Lecture 4.1. Plausibility Models for Belief Revision

chapter "**Epistemic Logic and Information Update**", in
Handbook of Philosophy of Information

False Beliefs and The Problem of Belief Revision

If I start with a false belief, how can ever get rid of it??

Updates (including Product Update) only add NEW BELIEFS, but never erase the old ones!

If an agent *first believes* φ, and then *learns that he was wrong* (i.e. that $\neg\varphi$ was true), performing an update $!(\neg\varphi)$ will NOT help:

it will simply **delete all his possible worlds** (from his accessibility relation), and we end with a **non-doxastic (non-serial) model**, in which our agent believes EVERYTHING (including contradictions).

Our agent got "crazy": instead of correcting his false belief, he got **inconsistent beliefs**!

Example 1: The Coin

Agent a flips a coin, and takes a quick glimpse at its upper face, before covering it. He **believes he saw the coin lies Tails up** (though in fact in fact he is wrong). The doxastic model is

Now he uncovers the coin: **it is Heads up**!

What does he do?! What does he believe now?

The update !H gives a *non-serial (hence, non-doxastic) model*:

In such a model, $\Box_a false$ is true, so a believes a contradiction!

Example 2: The Surprise Exam

The students in Comp Logica class **know** that *there will an exam (and only one exam) in one of the 5 working days of next week.* It's Exam Week!

But the Professor wants to *keep the exam a surprise*: the day of the exam is NOT announced. Instead, the students are asked to come to class every day (in the morning, at 9:00 a.m). If there is no exam that day, they can leave after 5 minutes. If not, then... the exam starts!

Based on past experience, the students **believe** that *the exam will be either on Monday or on Tuesday* (it always was in one of these. But they don't know for sure!

EXERCISE **Draw a doxastic model, with 5 possible worlds, representing the students' beliefs**.

An Update

Let atomic sentences i (with $1 \leq i \leq 5$) mean that *the exam will be in day i of the week*.

Suppose that students come to class on **Monday**, but **they are dismissed** after 5 minutes. So they learn that atomic sentence 1 was *false*: this is a *truthful public announcement* $!(\neg 1)$.

Draw the updated model. **What do the students believe now** (*after* they were dismissed)?

The students come again to class on **Tuesday**, but **they are dismissed again** after 5 minutes. This is a new *truthful public announcement* $!(\neg 2)$.

Draw the updated model. **What do the students believe now??**

Example 3: Cheating and the Failure of Product Update

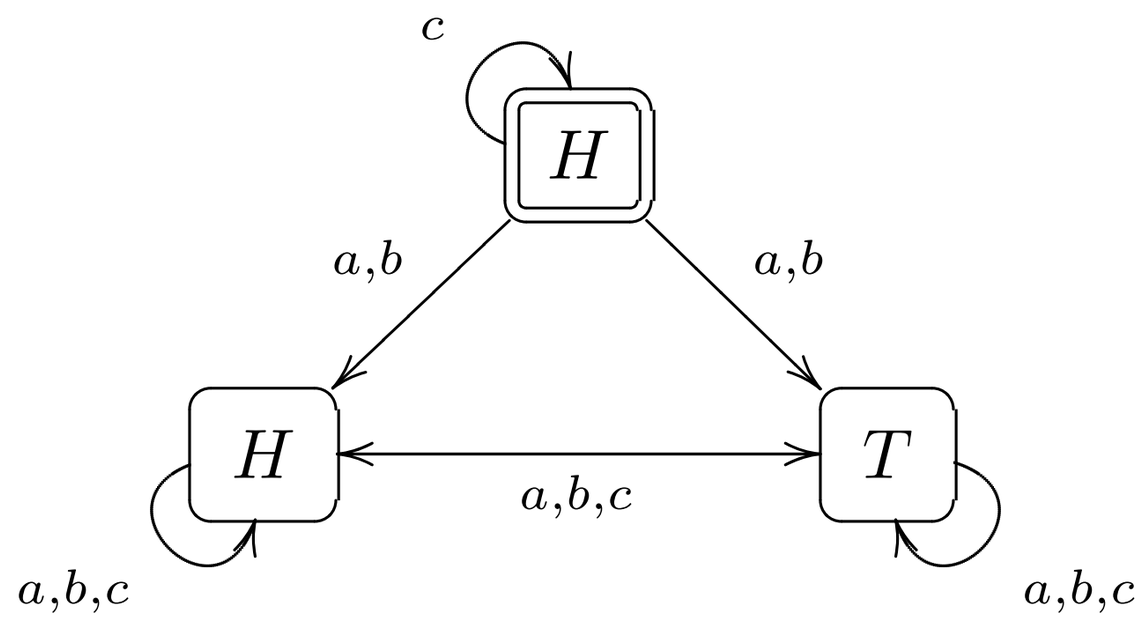
Like the update produced by public announcements $!\varphi$, our *product update* works very well when dealing with *"knowledge"*, or even with *(possibly false) beliefs*, **as long as these false beliefs are never contradicted by new information**.

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 *agent c secretly took a peek at the coin*, i.e. the output of Scenario 4:



So, now, c privately **knows** that the coin lies Heads up, but **the others believe he doesn't know**.

Counterexample Continued

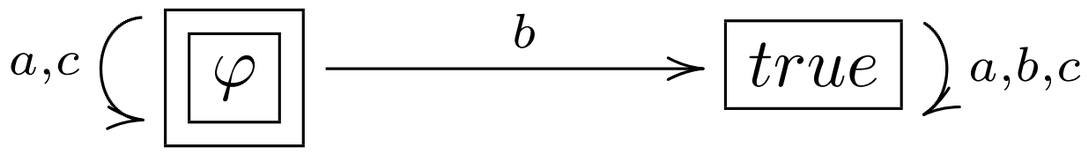
Suppose now that our Scenario 5 happens *after* the cheating in Scenario 4: agent c sends a secret announcement to his friend a (**who has not suspected any cheating** till now!), saying

"I **know** that H ".

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

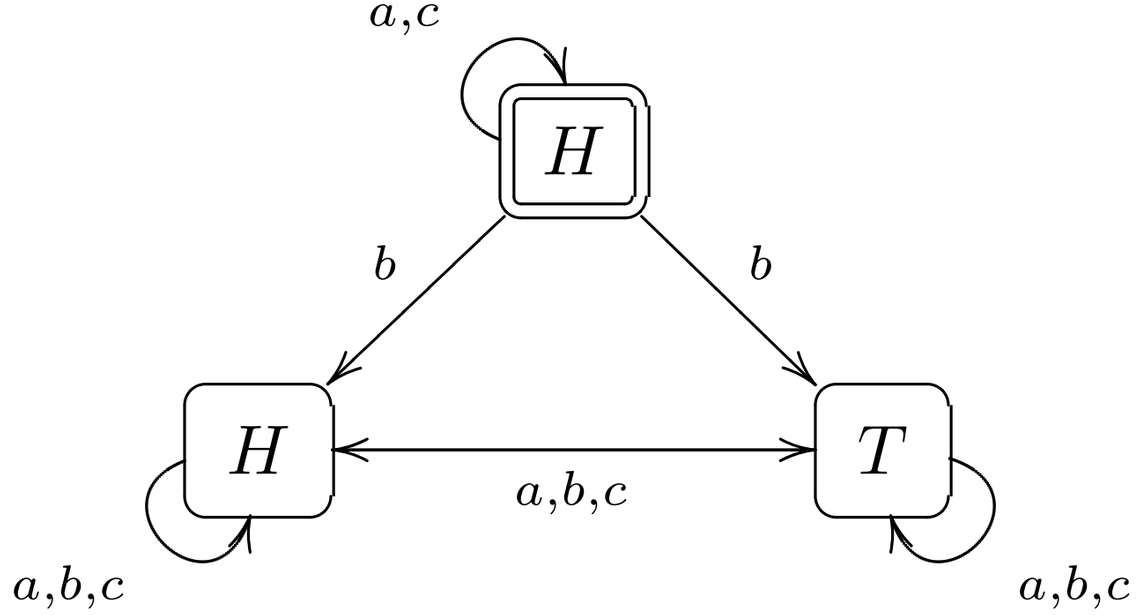
$$\varphi := K_c\mathsf{H},$$

i.e. with event model

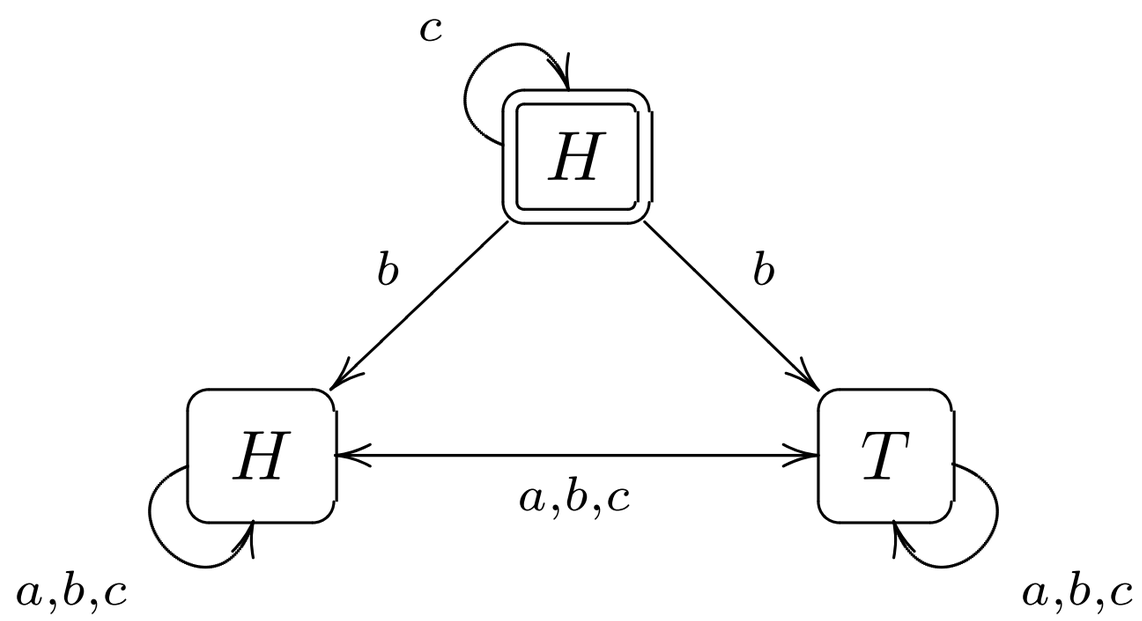


According to our intuition, after agent a reads the message, she knows everything that agent c: she knows it's Heads up, and she knows that agent c already knew that before this. Essentially, both a and c know the real world now.

So intuitively, the updated model for the situation *after* this private announcement SHOULD be:



However, the update product gives us:



This model is NOT serial: there are NO surviving a-arrows originating in the real world.

According to our semantics, *a will believe everything* after this: **agent a simply gets crazy**!

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

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: Newton, Gravitation and the Apple

And it gets worse!

I used to believe that **Newton was the first to discover the laws of gravitation after being inspired by being hit on the head by a falling apple**.

But then **I learned that this was a myth: this never happened to Newton!**

Now I'm a bit puzzled: it seems that my beliefs about the history of apples and Physics were based on myths.

Should I **continue to believe that Newton discovered gravitation, but that the discovery was not due to him being hit by an apple**?

Or should I rather stick with my **belief that the discovery of gravity was inspired by being hit by an apple**, and instead **just stop believing that it happened to Newton**?! Maybe it was actually Robert Hook the one who, after being hit by an apple, discovered gravitation, way before Newton! (In fact, Hook did claim something like this, by the way...)

Or I should maybe **give up both my beliefs**?! Maybe Hook did it, without any apple being involved.

Example Formalized

Let p be the sentence "*Newton discovered the law of gravitation*", and q be the sentence "*The discovery of law of gravitation was triggered by being hit by a falling apple*".

I used to believe p and q, and (by logical closure) also their conjunction $p \wedge q$. So my belief database was the following

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

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?! It all depends on which of them seems MORE PLAUSIBLE to me.

Is it *more plausible* that Newton discovered gravity or that the discovery was triggered by an apply? Or maybe these are *equally plausible*?

Judging this cannot be done by Logic alone: it depends on my knowledge of history of physics, but also maybe psychology, common sense etc.

Plausibility Models

A **(multi-agent) plausibility model**:

$$\mathbf{M} = (W, \leq_1, \ldots, \leq_n, \sim_1, \ldots, \sim_n, \nu)$$

- W a set of **possible "worlds"** ("states");
- $\mathcal{A} = \{1, 2, \ldots, n\}$ a (finite) set of **agents**;
- $\leq_a$ *preorders* (=reflexive and transitive relations) on W, on for each $a \in \mathcal{A}$: "**agent a's plausibility**" relation, meant to model *(revisable) beliefs*;
- $\sim_a$ *equivalence relations* on W: **a's possibility relation** ("*indistinguishability*") , meant to model *knowledge*;
- ν a valuation map for a set Φ of atomic sentences,

subject to a number of *additional conditions* (to be given later later).

Reading

Put $s < t$ iff: $s \leq t$ but $t \not\leq s$.

We read $s <_a t$ as saying that:
agent a considers world t is **"more plausible"** than world s.

$s \leq_a t$ is the non-strict version:
world t is **"at least as plausible"** as world s.

Information Partition and Information Cell

Each epistemic equivalence relation $\sim_a$ divides the possible worlds into disjoint sets, forming a **partition** of the state space into "cells" (= mutually disjoint sets of worlds):
two worlds w, s *belong to the same cell* iff $w \sim_a s$.

This is called **agent** a**'s information partition**.

For any world w, there is unique cell in a's partition that contains w, namely the equivalence class

$$w(a) := \{s \in W : w \sim_a s\},$$

which is also called w**'s information cell in agent** a**'s partition**.

The Conditions

The *conditions* required for plausibility relations are the following:

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

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

2. **"Comparability"**: **any two possible states are comparable (one is at least plausible as the other)**:

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

3. We consider W to be "**converse well-founded**": there NO infinite ascending chains

$$s_0 <_a s_1 <_a \cdots s_n <_a \cdots$$

of "better and better" (more and more plausible) worlds.

Most Plausible States

NOTE 1. If W is **finite** then the last condition is NOT NEEDED: **any** strict relation $<_a$ on finite W is converse well-founded.

NOTE 2. Essentially, converse-wellfoundedness amounts to requiring that **in every (non-empty) set of worlds there are some MOST PLAUSIBLE ones**:

i.e., if $P \subseteq W$ **is non-empty** set of possible worlds, then the set of "*most plausible P-worlds*"

$$best_a P = Max_{\leq_a} P := \{w \in P : w \not<_a s \text{ for any } s \in P\}$$

is also nonempty.

Plausibility encodes Possibility!

Given these conditions, it immediately follows that **two worlds 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

$$(W, \leq_1, \ldots, \leq_n, \nu),$$

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

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

This means we have to **draw only arrows only for the plausibility relations**: the knowledge/indistinguishability relations can be recovered from them!

Drawing Conventions for Plausibility Models

The plausibility arrows are *assumed to be ALWAYS REFLEXIVE and TRANSITIVE*, but for simplicity in the drawings **we skip all the loops and the arrows that can be obtained by transitivity** (by composing arrows of the same type).

An a**-arrow from the world** s to worlds t indicates that $s \leq_a t$ (i.e. **world** t **is at least as plausible as** s, or equivalently **world** s **is at most as plausible as** t).

This means that we have **one-directional arrows** going from worlds s to **(strictly) more plausible** worlds $t >_a s$.

It also means that we have **bi-directional arrows** between worlds that are **equally plausible** ($s \leq_a t \leq_a s$).

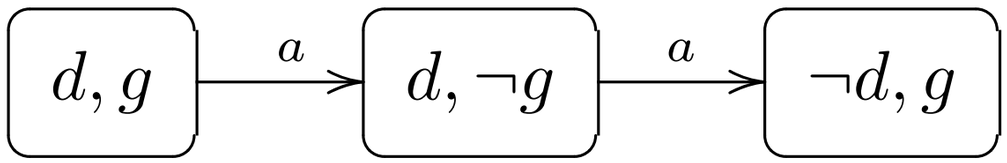
Example: Prof Winestein

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**.

The Plausibility 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) he's drunk or a genius, so we did **NOT** include any world $(\neg d, \neg g)$.

Knowledge

The **notion of knowledge** is defined as in epistemic models, as the Kripke modality for $\sim_a$:

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

EQUIVALENTLY:

$$\|K_a\varphi\|_{\mathbf{M}} = \{s \in W : s(a) \subseteq \|\varphi\|_{\mathbf{M}}\},$$

where recall that $s(a) = \{t \in W : s \sim_a t\}$ is s's "information cell" (in a's information partition).

Belief in plausibility models

Belief is now defined by quantifying over the **most plausible** worlds that are consistent the agent's information.

φ is believed (by agent a) at world $s \in W$ (in model $\mathbf{M}$) if **φ is true at ALL THE MOST PLAUSIBLE worlds in a's information cell**:

$$s \models B_a\varphi \quad \text{iff:} \quad t \models \varphi \text{ for all } t \in best_a s(a).$$

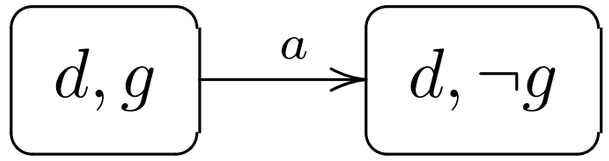
Equivalently:

$$\|B_a\varphi\|_{\mathbf{M}} = \{s \in W : best_a s(a) \subseteq \|\varphi\|_{\mathbf{M}}\}.$$

Updates in Plausibility Models

Suppose **new information** comes in (by a public announcement from an "infallible" source): say, *the Police does a blood test and shows Albert the result*. It definitely proves that **he is drunk**.

The test is a public announcement $!d$, which **deletes all the non-drunk worlds**. The resulting model is:



Now, Albert (correctly) believes (in fact, he **knows**) that *he is drunk*!
He has successfully revised his belief, without getting crazy!
(Unfortunately, he *also (wrongly) believes he's not a genius*!)

Conditional (Hypothetical) Beliefs

But, even BEFORE Police does the test, Albert can still reason **hypothetically**:

he can still believe that, **IF he is in fact drunk (contrary to his current belief) THEN he is not a genius**.

This "IF THEN" is NOT the same as usual (material conditional) $d \Rightarrow \neg g$:

Indeed, if φ is believed to be false, then $B(\varphi \Rightarrow false)$ holds, so $B(d \Rightarrow \psi)$ holds for EVERY ψ (including $B(d \Rightarrow g)$).

Conditional Beliefs as Contingency Plans

We can think of such a hypothetical belief $B^{\varphi}\psi$ as **"contingency" plan for belief change**:

(even though now I may believe $\neg\varphi$), in case later I find out that φ was the case, I will believe that ψ was the case.

Conditional Belief

At world $s \in W$, we say that ψ **is believed by agent** a **given** φ (or "*conditional on* φ"), and write $B_a^{\varphi}\psi$, if:
ψ **is true in all the most plausible** φ**-worlds in** a**'s information cell**.

Formally:

$$s \models B_a^{\varphi}\psi \quad \text{iff:} \quad t \models \psi \text{ for all } t \in best_a\left(\|\varphi\|_{\mathbf{M}} \cap s(a)\right).$$

Equivalently:

$$\|B_a^{\varphi}\psi\|_{\mathbf{M}} = \{s \in W : best_a\left(\|\varphi\|_{\mathbf{M}} \cap s(a)\right) \subseteq \|\psi\|_{\mathbf{M}}\}.$$

ALTERNATIVE NOTATION: $B_a(\psi|\varphi)$ instead of $B_a^{\varphi}\psi$.

Belief as a Special Case

It is easy to see that **simple (unconditional) belief B_a is a special case of conditional belief**:

Take *true* to be any tautological proposition (i,e., which is TRUE in all possible worlds, e.g. $p \vee \neg p$).

Then we always have:

$$B_a\varphi \;\Leftrightarrow\; B_a^{true}\varphi.$$

Knowledge as Another Special Case

But **knowledge K_a is also a special case of conditional belief!**

Take $false$ to be any contradictory proposition (i.e., which is FALSE in all possible worlds, e.g. $p \wedge \neg p$).

Then the following is valid in all plausibility models:

$$K_a\varphi \;\Leftrightarrow\; B_a^{\neg\varphi} false.$$

Equivalently:

$$K_a\varphi \;\Leftrightarrow\; B_a^{\neg\varphi}\varphi.$$

The agent believes ANYTHING given something that she KNOWS TO BE FALSE.

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" B_a is NOT the Kripke modality for the plausibility relation $\leq_a$.**

Nevertheless, belief (in a plausibility model) CAN be made into a Kripke modality by choosing another, appropriate accessibility relation.

Exercise:

Define binary an accessibility relation R_a on the set S in a plausibility model, such that the operator B_a is the Kripke modality for this relation; i.e. we must have

$$s \models B_a\varphi \quad \Leftrightarrow \quad \forall t(\, sR_at \Rightarrow t \models \varphi\,)$$

Solution: Belief as a Kripke Modality

In a plausibility model, we can define a **doxastic accessibility relation** R_a by putting:

$$sR_at \text{ iff } t \in best_a s(a).$$

Then *"belief" (as defined above) IS the Kripke modality for* R_a:

$$s \models B_a\varphi \text{ iff } \forall t\ (sR_at \ \Rightarrow\ t \models \varphi\,).$$

Plausibility models become doxastic $(KD45)$ **models, when endowed with this relations** R_a: so "belief" in these models still satisfies the axioms of Consistency and Positive and Negative Introspection.

Conditional Belief as a Kripke Modality

We can similarly define a **conditional doxastic accessibility relation** $sR_a^{\varphi}t$, given by

$$sR_a^{\varphi}t \text{ iff } t \in best_a\left(\|\varphi\|_{\mathbf{S}} \cap s(a)\right),$$

and then *conditional belief is a Kripke modality for this relation*:

$$s \models B_a^{\varphi}\psi \text{ iff } \forall t\,(sR_a^{\varphi}t \Rightarrow t \models \psi\,).$$

However, *the conditional doxastic arrows are not necessarily serial*: this means that **conditional belief is not necessarily consistent**.

More precisely: **the conditional doxastic arrows R_a^{φ} are serial ONLY if the condition φ is consistent with the agent's knowledge**, i.e. if $s \models \neg K_a \neg\varphi$ (equivalently, if $\|\varphi\|_{\mathbf{S}} \cap s(a) \neq \emptyset$).

Interpretation: Revision is restricted by Knowledge

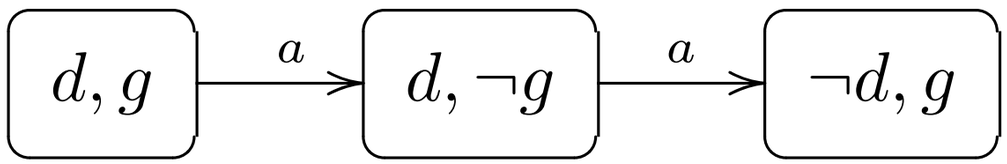
If φ is **known to be false** then the agent should **NOT be able to revise with** φ: revision only applies to beliefs that are not actually known.

This is reflected by the fact that *conditional beliefs may be inconsistent*: this means that the **corresponding revision fails**.

Indeed, if $K_a \neg\varphi$, then $B_a^{\varphi}\psi$ holds for **every** ψ

Examples

In the Winestein model



the following hold at all worlds:

$$K_a(d \vee g)$$

$$B_a(\neg d \wedge g)$$

$$B_a^d \neg g$$

The Exam Revisited

The students *believe that the exam is either Monday or Tuesday.*

If we ask them to be more specific, say they *cannot decide between Monday or Tuesday*: they have no reason to believe one or the other.

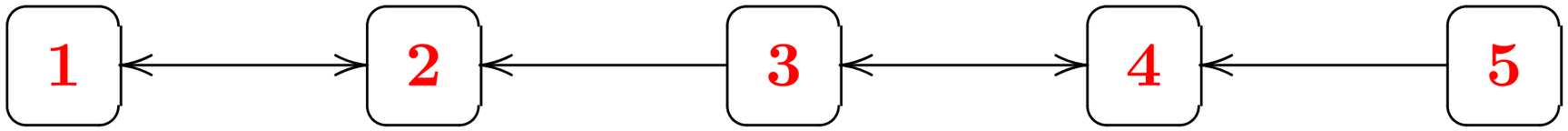
Now we ask them: "*But what if you are wrong? What if the exam is neither Monday or Tuesday? In that case, when will the exam be, according to you?*"

They answer "*In that case, I think it's either Wednesday or Thursday*"

(Why? Professor wants to keep the day of exam a surprise, scheduling it on Friday would spoil the surprise: by Thursday at noon, they will know it's on Friday!)

Again, if we ask them to be more specific, they *cannot decide between Wednesday or Thursday*: they seem *equally plausible.*

A Plausibility Model for the Exam Example



In this model, Monday and Tuesday are equally plausible, but both are more plausible than Wednesday and Thursday; which are themselves equally plausible, but more plausible than Friday.

Check that the following formulas are true:

$$B(1 \vee 2)$$

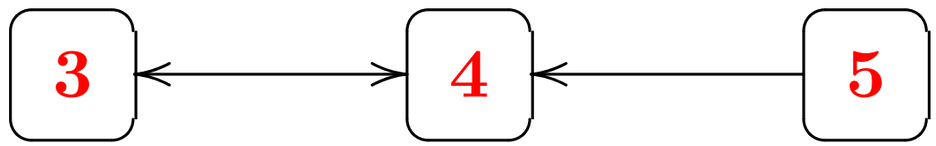
$$B^{\neg(1\vee 2)}(3 \vee 4)$$

$$B^{\neg(1\vee 2)\wedge\neg(3\vee 4)}5$$

Day-passing as Update

Represent the passing of each day i with no exam, as an **update (public announcement)** $!(\neg i)$, that deletes the non-i worlds.

Suppose **Monday and Tuesday pass with no exam**. After the updates $!(\neg 1)$, followed by $!(\neg 2)$, we get:



What do the students believe now?

Exam Example Continued

Suppose **Wednesday and Thursday also pass with no exam**. By performing $!(\neg 3)$ then $(!\neg 4)$, we get:

What do the students believe now?

NOTE: Recall that these are plausibility models, so the relations are in fact reflexive and transitive, and beliefs are given by quantifying over most plausible worlds. So, at each step, students have **consistent beliefs**! No contradiction!

Lecture 4.2. Plausibility Models, Safe Belief, Updates

chapter "**Epistemic Logic and Information Update**", in
Handbook of Philosophy of Information

Common Knowledge

Common knowledge Ck is defined *as in epistemic models*, by **quantifying** over all worlds that are accessible by any **concatenations of indistinguishability relations for any agents**:

$$s \models_{\mathbf{S}} CK\varphi \text{ iff } t \models_{\mathbf{S}} \varphi \text{ for every t and every finite chain}$$
$$\text{(of length } n \geq 0\text{) of the form } s = s_0 \sim_{a_1} s_1 \sim_{a_2} s_2 \cdots \sim_{a_n} s_n = t.$$

Common True Belief and Common Belief

Common true belief Ctb is similarly defined, by **quantifying** over all worlds that are accessible by any **concatenations of doxastic relations for any agents**:

$$s \models_{\mathbf{S}} Ctb\,\varphi \text{ iff } t \models_{\mathbf{S}} \varphi \text{ for every t and every finite chain}$$

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

where R_a are the *doxastic* arrows introduced above:
sR_aw iff $w \in best_a s(a)$.

Note that $Ctb\varphi$ implies φ, since the quantifier includes **the real world** (accessible from itself by a chain of length 0): that's why is called **common TRUE belief**.

If we restrict the definitions only to chains of any length $n \geq 1$, we obtain the notion of **common belief** Cb.

EXAMPLE: Adding Mary Curry to the Winestein story

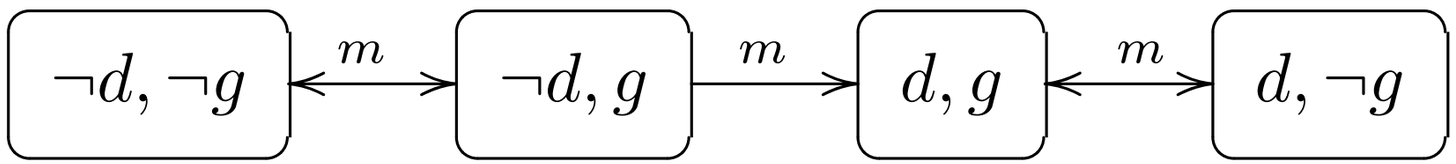
Albert Winestein's best friend is Prof. Mary Curry.

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 she could be wrong**: it is in principle possible that Albert is not drunk, despite the presence of the usual symptoms.

The model for Mary alone:



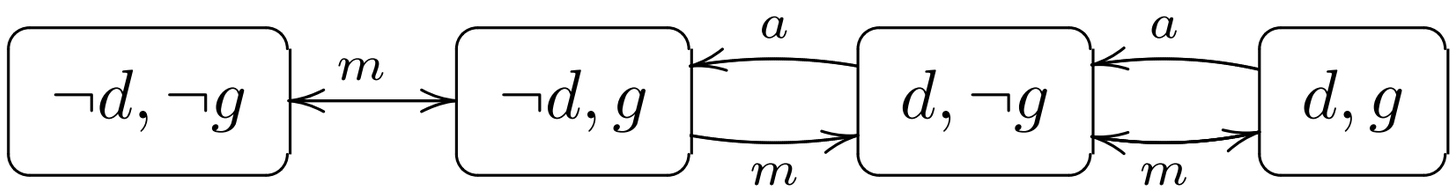
A Multi-Agent Model M

To *put together Mary's order with Albert's order*, we need to know *what do they know about each other*.

Let's now suppose that **all the assumptions we made about Albert and Mary are common knowledge**, **EXCEPT for the real world** (i.e. *whether or not Albert is really drunk, and whether or not he is really a genius*).

More precisely: **all Mary's opinions (knowledge, beliefs, conditional beliefs, as described above) are common knowledge**. It is also **common knowledge that**: *if Albert feels he's a genius, then he's either drunk or a genius; Albert knows what he feels (about being or not a genius); if Albert is drunk, then he feels is a genius; if Albert is a genius, then he feels he is a genius; if Albert feels he's a genius, then he believes he's a sober genius, but if he'd learn that he's drunk, he'd believe that he's not a genius.*

Then we obtain the following multi-agent plausibility model **M**:

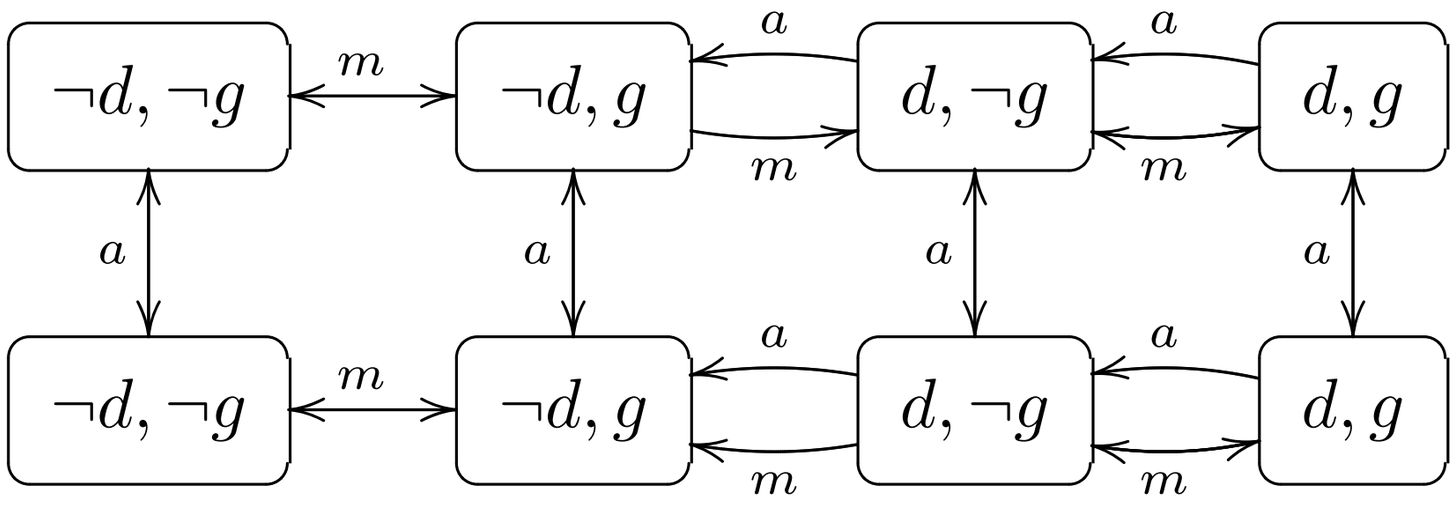


Relaxing the Assumptions: Another Multi-Agent Model

Let's relax our assumptions about agents' mutual knowledge: we now **drop** the assumption that Mary's opinions are common knowledge, while **keeping all the other assumptions**.

In addition, we now assume that **it is common knowledge that Mary has no opinion on Albert's genius** (i.e. that she *considers genius and non-genius as equi-plausible*), but that **she has a strong opinion about his drunkness**: she can see him, so judging by this she **either strongly believes he's drunk or she strongly believes he's not drunk**. (But her actual opinion about this is unknown to Albert, who thus *considers both opinions as equally plausible*.)

The resulting model is:



where **the real world** is represented by **the upper** (d, g) **state**.

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 any child is clean rather than dirty*, **whenever she has any doubt about that.**(So, of course, if one *knows* that the other is dirty then this plausibility assumption doesn't apply). 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.

Check that in the real world (d, d), we have:

$$(d, d) \models B_1(c_1 \wedge d_2) \wedge B_2(c_1 \wedge c_2)$$

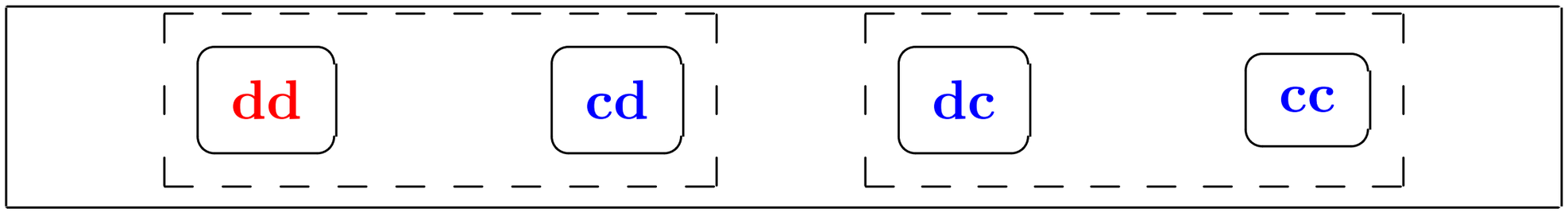
but in a different world (d, c) we have

$$(d, c) \models B_1(c_1 \wedge c_2) \wedge B_2(c_1 \wedge c_2)$$

Is c_1 a *common belief* in (d, d)? Is it a common *true* belief?

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.

Validities on Plausibility Models

The following principles are valid on the class of all plausibility models:

Persistence of Knowledge: Knowledge Implies (Conditional) Belief (Under Any Condition).

$$K_a\psi \Rightarrow B_a^\theta\psi$$

Strong Positive Introspection of (Conditional) Belief: Agents KNOW Their (Conditional) Beliefs.

$$B_a^\varphi\psi \Rightarrow K_a B_a^\varphi\psi$$

Strong Negative Introspection of (Conditional) Belief: Agents KNOW What They DON'T (Conditionally) Believe.

$$\neg B_a^\varphi\psi \Rightarrow K_a\neg B_a^\varphi\psi$$

Of course, these hold in particular for *simple (unconditional) belief*.

More Validities

Normality (Kripke's Axiom)

$$B_a^\theta(\varphi \Rightarrow \psi) \Rightarrow (B_a^\theta \varphi \Rightarrow B_a^\theta \psi),$$

$$K_a(\varphi \Rightarrow \psi) \Rightarrow (K_a \varphi \Rightarrow K_a \psi).$$

All Standard Properties of Knowledge:

Truthfulness, Positive and Negative Introspection of K.

Success: Given φ, one believes φ.

$$B_a^\varphi \varphi$$

Knowledge implies True Belief

From our definitions it is easy to see that we always have

$$K_a\varphi \Rightarrow (\varphi \wedge B_a\varphi).$$

If something is known then it is believed, and moreover it is true.

The converse however is false in our setting!

True Belief is not Knowledge

Albert believes he's a genius

$$(d, g) \models B_a g,$$

but **he doesn't "know" he's a genius**, since there exist possible worlds e.g. $(d, \neg g)$, in which he wouldn't be a genius.

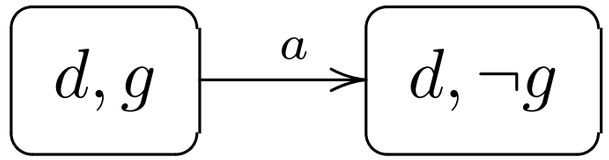
In this case it is even more clear that Albert's true belief is not knowledge: **unlike knowledge, it can be easily lost**! *If he learned that he was drunk, he'd no longer believe he is a genius.*

So **not every belief that happens to be true qualifies as "knowledge".**

Unsafe Belief

Not only that: Albert loses his (true) belief in genius AFTER he learns some TRUE information (namely that he is drunk)!

After $!d$, we get



and NOW Albert believes he is NOT a genius!

We say that his belief in g (although true) is **un-safe**.

Safe Belief: intuition

Intuitively, *a belief is "safe" if you cannot lose it by learning any new TRUE fact.*

Beliefs that are un-safe (even if true) can be easily lost, , just by learning some other (true) information.

A person holding an un-safe belief can be easily *manipulated* by somebody else, **without any lies or deceit**: just feed him some information!

Safe Belief: formal definition

Recall that belief in plausibility models was NOT the Kripke modality for the plausibility relation $\leq$.

Let us now **define "safe belief" $\Box_a$ as the Kripke modality for the plausibility relation**:

$$s \models \Box_a \varphi \qquad \text{iff:} \qquad t \models \varphi \text{ for all } t \text{ such that } s \leq_a t.$$

From now on, in the rest of the course, $\Box_a$ **will denote safe belief** (which belief will only be denoted by B_a, and knowledge by K_a).

Why Call it Safe ?

But what does this formal definition have to do with the intuitive notion of safety?

To motivate this, note the following (easily verifiable) equivalence:

$s \models \Box_a \varphi$ iff: $s \models B_a^P \varphi$ for all P such that $s \models P$.

A belief in φ is "safe" (i.e. $\Box_a \varphi$ holds) iff it CANNOT be lost whenever any additional TRUE evidence is given: so φ is believed, and continues to be believed, conditional on EVERY TRUE information.

Safe Beliefs are True Beliefs

Since plausibility relations are *reflexive*, we have the following validity in all plausibility models:

$$(Truthfulness) \qquad \Box_a \varphi \Rightarrow \varphi$$

Safe beliefs are true beliefs.

But the converse is FALSE: *not all true beliefs are safe*.

COUNTEREXAMPLE: Albert's belief in genius g is true, but un-safe.

$\Box$ is positively, but NOT negatively introspective

Since the plausibility relation $\leq_a$ is transitive, but NOT necessarily symmetric (nor Euclidean), $\Box$ **is positively, but NOT necessarily negatively introspective**:

$$\Box_a \varphi \Rightarrow \Box_a \Box_a \varphi$$

is valid in all plausibility models, but

$$(??)\ \neg \Box_a \varphi \Rightarrow \Box_a \neg \Box_a \varphi$$

is NOT valid.

In fact, **agents always believe that their own beliefs are safe (even if they are not)**:

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

is valid.

Relationships: knowledge, safe belief and (conditional) belief

Knowledge implies safe belief:

$$K_a\varphi \Rightarrow \Box_a\varphi$$

is valid on plausibility models.

Safe Belief implies belief:

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

is valid on plausibility models.

Moreover, **safe belief implies conditional belief given any TRUE condition**:

$$(\Box_a\varphi \wedge \theta) \;\Rightarrow\; B_a^\theta\varphi$$

Belief, in Terms of Safe Belief

An important observation, is that, in a plausibility model, *belief can in fact be "defined" in terms of safe belief*, because of the following validity:

$$B_a\varphi \;\; \Leftrightarrow \;\; \Diamond_a \Box_a \varphi \,,$$

where $\Diamond_a\varphi = \neg\Box_a\neg\varphi$ is the dual Diamond modality for $\Box_a$..

EXERCISE: Prove this semantically.

Conditional Belief, in terms of K and $\Box$

Conditional Belief can also be thus "defined", but ONLY if we use BOTH knowledge K and safe belief $\Box$.

$$B_a^{\varphi}\psi \quad \Leftrightarrow \quad \Big(\tilde{K}_a\varphi \Rightarrow \tilde{K}_a(\varphi \wedge \Box_a(\varphi \Rightarrow \psi)) \Big),$$

where $\tilde{K}_a\varphi := \neg K_a \neg\varphi$ is the Diamond modality for K_a.
(You can read $\tilde{K}_a\varphi$ as "φ is consistent with agent a's knowledge", or "agent a consider φ possible".)

A new notion: Strong Belief

We say a sentence φ is **strongly believed** by agent a in world w (of model $\mathbf{M}$), and write $w \models Sb_a\varphi$, if the following two conditions hold

1. φ **is consistent with the agent's knowledge** at $w(a)$:

$$\|\varphi\|_{\mathbf{M}} \cap w(a) \neq \emptyset,$$

2. **within the agent's information cell $w(a)$, all φ-worlds are (strictly) more plausible than all non-φ-worlds**:

$$s >_a t \text{ holds} \quad \text{for every } s \in w(a) \cap \|\varphi\|_{\mathbf{M}} \text{ and every } t \in w(a) - \|\varphi\|_{\mathbf{M}}.$$

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

$$Sb_a\varphi \Rightarrow B_a\varphi.$$

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**!

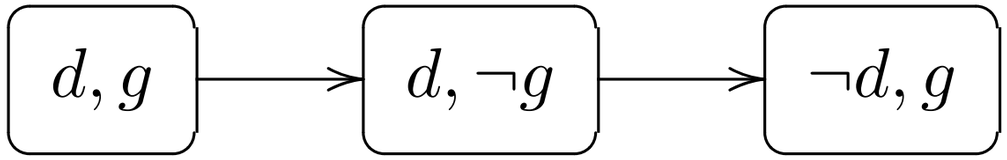
More precisely:
φ is **strongly believed** iff φ **is believed and is also conditionally believed given any new evidence (truthful or not) EXCEPT if the new information is known to contradict** φ; i.e. if:

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

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).

Non-closure of Strong Belief

Unlike knowledge and belief, **the set of strong beliefs is not closed under logical inference**.

In particular, in the Winestein example, **Albert strongly believes $g \wedge \neg d$, but he does not strongly believe g.**

This shows more, namely that

$$Sb_a(\varphi \wedge \psi) \not\Rightarrow Sb_a\varphi \wedge Sb_a\psi.$$

COUNTER-EXAMPLE: see previous slide

But the converse implication DOES hold.

Strong Belief, in terms of K and $\Box$

NOTE: Strong belief can also be "defined" in terms of knowledge K and safe belief $\Box$ (using the "definition" of belief B in terms of these two notions):

$$Sb_a\varphi \;\Leftrightarrow\; (\, B_a\varphi \;\wedge\; K_a(\varphi \Rightarrow \Box_a\varphi)\,)$$

is valid on plausibility models.

Lecture 5.1. Learning in Single-Agent Plausibility Models

chapter "**Epistemic Logic and Information Update**", in *Handbook of Philosophy of Information* *Stanford Encyclopedia of Philosophy* (available online at *https://plato.stanford.edu/*), search for entry on "**Dynamic Epistemic Logic**"

"Dynamic" Belief Revision

Suppose we have a **single-agent plausibility model**, represent *some (un-named) agent's current beliefs, conditional beliefs, strong beliefs* etc.

Suppose now that some **new information** φ comes from some given **source** (which can be another agent, or a book, a database, a library etc).

How can we compute the agent's new beliefs, conditional beliefs etxc (after learning the new information φ) **about the new state of the world (as it is AFTER the learning)**?

But isn't this already solved?! We have one way to do it:
updates $!\varphi$, by which the *all non-*φ *are deleted*.

Hard versus Soft Information

But an update is a very *special* kind of learning, that comes with "*warranty of truthfulness*": it assumes as that the new information is **absolutely certain**.

We call this "**hard information**". **Once learned, it'll never have to be revised:** *once we deleted some worlds, we cannot "un-delete" them* (add them back)!

But **new information may be "soft"**: not so fully certain, just *very plausible (believable)*. Or sometimes not even that!

Our agent should still **learn** this information: if it is believable, then she should **come to believe it**;
but in a way that is **reversible**: if later *she might learn that this information was false after all*, she **needs to have a way to "un-learn" it.**

Generalization: Doxastic Upgrades

So we want to *generalize* our dynamics to represent **"softer" forms of learning**.

Like updates, *each type of learning will correspond to a type of model transformation*: a **doxastic "upgrade"** (or "**belief upgrade**").

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

Upgrades as Model Transformers

A **belief upgrade** **is a *model transformer*** T, i.e. a partial map from single-agent plausibility models $\mathbf{S} = (S \leq, \|\cdot\|)$ to *new* (single-agent) plausibility models $T(\mathbf{S}) = (S', \leq', \|\cdot\| \cap S')$, 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 preorder $\leq'$ on S'.

The upgrade T is **defined** ONLY *if the real world survives it*, i.e. IFF $s_* \in S'$. *Otherwise, it cannot be performed.*

Hard and Soft Upgrades

NOTATION: The **domain** of T in a model $\mathbf{S}$ is defined as

$$Dom_{\mathbf{S}}(T) := S' \,,$$

i.e. it consists of *all states in which* T *can be performed*.

An upgrade T is called **soft** iff it's a **total** map; equivalently, iff

$$Dom_{\mathbf{S}}(T) = S \,, \quad \text{for all plausibility models} \mathbf{S}.$$

A soft upgrade **doesn't add anything to the agent's "hard" 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

As for updates, we can add to the language dynamic operators $[T]\psi$, for each upgrade T, to express the fact that ψ **will surely be true** (in the new model) **AFTER the upgrade** T.

The semantics is given as before by *performing the transformation* T and checking whether the post-condition ψ *holds in the new, upgraded model* whenever the upgrade can be performed:

$$s \models_{\mathbf{S}} [T]\psi \quad \text{iff:} \quad s \in Dom_{\mathbf{S}}(T) \text{ implies } s \models_{T(\mathbf{S})} \psi\,.$$

Again, for the dual existential modality $< T > \psi$, this can be easily given in terms of interpretation map:

$$\| < T > \psi \|_{\mathbf{S}} \;=\; \|\psi\|_{T(\mathbf{S})}\,.$$

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 Revision" with φ):
all φ-worlds (in any given information cell) become "better" (more plausible) than all $\neg\varphi$-worlds (in THE SAME information cell), and *within the two zones, the old ordering remains.*

(3) Conservative upgrade $\uparrow\varphi$ ("Minimal Revision" with φ):
the "best" (=most plausible) φ-worlds (in a given information cell) become better (strictly more plausible) than all other worlds (in THE SAME information cell), and *in rest the old order remains.*

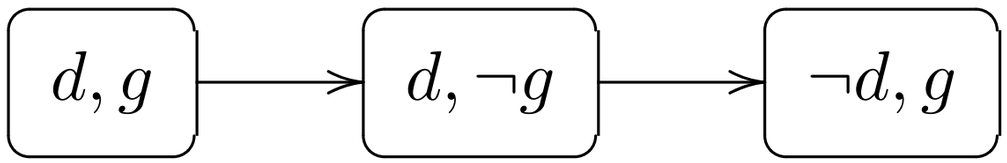
Different attitudes towards the new information

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

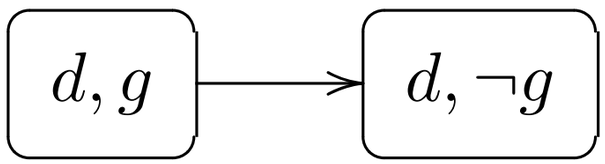
- **Update**: an **infallible** source. The source is *"known" (guaranteed) to be truthful.*
- **Radical (or Lexicographic) upgrade**: the source is **strongly believed to be truthful**. It'll be very hard to give up this belief! So, although the source may be *fallible*, it is **highly trusted**.
- **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!

Recall Updates: Learning that you're drunk

Remember that, in the Weinstein story, the event of Albert *seeing the result of a blood test, that proved that he is drunk beyond any doubt* was modeled as an **update** $!d$, taking the original model



into the model



NOTE: After this, Albert, not only *believes*, but in fact **knows** that he is drunk!

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*!"

NOTE: Mary is not (yet) an agent!

We don't **really** want yet to have Mary in our formal account: we are still concerned here with **single-agent belief revision**.

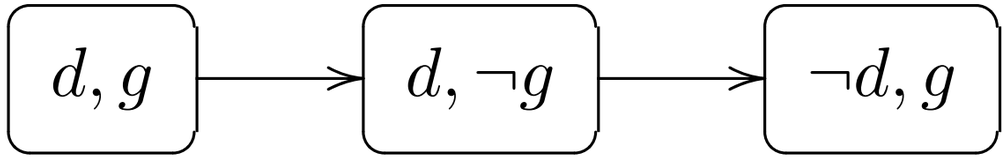
So for the moment, we will **NOT treat Mary as an "agent"** that can possess information, hold beliefs and revise them, but only as **the "source" of information**.

In our formal model there is still *only one "believer"*: Albert Winestein.

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

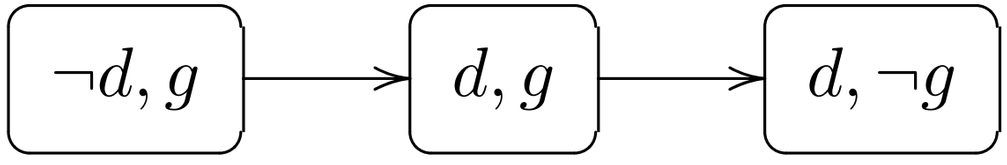


to capture Albert's new beliefs?

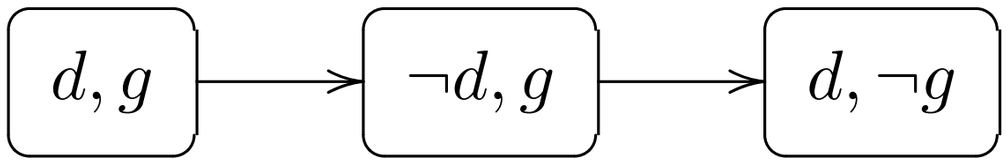
There are two drunk-worlds (d, g) and $(d, \neg g)$. **Which one should we promote ahead of all the others?**

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!

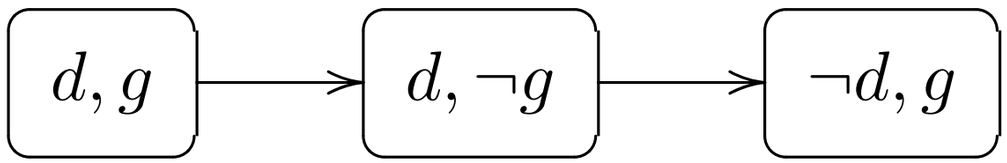
Radical Upgrade

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.*

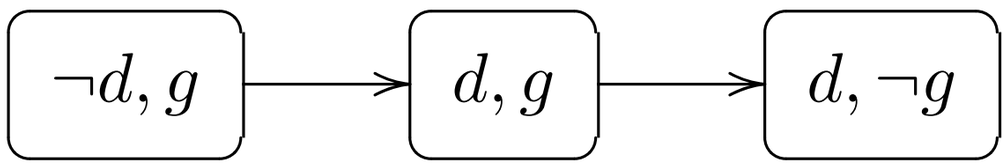
This corresponds to **radical upgrade**: it involves a rather radical revision of the prior beliefs, based on a strong belief in the correctness of the new information.

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

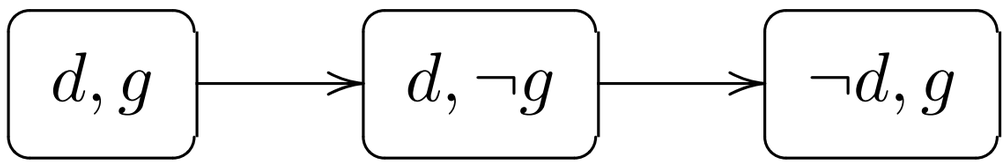
In this case, Albert will have to choose the second option: **promote only the most plausible d-world, leaving the rest the same.**

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).

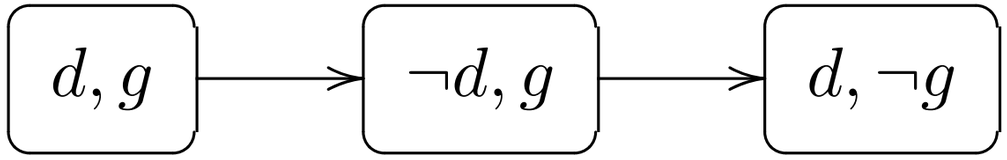
This corresponds to **conservative upgrade**.

Example of Conservative Upgrade

In the original Winestein situation



a conservative upgrade $\uparrow d$ produces the model



In this new model we have: $\quad Bd \wedge B^g \neg d.$

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!

Updates give you knowledge

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

For atomic facts (which don't change their truth value after being learnt), "was" can be replaced by "is":

$$[!p]Kp$$

is valid.

"**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.

(Radical) Upgrades Induce (Strong) Belief

In contrast, *soft upgrades can only induce belief*. More precisely:

After a *conservative upgrade*, the agent comes to **believe** that φ WAS the case (before the upgrade), **UNLESS he already knew** $\neg\varphi$ (before the upgrade).

Again, for atomic sentences, we can replace "was" by "is", hence the validity:

$$\neg K\neg p \Rightarrow [\uparrow p]Bp.$$

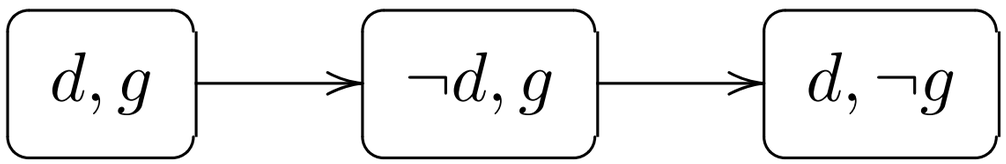
After a *radical upgrade*, the agent comes to **strongly believe** that φ WAS the case, **UNLESS he already knew** $\neg\varphi$ (before the upgrade):

$$\neg K\neg p \Rightarrow [\Uparrow p]Sbp.$$

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.

Updates are closed under composition

Recall that *updates on epistemic/doxastic models were closed under sequential composition.*

The same happens with updates on *plausibility models*:

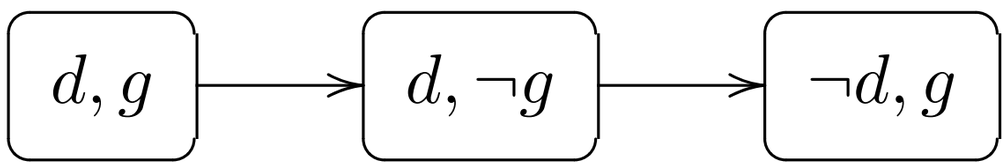
A **sequence of successive updates is equivalent to only one update on any plausibility model**:
the effect of doing first an update $!\varphi$ *then an update* $!\psi$ *is the same as the effect of doing the update* $!(\varphi \wedge [!\varphi]\psi)$

In other words: instead of **first announcing you that** φ **is the case then announcing you that** ψ **is the case**, I just announce you from the start that φ **is the case AND that** ψ **WOULD be the case AFTER I'd announce you** φ.

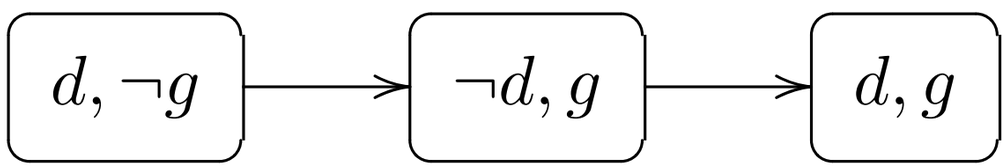
Radical upgrades are NOT closed under composition

But *this is NOT the case for soft announcements*: e.g. **a sequence of two radical upgrades is NOT a radical upgrade!**

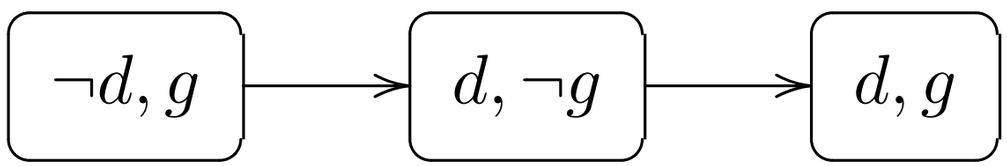
Counterexample: Prof Winestein



Tell Albert **he's a drunk genius**: do $\Uparrow (d \wedge g)$.



Then announce **he's drunk**, by an upgrade $\Uparrow d$:



NO single (radical or conservative) upgrade can get us from the first to the last model!

Reduction laws

The **reduction law for belief under updates** makes essential use of *conditional belief*:

$$[!\psi]B\varphi \quad \Longleftrightarrow \quad \big(\psi \Rightarrow B^{\psi}[!\psi]\varphi\big)\,.$$

Of course, for a complete axiomatization, we need to get in turn a *reduction law for conditional belief*!

$$[!\psi]B^{\theta}\varphi \quad \Longleftrightarrow \quad \Big(\psi \Rightarrow B^{\psi\wedge[!\psi]\theta}[!\psi]\varphi\Big)\,.$$

Similar, but more complicated, axiomatizations can be given for the other types of upgrades.

Other Model Transformers: "Negative" Attitudes

All upgrades considered till now corresponded to **positive** attitudes towards the source of information:the new information was at least believed after the upgrade (unless it was known to be false).

In contrast, there also exist **negative** ones:

Negative Update $!^{-}\varphi$: the source is known to be **infallibly wrong** (-it always lies);

Negative Radical upgrade $\Uparrow^{-} \varphi$: the source is **highly distrusted**. The listener *strongly believes the speaker is lying.*

Negative Conservative upgrade $\uparrow^{-} \varphi$: the source is plainly **distrusted**. The listener (barely) *believes the speaker is lying.*

Negative Update $!^{-}\varphi$ is equivalent to an update $!(\neg\varphi)$ with the negation of the received information: **all the** φ **states are deleted** and *the same relations are kept between the remaining states.*

Negative Radical upgrade $\Uparrow^{-}\varphi$ is equivalent to a radical upgrade $\Uparrow(\neg\varphi)$ with the negation of the received information: **all** $\neg\varphi$**-worlds become "better" (more plausible) than all** φ**-worlds**, and *within the two zones, the old ordering remains.*

Negative Conservative upgrade $\uparrow^{-}\varphi$ is equivalent to a conservative upgrade $\uparrow(\neg\varphi)$ with the negation of the received information: **the "best"** $\neg\varphi$**-worlds become better than all other worlds**, and *in rest the old order remains.*

These negative attitudes are also subsumed under the generic name of upgrades (though they might better be called "downgrades"!).

Another Example: Neutrality

(7) Doxastic Neutrality $id\varphi$ is the attitude according to which the source cannot be trusted nor distrusted: the listener *simply ignores* the new information φ, keeping her old plausibility order as before. This is the **identity map** id on plausibility models.

Other Examples: Mixed Attitudes

An agent's attitude towards a source of information might **depend on the type of information** received from that source: she might treat differently different types of information. She might **mix** two or more basic transformers, using **semantic or syntactic conditions** to decide which to apply.

For instance, the agent may consider the source to be infallibly right about sentences belonging to a given sublanguage L_0, while she may only barely trust it with respect to any other announcements. This attitude could be denoted by $!_{L_0}\uparrow$.

Example

If the "source" is a well-known Professor of Mathematics, our agent may accept him as *an infallible source of mathematical statements*, and thus perform an update $!\varphi$ whenever the professor announces a sentence φ about Mathematics.

In any other case (say when the Professor talks about food, or love), our agent might treat the new information coming from the Professor more cautiously (say, barely believing it $\uparrow \varphi$, or even ignoring it, and thus applying id):
indeed, a typical mathematician may be utterly unreliable concerning any other area of conversation except for Mathematics!

Lecture 5.2. Learning in Multi-Agent Plausibility Models

chapter "**Epistemic Logic and Information Update**", in *Handbook of Philosophy of Information* *Stanford Encyclopedia of Philosophy* (available online at *https://plato.stanford.edu/*), search for entry on "**Dynamic Epistemic Logic**"

Joint Upgrades and Updates

THE SIMPLEST way to generalize upgrades to multi-agent plausibility models is to apply the SAME update or upgrade operations *simultaneously to all the relations*.

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

some information φ is PUBLICLY, 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, or a conservative upgrade etc).

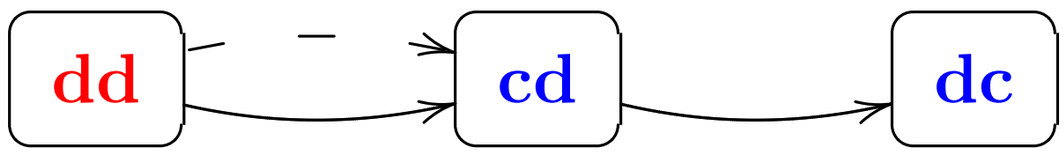
Muddy children example: A Joint Update

Start with the "Two Muddy Children Standing in a Line" example:

Now, Father, who is an **infallible** source (-he never lies!). announces:

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

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

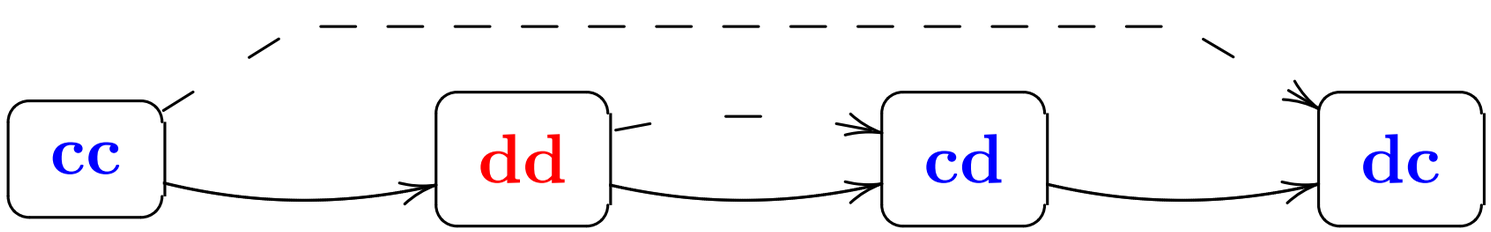


Muddy children example : Joint Radical Upgrade

Alternatively, an older sister announces: "*At least one of you is dirty*". She is a **highly trusted source, though not infallible**: so we have a *radical upgrade* $\Uparrow (d_1 \vee d_2)$.

Applying this upgrade to the original model

we obtain the upgraded model:



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* $\uparrow (d_1 \vee d_2)$, which changes the initial model

into

"Publicly Announced" Private Upgrades

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

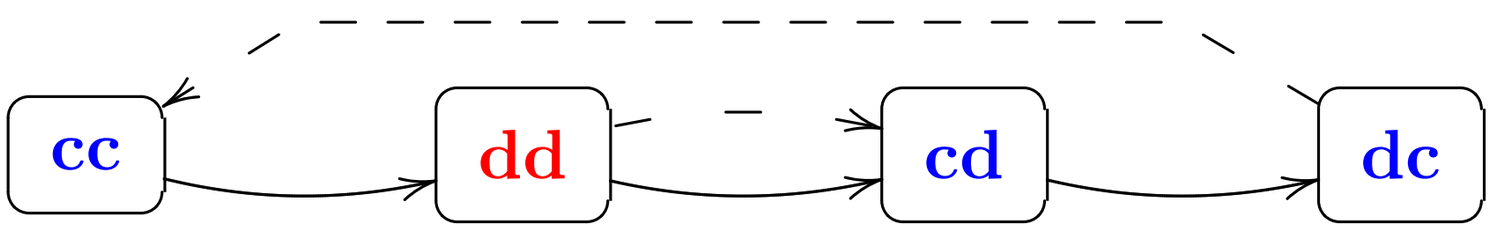
it is common knowledge that a single agent a upgrades/updates with φ, but also that the others do NOT upgrade/update at all with φ.

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.*

Example of "Publicly Announced" Private Upgrade

In the same initial situation (Two Muddy Children in a Line)

the older sister makes the same announcement, but now we assume it is *common knowledge* that **only child 2 highly trusts the sister**, and 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:



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**:

for instance, *one agent may highly trust the source*, and thus apply *radical upgrade*, while *another agent may only barely trust it*, and so apply *conservative upgrade*.

Beyond Common Knowledge of Attitudes

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.

Multi-Agent Belief Dynamics: the general case

To go beyond that, we'll need **a more sophisticated setting**, *borrowed from (and extending) "classical DEL"*:

event-plausibility models.

Models for 'Events'

Until now, our models captured 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 plausibility models to represent such types of *epistemic events*, in a way that is similar to the representations we have for epistemic states.

Event Plausibility Models

A **multi-agent event plausibility model**

$$\mathbf{\Sigma} \quad = \quad (E, \sim_1, \ldots, \sim_1, \leq_1, \ldots, \leq_n, pre)$$

is just like a multi-agent state plausibility model, except that its elements $e \in E$ are now called **events** (or *actions*), and instead of the valuation we have a **precondition map** pre, associating a sentence pre_e to each action e: *pre_e is true in a world iff e can be performed.*

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

Simplified description of Event Plausibility Models

As before (for state plausibility models), the epistemic indistinguishability relations $\sim_a$ are definable in terms of $\leq_a$:

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

So we can **skip** $\sim_a$ when describing the event plausibility models, obtaining a simplified presentation:

$$\mathbf{\Sigma} \quad = \quad (E, \leq_1, \ldots, \leq_n, pre).$$

So, most of the time, we will **draw only the plausibility arrows**, with the same conventions as for state plausibility models (we skip the loops and the arrows that can be obtained by transitivity).

Equi-Plausibility Relation

We also introduce a special notation

$$\cong_a := \leq_a \cap \geq_a$$

for the *equi-plausibility relation* between events.

Two actions e and f are **equally plausible** (or "*equi-plausible*") if each of them is at least as plausible as the other one:

$$e \cong_a f \quad \text{iff:} \quad e \leq_a f \text{ and } f \leq_a e$$

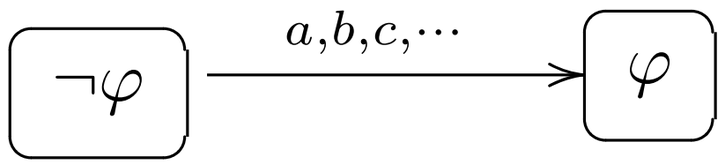
EXAMPLE: event model for joint update

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 **there are loops for every agent**: since the announcement is *public*, and it is *common knowledge that everybody does this update*, since everybody considers the source to be infallible.)

EXAMPLE: event model for joint radical upgrade

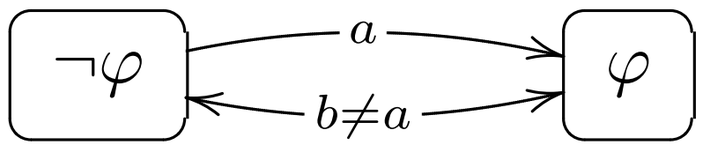
The event model for a **joint radical upgrade** $\Uparrow \varphi$ is:



This is still a **public announcement**: everybody hears φ. The action **on the right** represents the case that the announcement is **truthful** (i.e. φ is in fact true), while the action **on the left** represents the case that the announcement is **false**). Every agent considers the action on the right to be more plausible than the one on the left (-so they **all strongly believe the source is telling the truth**), and moreover **this fact is common knowledge**.

EXAMPLE: event model for publicly-announced private upgrade

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



Though *the attitudes towards the source* ***differ*** *now* (only a **strongly trusts** the source the source, the others are **neutral**, neither trusting nor distrusting the source), ***these attitudes are common knowledge***: they all know (and know the others know etc) that a strongly trusts the source, and that the others dismiss it.

Example: lack of common knowledge of attitudes

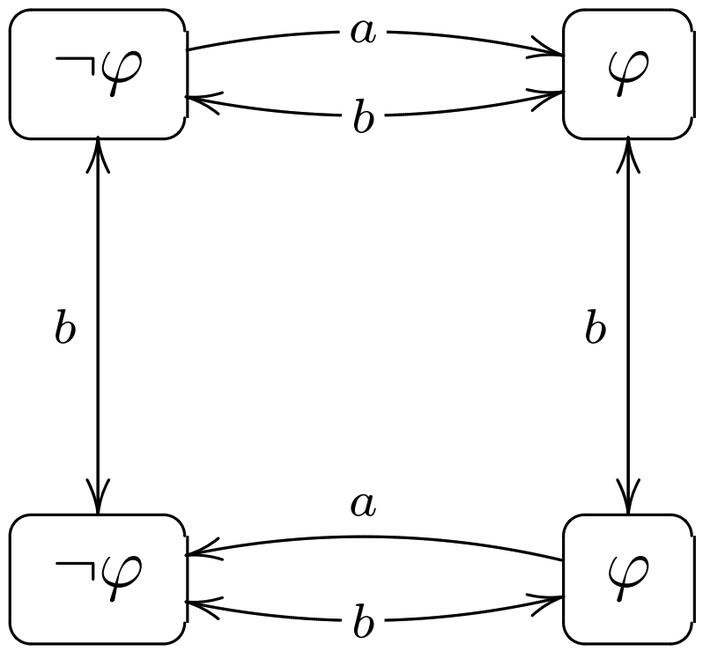
Suppose that everything goes on as in the above example (a sentence φ is ***publicly announced***, and it is still *common knowledge* that agent b is **neutral** towards the source (-neither trusts or distrusts the source, so he applies the identity upgrade id, and hence doesn't change his plausibility).

But this time **agent a's attitude is NOT common knowledge.**

Instead (to keep it simple) it is **common knowledge that EITHER a highly trusts the source ($\Uparrow$) OR she highly distrusts the source($\Uparrow^{-}$).**

(And, of course, we assume that **each agent knows his/her own attitude!**)

The event plausibility model is now the following:



where for simplicity I skipped some of the arrows that can be obtained by transitivity:

e.g. there are bidirectional b-arrows between the $\neg\varphi$-event in the upper-left corner and the φ-event in the lower-right corner, and similarly there are bidirectional b-arrows between the φ-event in the upper-right corner and the $\neg\varphi$-event in the lower-left corner.

Exercise

Draw an event plausibility model for a scenario similar to the previous one, **except that neither of the two agents' attitudes are common knowledge**.

Instead, (to keep it simple) it is **common knowledge that each of the two agents either highly trusts the source ($\Uparrow$) or is neutral (id)**.

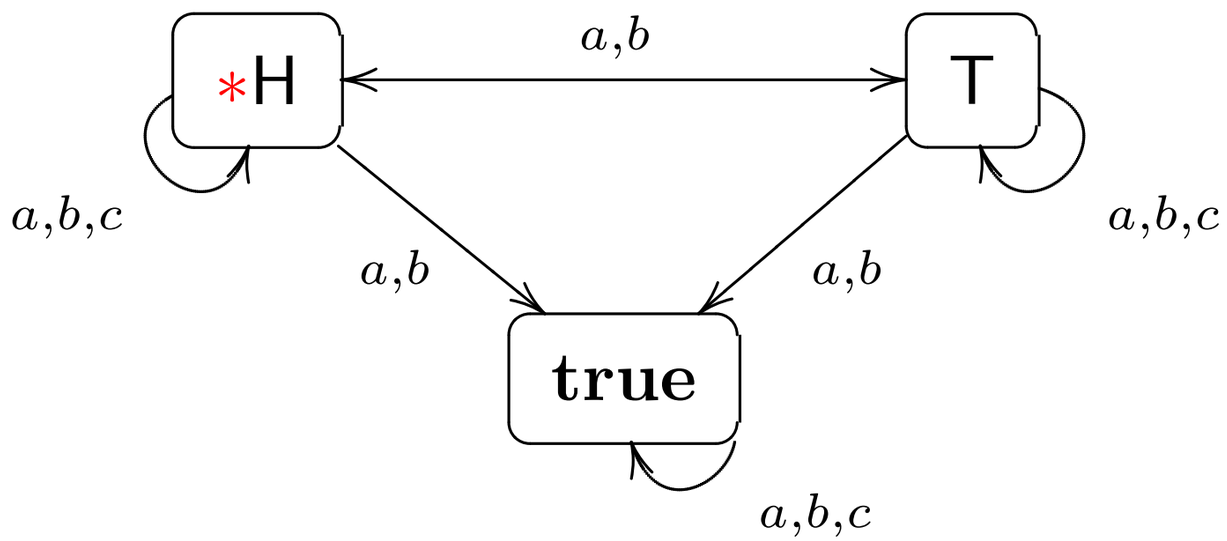
HINT: You will need to draw (a plane projection of) a cube!

Example: Secret (Fully Private) Learning

Let us consider again the "Concealed Coin" situation from the first Lecture: a coin was on the table, covered, so it was common knowledge that nobody saw its upper face.

But now, when nobody looks, Charles (agent c) **secretely takes a peek at the coin and sees it's Heads up**. Alice (a) and Bob (b) don't suspect anything: they *believe that nothing is really happening*.

A possible representation of this action as an event plausibility model is



The real event (marked with a star) is the one in which Charles secretely takes a peek and sees the coin Heads up. What Alice and Bob believe it's happening is "nothing" (their most plausible event has precondition "true" and it's completely public): i.e. that nobody learns anything new.

However, if later Alice and Bob would be told that this was NOT the case, and that in fact Charles took a peek, they would still not know whether he saw the coin Heads up or Tails up. So they'd equally plausible that Charles saw Heads (=the event in the upper left corner, marked with a star) and he saw Tails (the event in the upper right corner).

Looking for a General Update Rule

We want to *compose* any initial state plausibility model $\mathbf{M} = (W, \leq_1, \ldots, \leq_n, \|.\|)$ with any event plausibility model $\mathbf{\Sigma} = (E, \leq_1, \ldots, \leq_n, pre)$, in order to compute the *new state plausibility model* after the event:

i.e. find an operation $\otimes$, such that $\mathbf{M} \otimes \mathbf{\Sigma}$ correctly represent all the agents' information (beliefs, knowledge, conditional beliefs, strong beliefs etc) after the event.

A Product Construction

We think of the basic events of our event model as deterministic actions: for a given input-state w and a given event e, there is **at most one output-state**. This means that we can represent the **new (output) states** as **pairs** (w, e), where $w \models pre_e$.

Hence, the **new state space** after the event will be represented as **a subset of the Cartesian Product** $W \times E$.

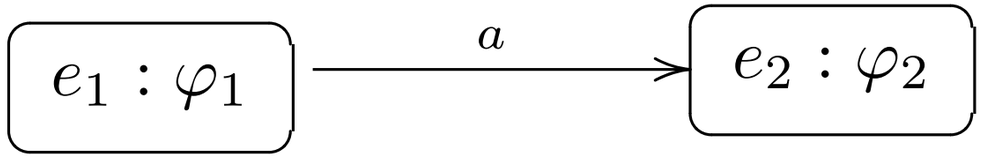
Our events are *purely epistemic/doxastic* (i.e. *learning or communication* actions), so they do **not change the valuation**:

$$(w, e) \models p \text{ iff } w \models p.$$

But how should we define the new plausibility $\leq_a$ **on output-pairs** (w, e) ?

First Case

Suppose the event model includes a **strict** plausibility order between two events e_1, e_2 with preconditions φ_1, φ_2:



IF before the event **worlds s_1 and s_2 were a-indistinguishable** ($s_1 \sim_a s_2$), and satisfying the corresponding preconditions ($s_1 \models \varphi_1$, $s_2 \models \varphi_2$), THEN (after the event) **world (s_2, e_2) should become strictly more plausible for a than** (s_1, e_1).

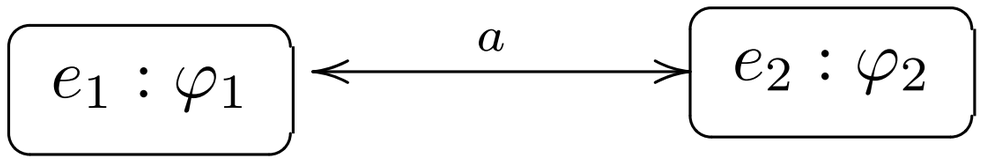
$$(s_1 \sim_a s_2 \text{ and } e_1 <_a e_2) \implies (s_1, e_1) <_a (s_2, e_2)$$

IF however, the **two worlds were distinguishable before the event**, than *one of the them was already known to be impossible*, so *it is still known to be impossible* (i.e. the **two pairs are still distinguishable**) *after the event*:

$$s_1 \not\sim_a s_2 \implies (s_1, e_1) \not\sim_a (s_2, e_2).$$

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: e_1 or e_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**.

$$(s_1 \leq_a s_2 \text{ and } e_1 \cong_a e_2) \implies (s_1, e_1) \leq_a (s_2, e_2).$$

Third Case

Finally, what if **the two events are epistemically distinguishable**: $e_1 \not\sim_a e_2$?

Then, *when one of them* (e_1) *happens, the agent* ***knows*** *it is the other one* (e_2) *is* ***impossible***.

By perfect recall, he *still knows after the event that* e_2 *did not happen.*
So **he can still distinguish the output pairs**:

$$e_1 \not\sim_a e_2 \implies (s_1, e_1) \not\sim_a (s_2, e_2).$$

The Action-Priority Rule

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

$(s_1, e_1) \leq_a (s_2, e_2)$ iff: either $(e_1 <_a e_2, s_1 \sim_a s_2)$ or $(e_1 \cong_a e_2, s_1 \leq_a s_2)$.

Interpretation

The Action-Priority Rule gives "priority" to the *action* plausibility relation.

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 way the current action appears to the agent changes the initial beliefs, and NOT vice-versa. **The past beliefs need to be revised by the current beliefs, and NOT the other way around!**

DEFINITION: "Action-Priority" Update

Given a state plausibility model $\mathbf{M} = (W, \leq_1, \ldots, \leq_n, \nu)$ and an event plausibility model $\mathbf{\Sigma} = (E, \leq_1, \ldots, \leq_n, pre)$, we define their *"action-priority" update*

$$\mathbf{M} \otimes \mathbf{\Sigma} \quad = \quad (W \otimes E, \leq'_1, \ldots, \leq'_n, \nu'), \quad \text{where}$$

1. The "new worlds" are represented as "consistent" world-action pairs: i.e. such that this action is executable in this world.

$$W \otimes E := \{(w, e) \in W \times E \ : w \models_{\mathbf{M}} pre_e \}.$$

2. For all $(w, e), (s, f) \in W \otimes E$ and all agents $a \in \mathcal{A} = \{1, \ldots, n\}$:

$$(w, e) \leq_a (s, f) \ \text{ iff: } \quad \text{either } (e <_a f, w \sim_a s) \text{ or } (e \cong_a f, w \leq_a s)$$

(where $\cong$ is the *equi-plausibility relation*).

3. The valuation stays the same: $\nu'(w, e) = \nu(w)$.

Special Case: Product Update

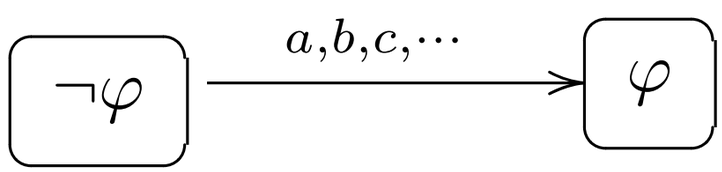
In particular, our rule has the consequence that the agents' knowledge relations (K) about the initial state and about the on-going event are composed according to the Product Update rule:

$$(w, e) \sim_a (s, f) \text{ iff } w \sim_a s \text{ and } e \sim_a f .$$

This special case is the "**Product Update**", that you encountered in **the first part of the course**.

EXAMPLE: joint radical upgrade again

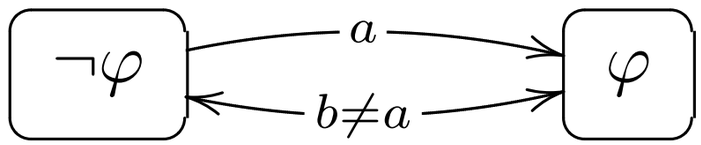
EXERCISE: Check that, if we take the Action Priority Update of any state model $\mathbf{S}$ with the event model $\mathbf{\Sigma}_{\Uparrow\varphi}$ for joint upgrade $\Uparrow\varphi$, given as above by



the resulting new state model $\mathbf{S} \otimes \mathbf{\Sigma}_{\Uparrow\varphi}$ is indeed (isomorphic to) the result of performing the joint radical upgrade $\Uparrow\varphi$ on $\mathbf{S}$.

EXAMPLE: publicly-announced private upgrade, again

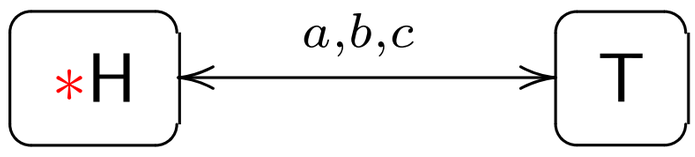
EXERCISE: Check that, if we take the Action Priority Update of any state model **S** with the event model for a *publicly-announced private (radical) upgrade by agent* a, given as above by



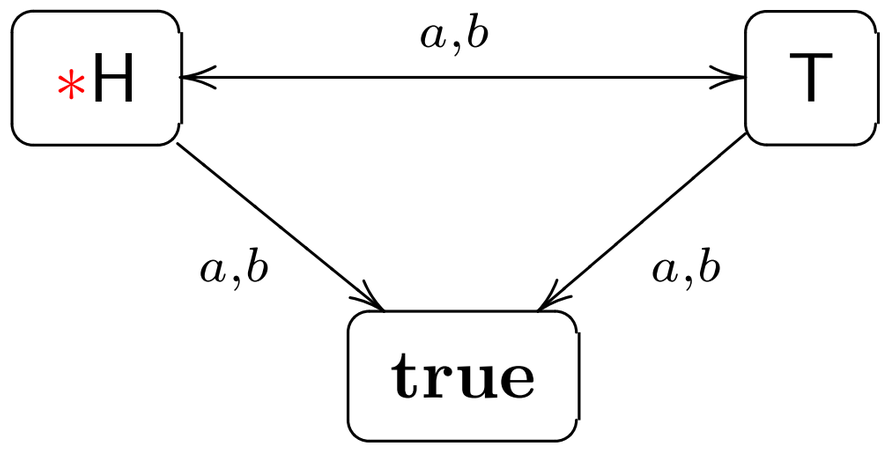
the resulting (updated) model is indeed the same as the result of performing the publicly-announced private radical upgrade with φ by agent a (i.e., applying $\Uparrow \varphi$ only to agent a's arrows).

Example: Secret Learning, again

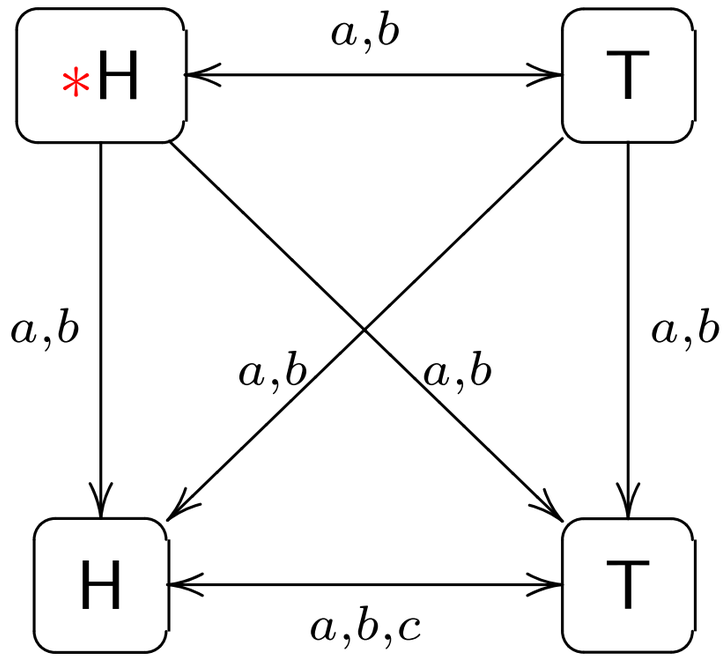
In the coin scenario in which *Charles is secretely taking a peek*, the Action-Priority update of the original state (plausibility) model with this event plausibility model



$$\otimes$$



gives us:

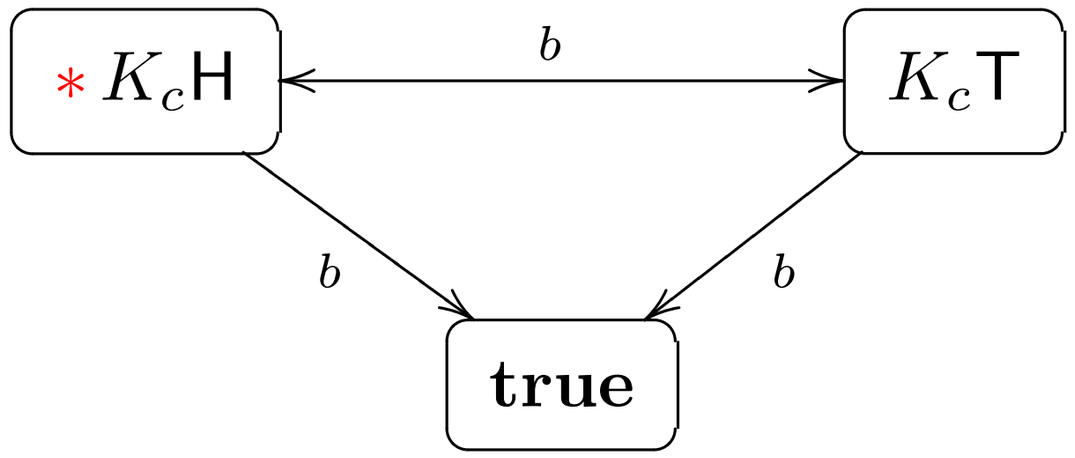


So e.g. Alice still believes that Charles doesn't know the face. However, if later she's given the information that he took a peek (without being told what he saw), she'd know that he knows the face; but as for herself, she'd still consider both faces equally plausible.

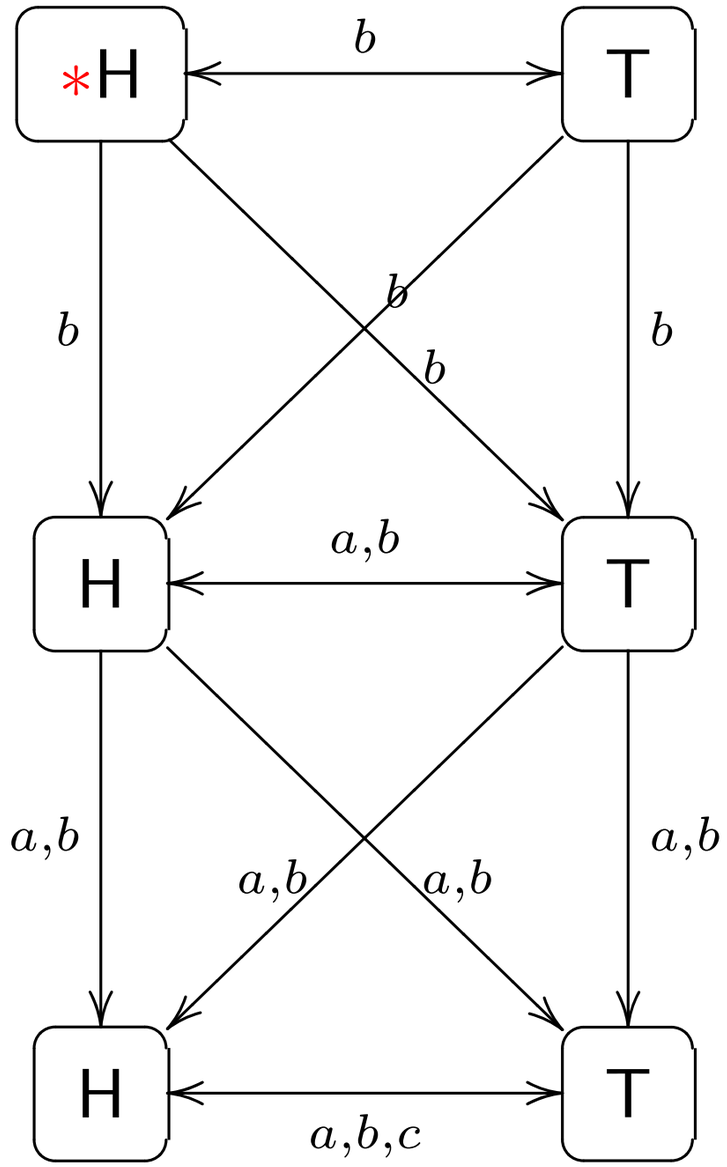
A Secret Communication Contradicting Prior Beliefs

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

This is a private knowledge announcement $!_a(K_cH)$:



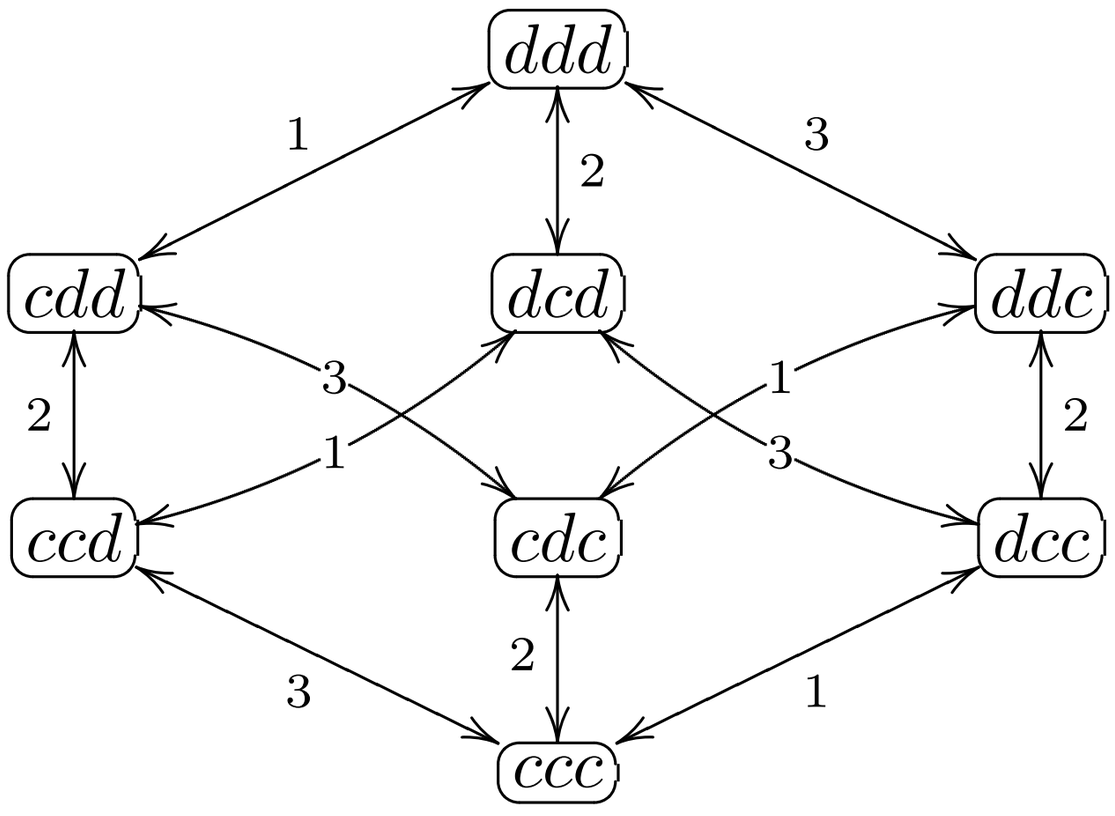
Alice previously didn't even suspect that Charles took a peek. So this private knowledge announcement is inconsistent with her prior beliefs. Nevertheless, after applying the Action Priority Product, we see that Alice, instead of going crazy, simply deduces that Charles somehow learned the face of the coin:



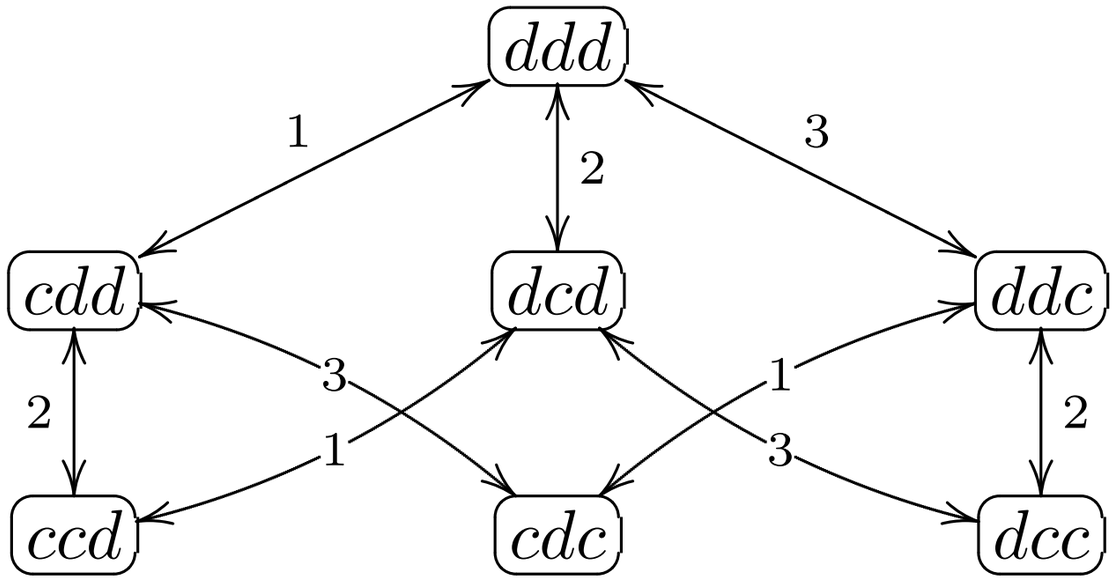
NOTE: I skipped some of the arrows that you can get by transitivity, e.g. there are b-arrows going from any of the two top worlds to any of the two bottom worlds.

Solving the standard "Muddy Children"

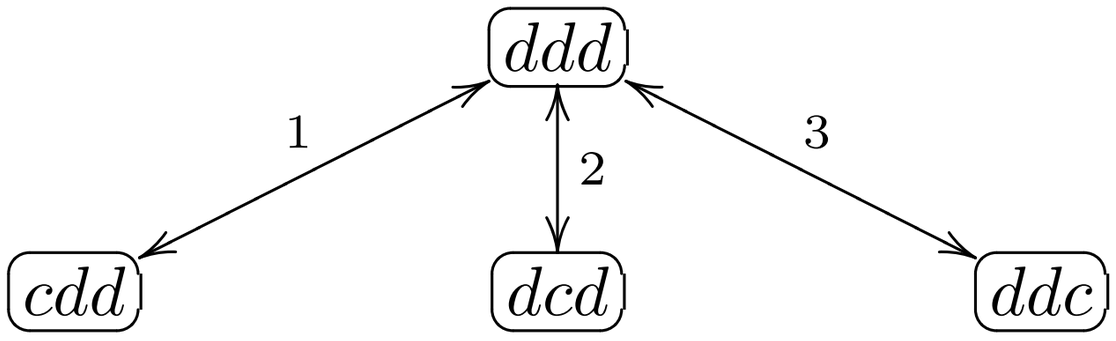
Three children, child 1 and child 2 are dirty. Originally, assume each child considers equally plausible that (s)he's dirty and that (s)he's clean:



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:



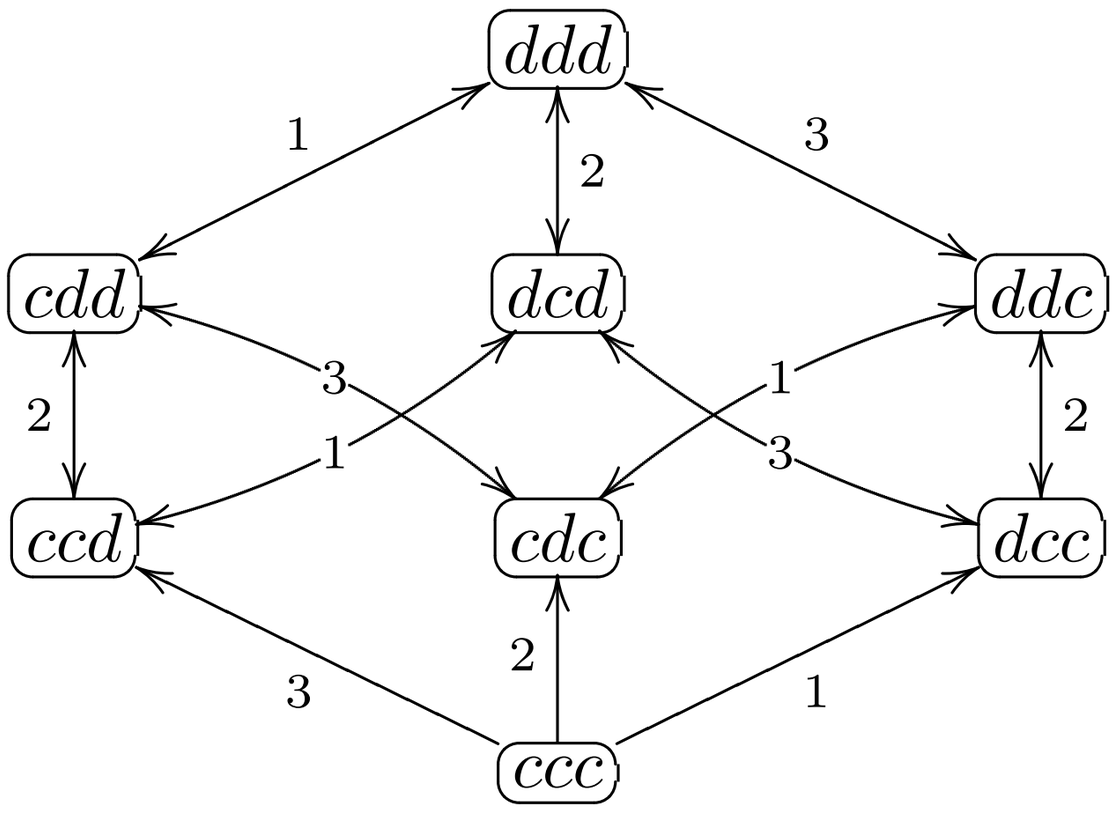
If the children answer "I don't know I am dirty", and they are infallible, then the update $!(\bigwedge_i \neg K_i d_i)$ produces:



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:

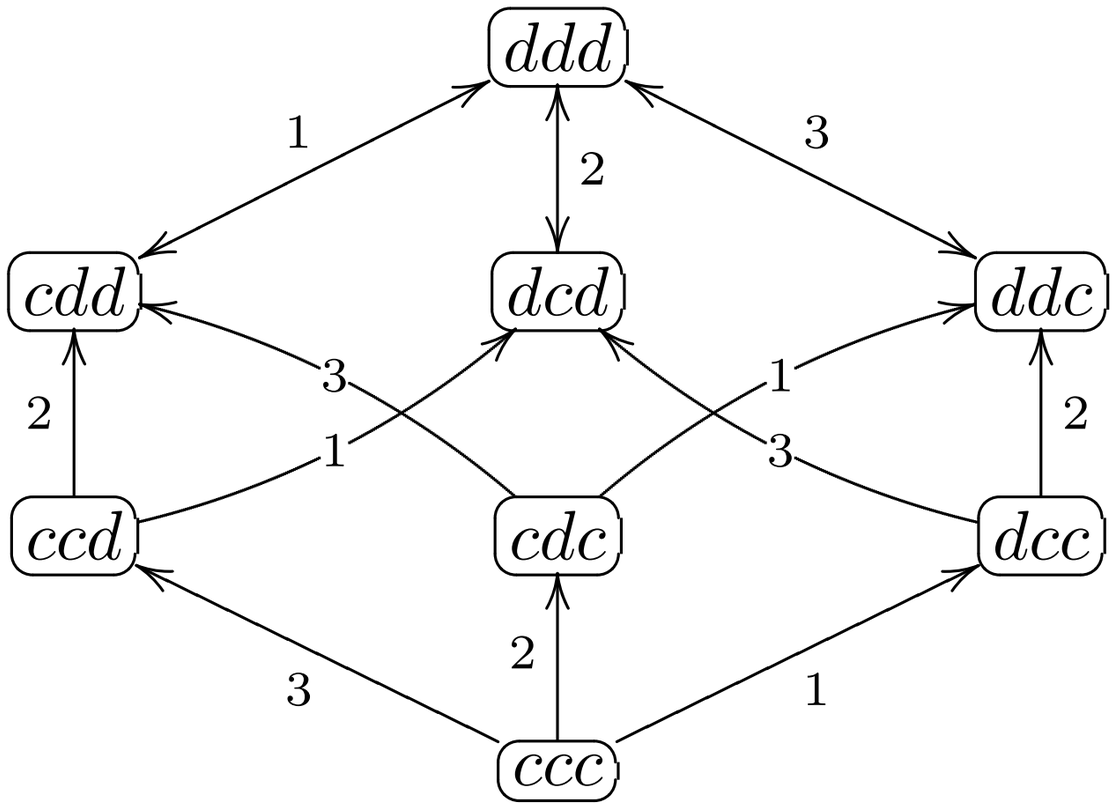


Do you believe you're dirty?

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*?

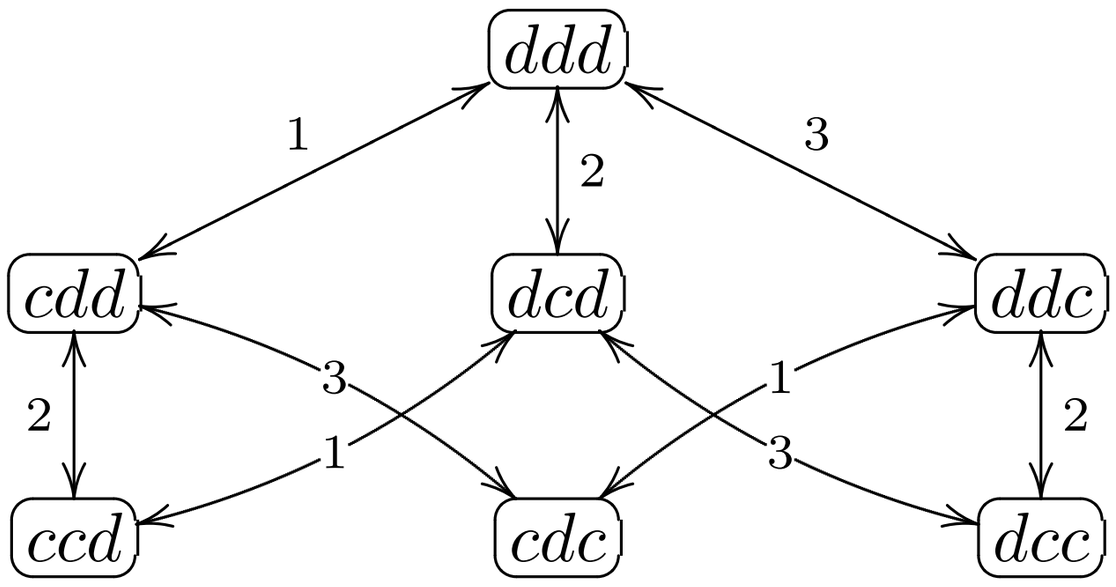
After a (radical or conservative) upgrade with the sentence $\bigwedge_i \neg B_i d_i$, we obtain:



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

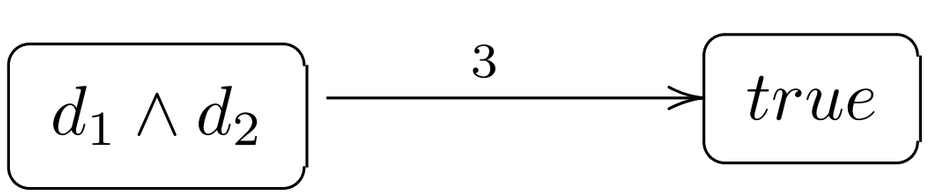
Cheating Muddy Children

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!