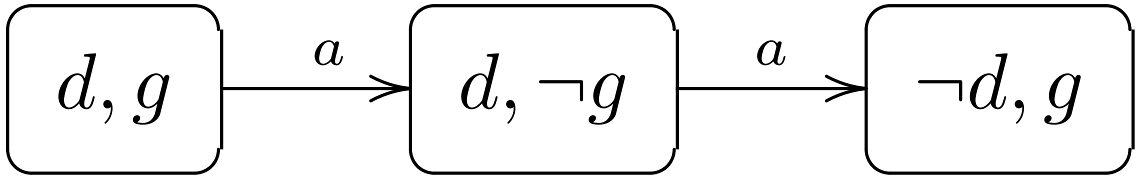
In this case, Albert will have to choose the second option: **promote only the most plausible d-world, leaving the rest the same**.

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).

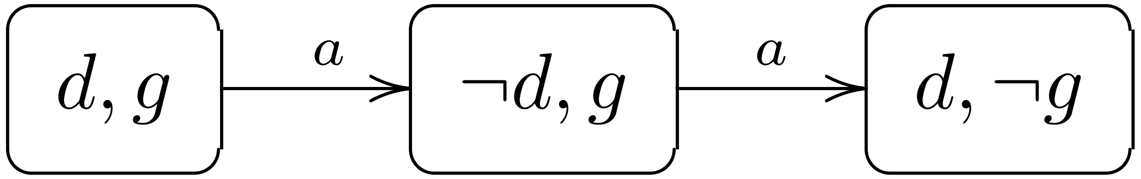
This corresponds to **conservative upgrade**.

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**:

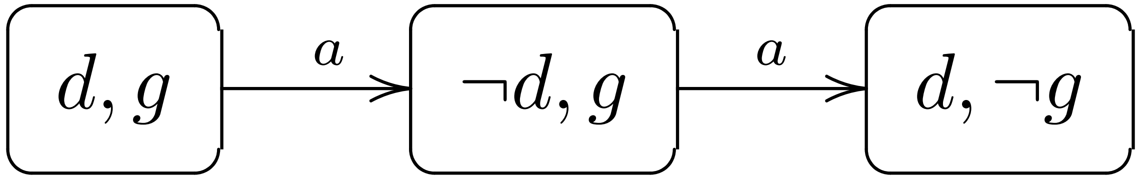
Indeed, after a *conservative or a radical upgrade*, the agent only comes to **believe** that φ (was the case), **UNLESS he already knew** (before the upgrade) that φ was **false**; i.e. we have the validity

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

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

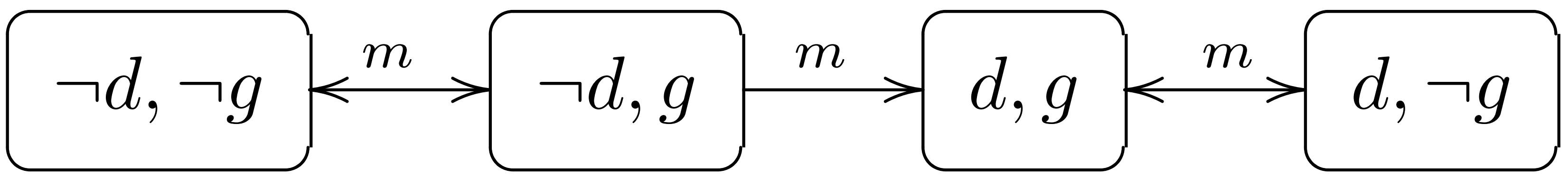
Albert Winestein's best friend is Prof. Mary Curry.

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

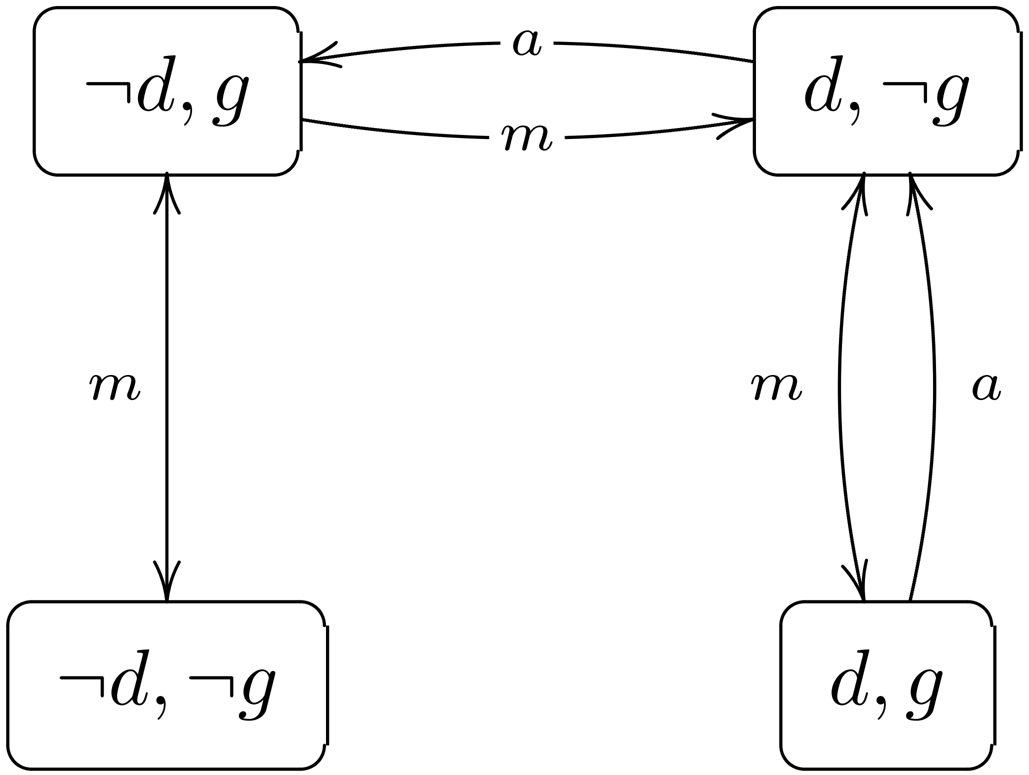
She's **completely indifferent with respect to Albert's genius**: she considers the possibility of genius and the one of non-genius as equally plausible.

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

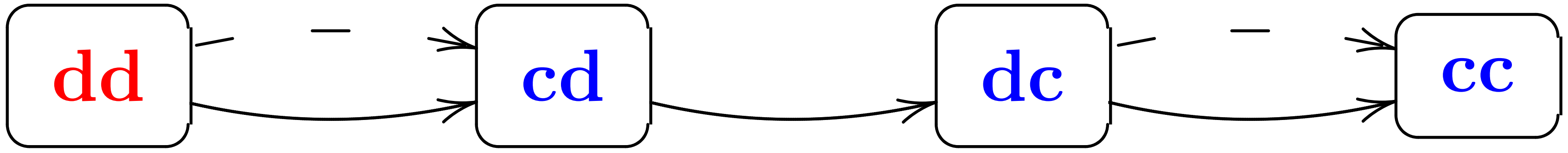


Muddy Children Example

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



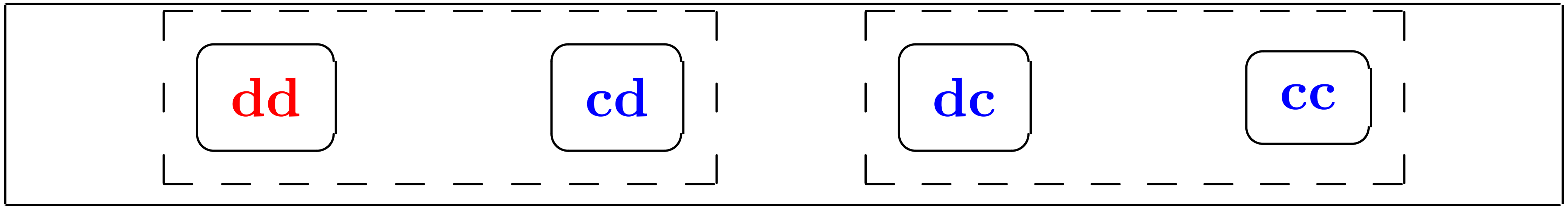
Dotted arrows: child 1's plausibility.

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.

3.5. 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*!

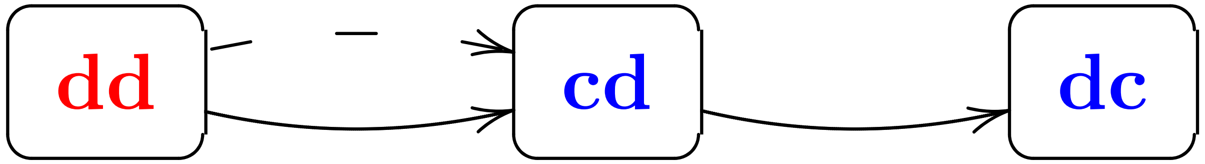
Muddy children example: A Joint Update

The Father announces:

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

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

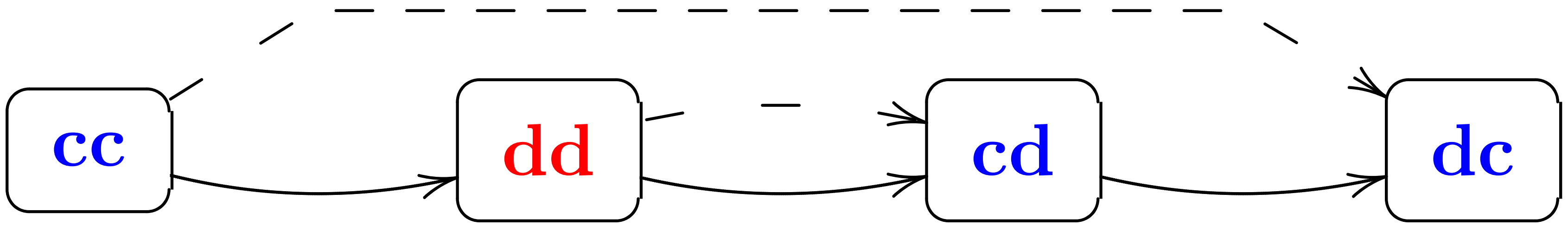
So this is an **update** $!(d_1 \vee d_2)$, yielding the updated model:



Muddy children example : Joint Radical Upgrade

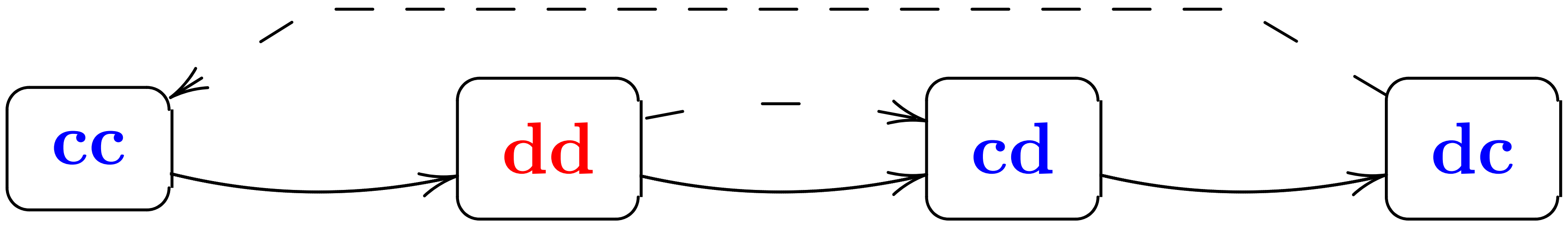
Alternatively, an older sister announces: "*At least one of you is dirty*". She is a **highly trusted source, though not infallible**:

This radical upgrade yields:



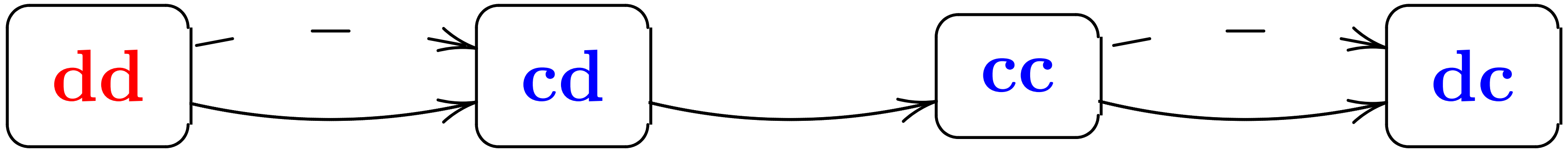
Children example: “Publicly Announced” Private Upgrade

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*:



3.6. 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.

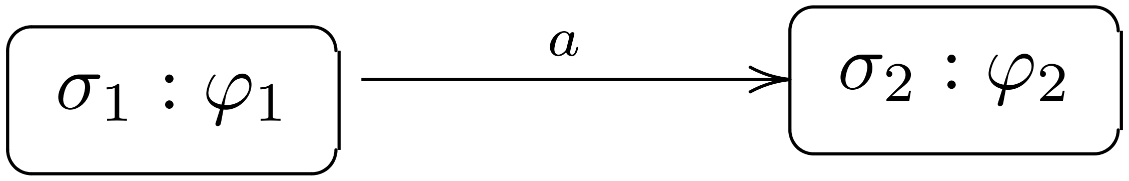
A number of other such rules were proposed and discussed by H. van Ditmarsch.

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:

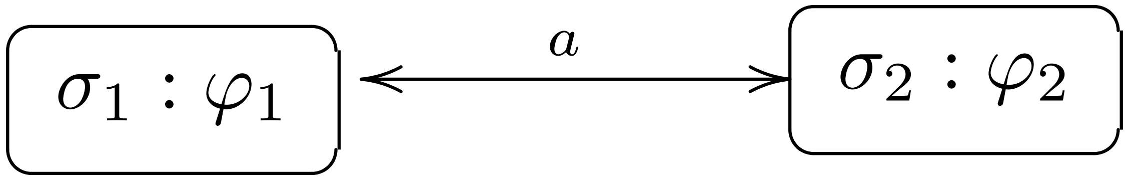
$$s_1 \sim_a s_2 \text{ and } \sigma_1 <_a \sigma_2 \text{ imply } (s_1, \sigma_1) <_a (s_2, \sigma_2),$$

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*.

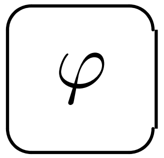
Interpretation

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

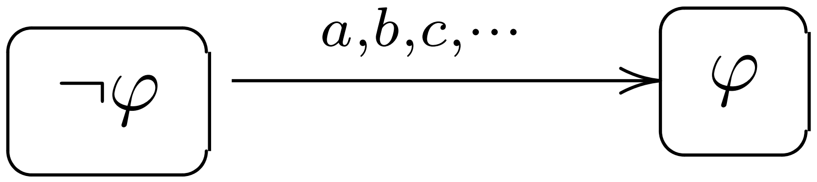
The event model for a joint radical update $!\varphi$ is essentially the same as in standard DEL (the event model for a "public announcement"):



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

EXAMPLE: joint radical upgrade

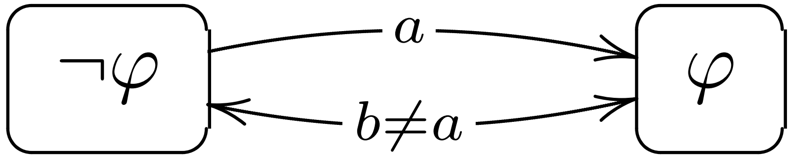
The event model for a joint upgrade $\Uparrow \varphi$ is:



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

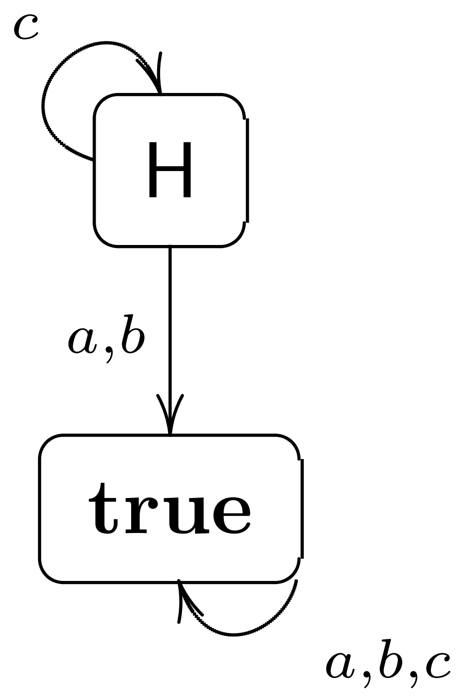
The event model for a publicly-announced private (radical) upgrade with φ is:



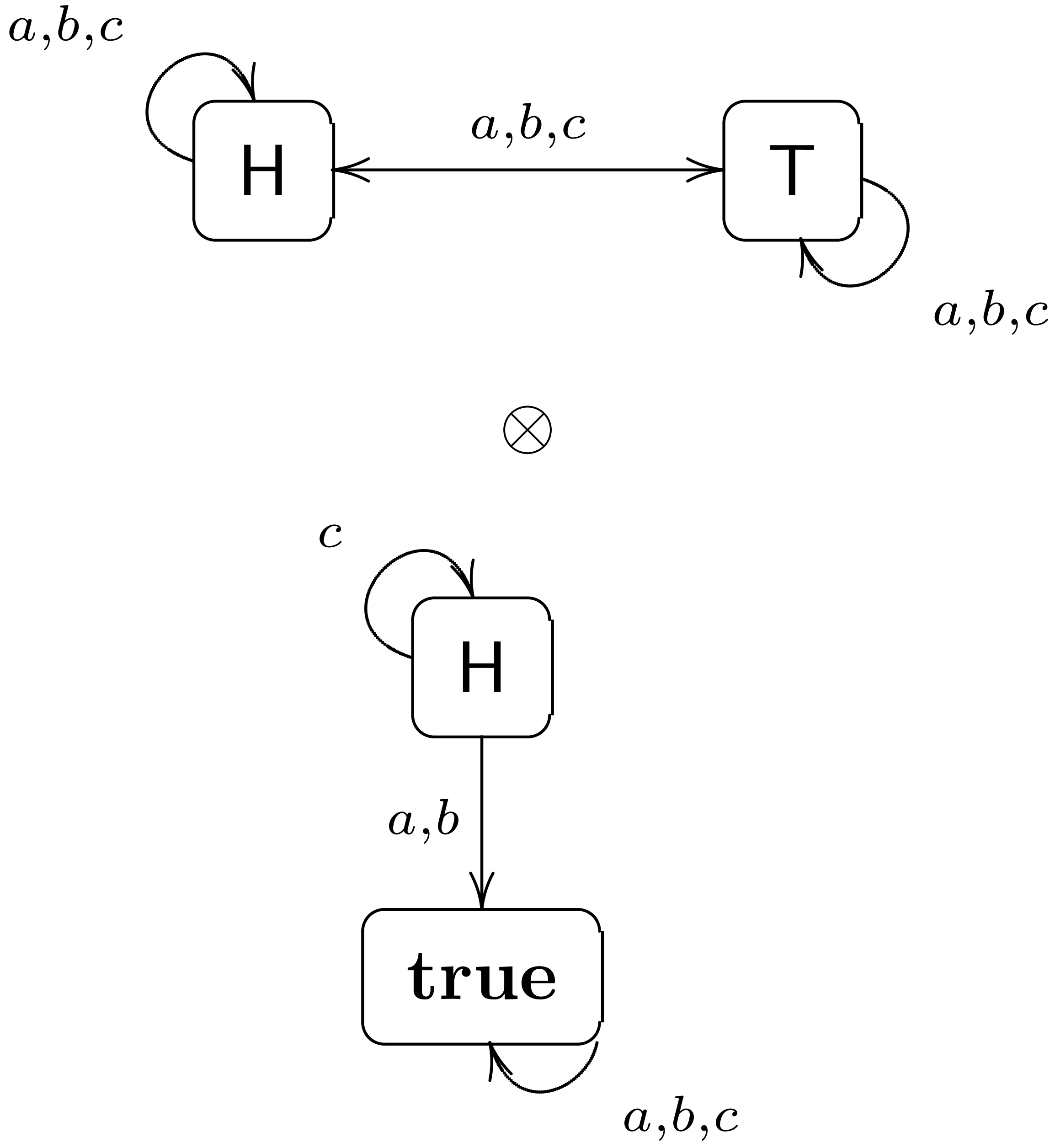
Example: Secret (Fully Private) Announcement

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.*

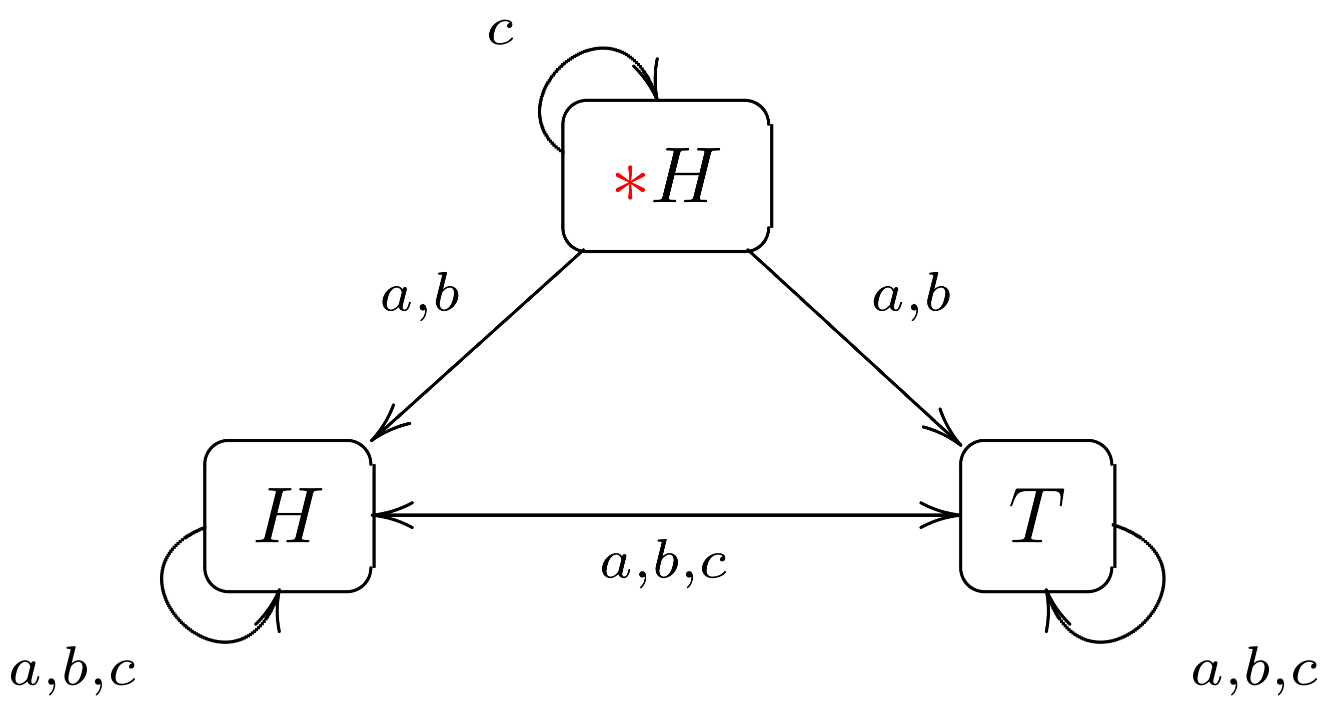
The DEL event model for this action was



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.

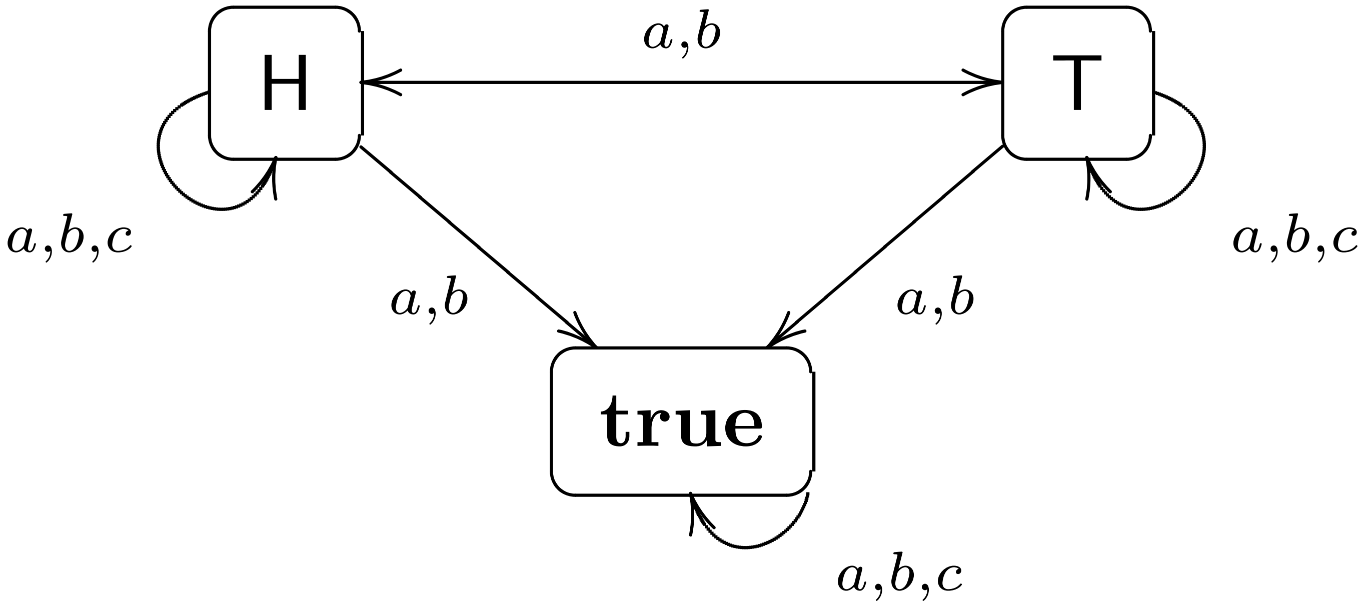
However, this is NOT the correct PLAUSIBILITY model for the new situation: it does NOT correctly reflect the agents' CONDITIONAL beliefs after the cheating.

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!

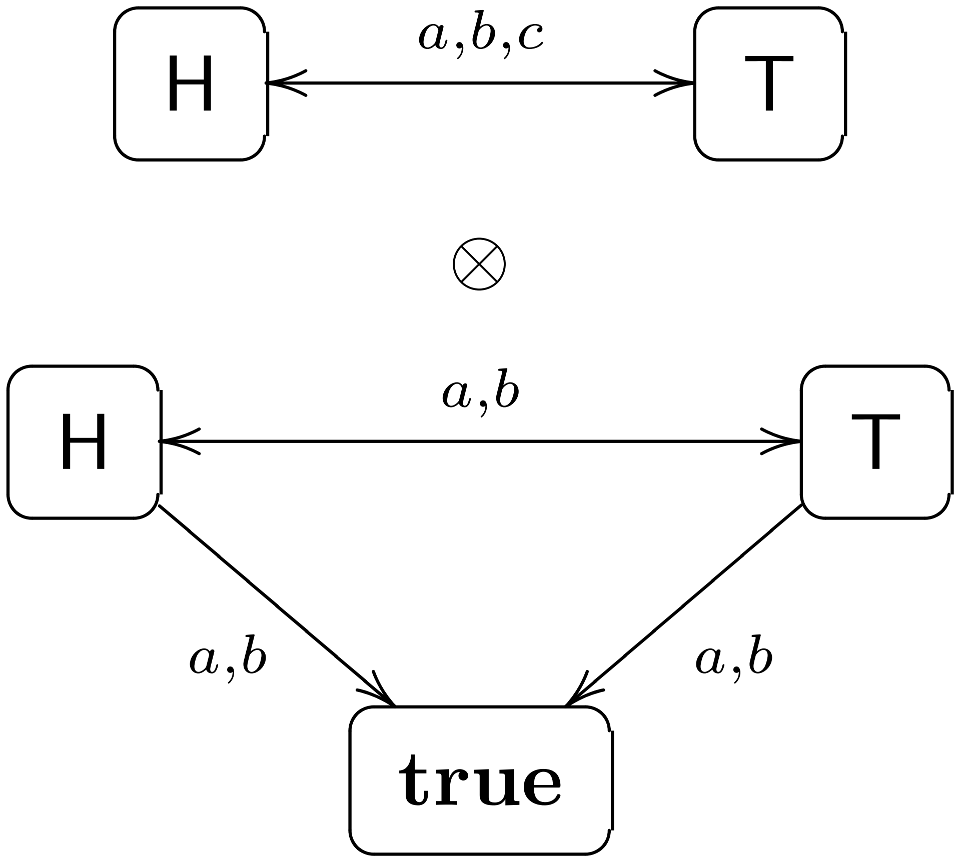
To compute the correct plausibility model, we need first to figure the correct event plausibility for the above action. For this, we still need to ask: *what does this event tell Alice (a) and Bob (b) about the face of the coin in case Charles (c) took a peek*?

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?

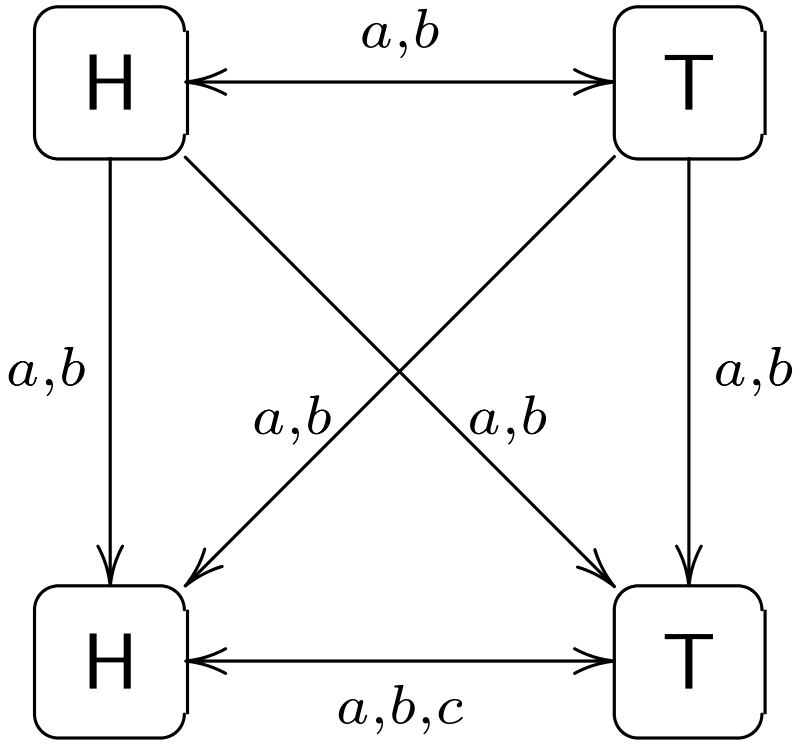
Clearly, this event *doesn't carry ANY new information* for Alice and Bob, so she should stick with whatever she believed before about the coin. Hence, the event model is



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



gives us:



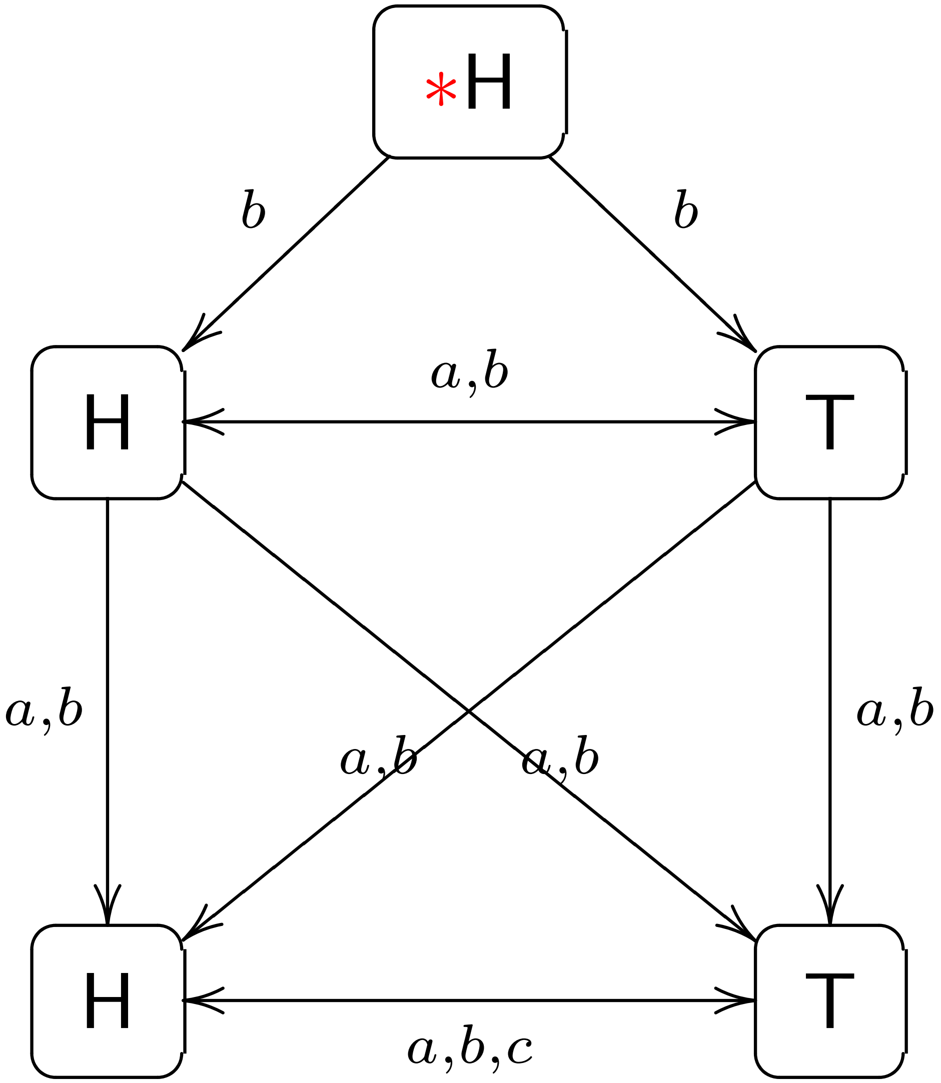
So e.g. a still believes that c doesn't know the face. However, if later she's given the information that he took a peek (without being told what he saw), she'd know that he knows the face; but as for herself, she'd still consider both faces equally plausible.

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!

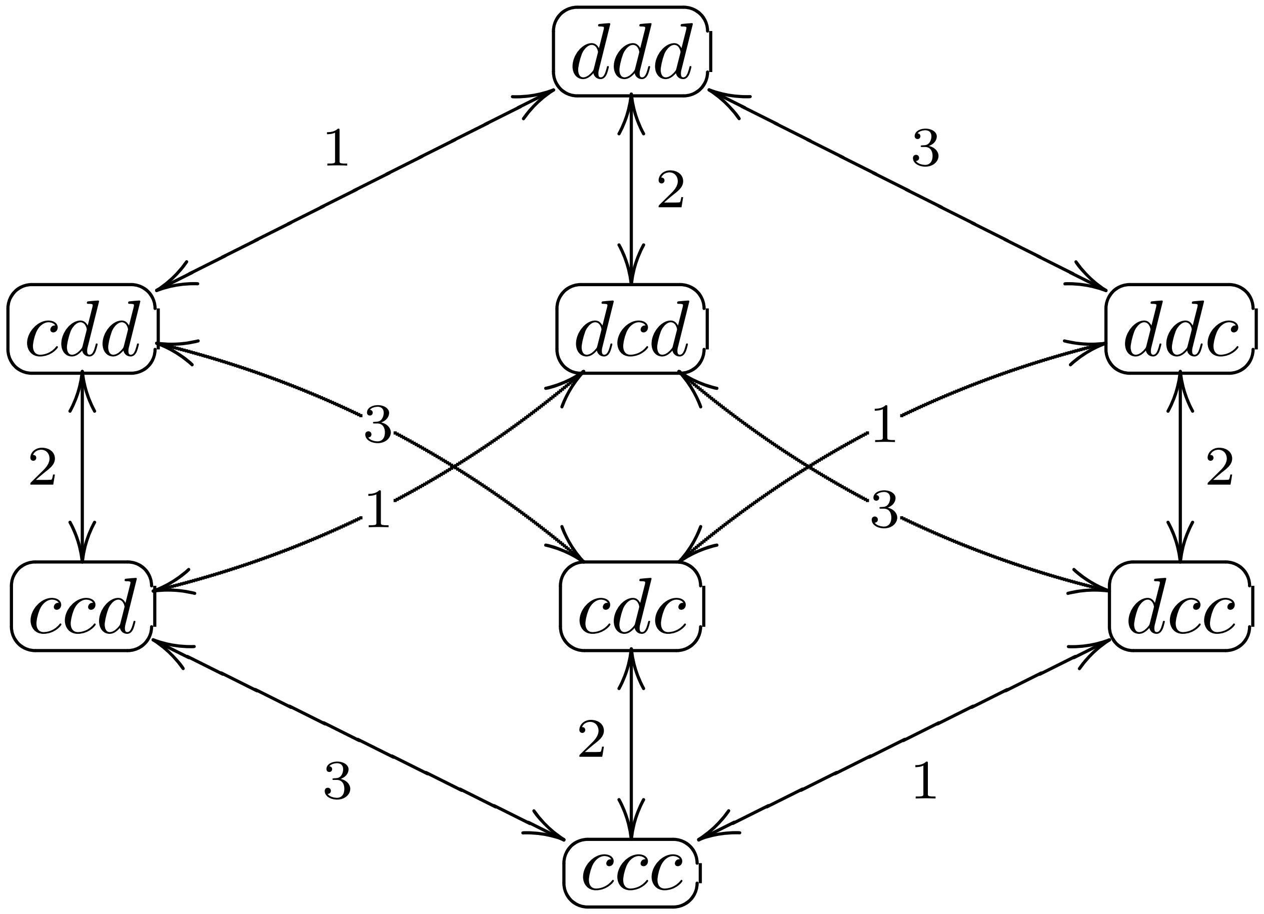
Now, things are better. The real world (in which Charles knows H) is still epistemically possible for Alice. So after the fully private announcement $!_a(K_c H)$, the plausibility model simply becomes:



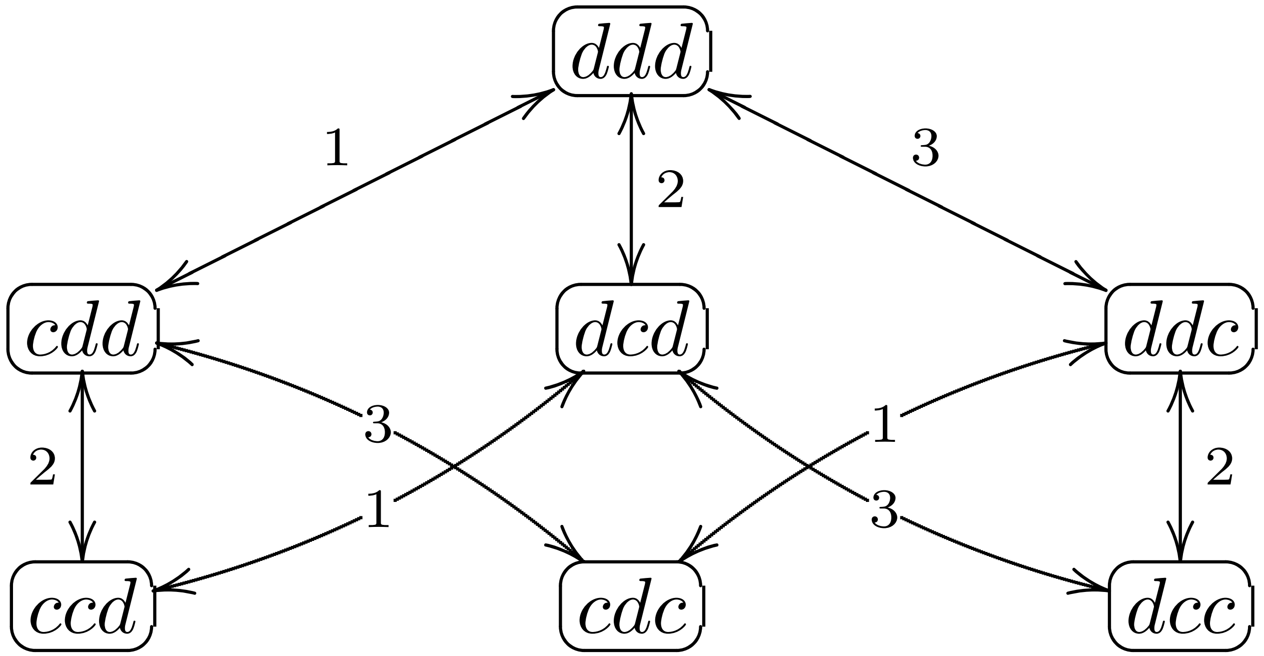
Exercise: what is the event model that gave us this plausibility model?

Solving the standard "Muddy Children"

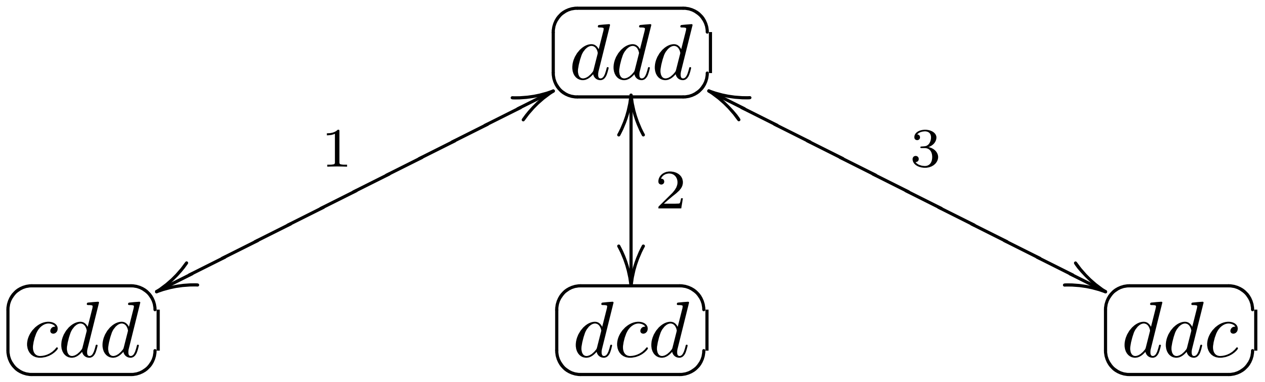
Three children, child 1 and child 2 are dirty. Originally, assume each child considers equally plausible that (s)he's dirty and that (s)he's clean:



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:



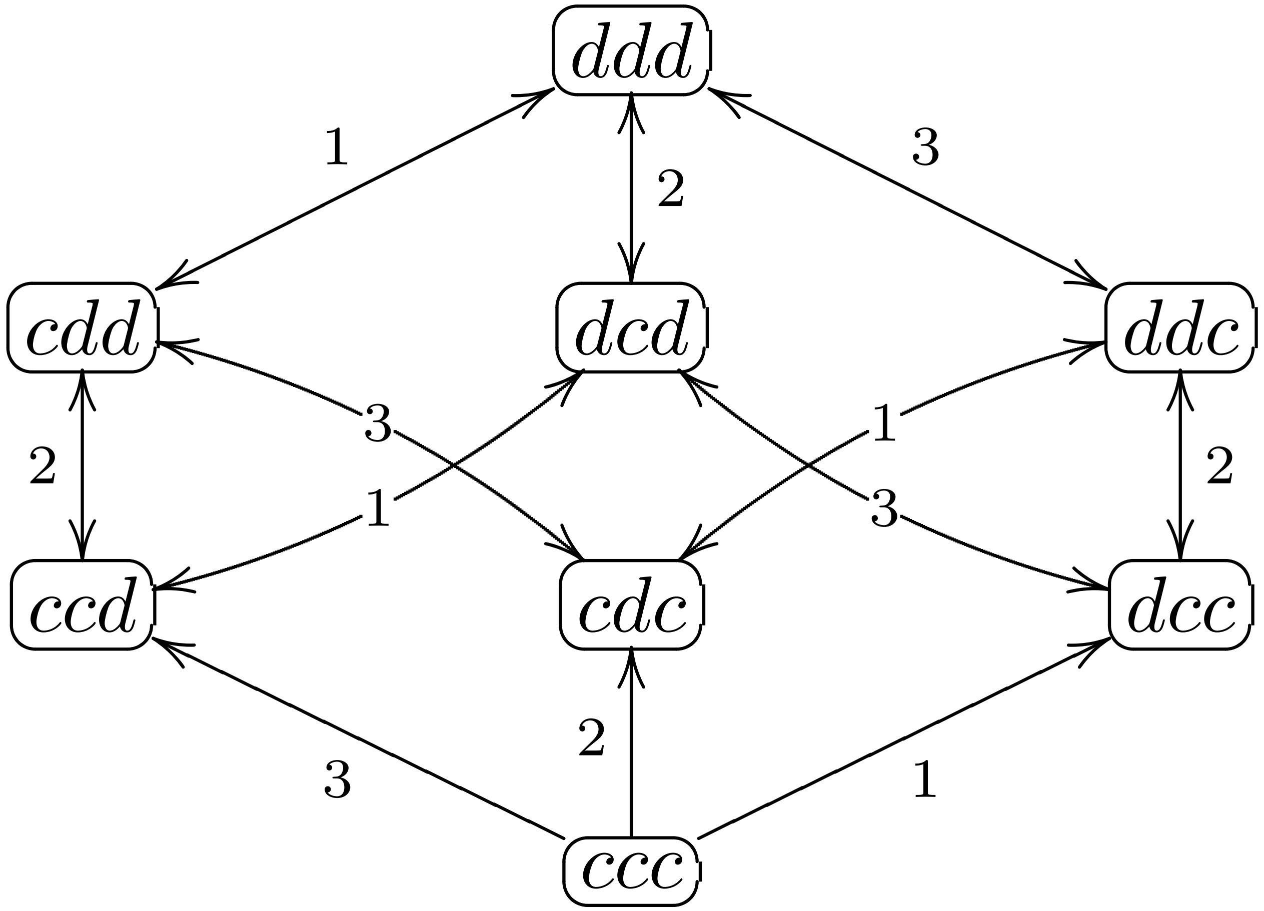
If the children answer “I don’t know I am dirty”, and they are infallible, then the update $!(\bigwedge_i \neg K_i d_i)$ produces:



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:

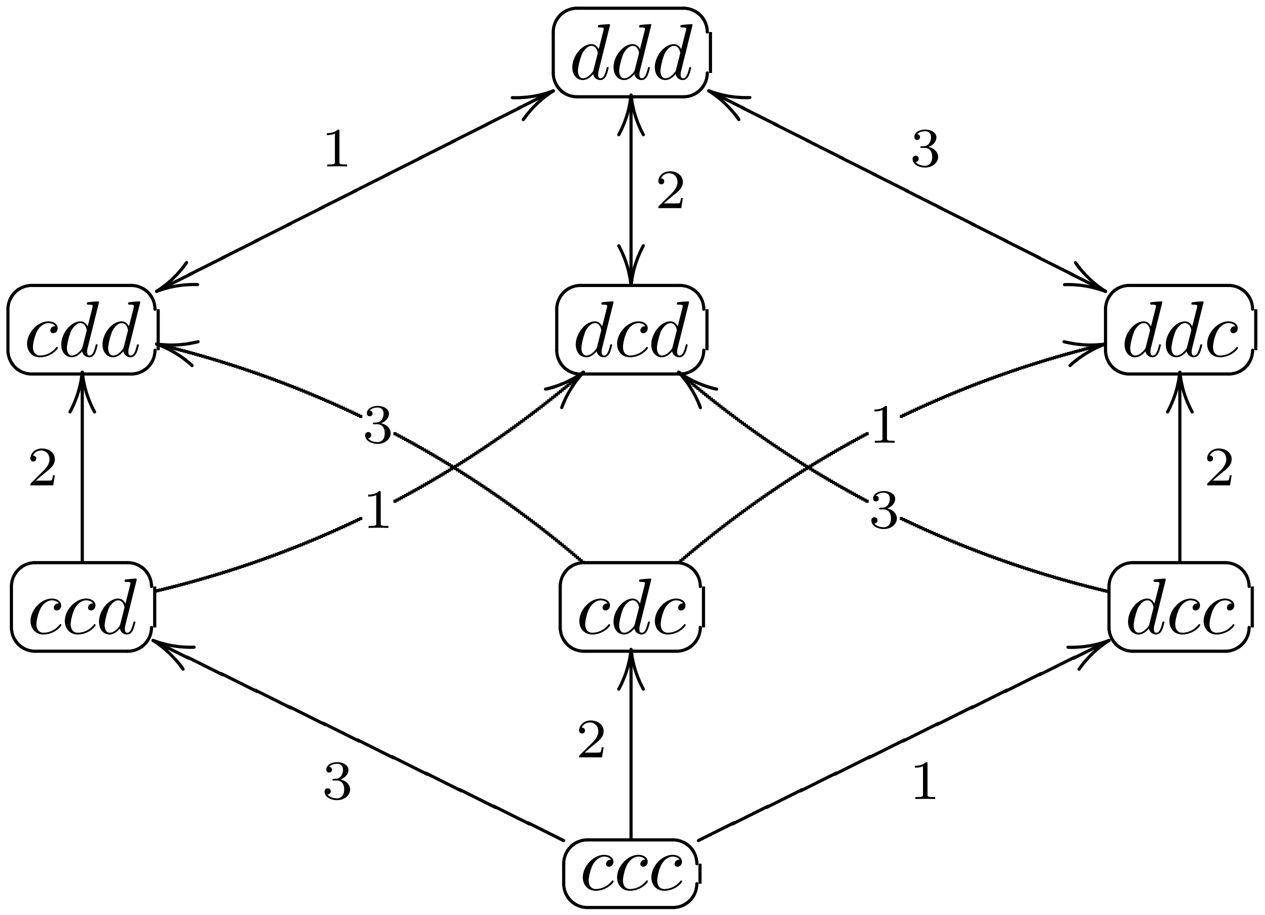


Do you believe you’re dirty?

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*?

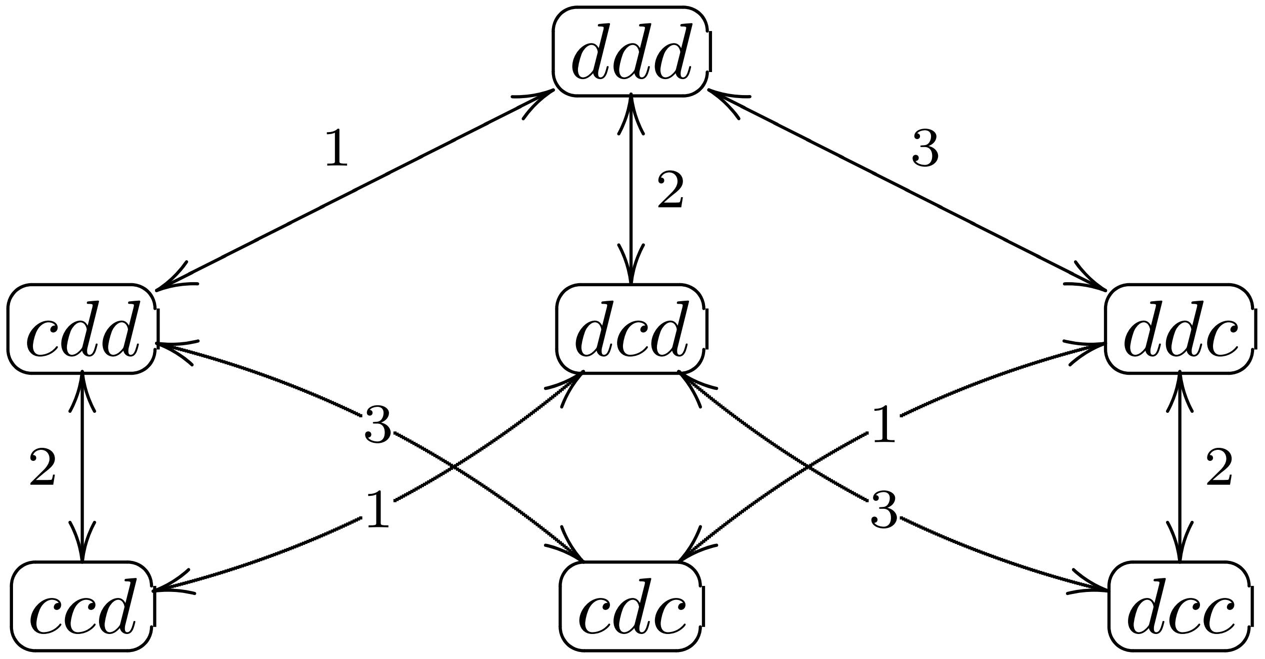
After a (radical or conservative) upgrade with the sentence $\bigwedge_i \neg B_i d_i$, we obtain:



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

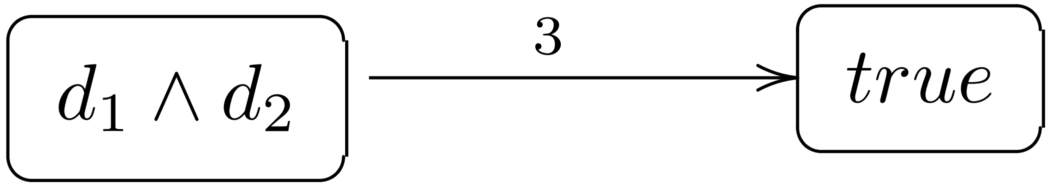
Cheating Muddy Children

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!