Dynamic Logics for Interactive Belief Revision

TUTORIAL:

Alexandru Baltag and Sonja Smets (ILLC, Amsterdam)

Plan of Lecture

1. Introduction
2. Plausibility Models and "Static" Doxastic Attitudes
3. Doxastic Transformers
4. Formalizing Dynamic Doxastic Attitudes
5. Fixed Points and Honesty
6. Persuasiveness
7. Reaching Doxastic Agreement
8. Belief Aggregation and Preference Merge
9. General DEL: Event Models and Action-Priority Product Update
10. Application to Game Theory: epistemic analysis of solution concepts

1. Introduction: Pirates of the Caribbean

Mullroy: *What's your purpose in Port Royal, Mr. Smith?*

Murtogg: *Yeah, and no lies.*

Jack Sparrow: *Well, then, I confess, it is my intention to commandeer one of these ships, pick up a crew in Tortuga, raid, pillage, plunder and otherwise pilfer my weasely black guts out.*

Murtogg: *I said no lies.*

Mullroy: *I think he's telling the truth.*

Murtogg: *Don't be stupid: if he were telling the truth, he wouldn't have told it to us.*

Jack Sparrow: *Unless, of course, he knew you wouldn't believe the truth even if he told it to you.*

Honest Lies versus Sincere Cheating

The following was an example of **dishonest sincerity**: a "*cheating truth*".

But there are reversed examples, of **honest lies**:

> *Everyone lies online. In fact, readers expect you to lie. If you don't, they'll think you make less than you actually do. So the only way to tell the truth is to lie.*
>
> (Brad Pitt's thoughts on lying about how much money you make on your online dating profile; Aug 2009 interview to "Wired" magazine)

Combining two paradigms: DEL and BR

TOPIC: Logics for reasoning about the **pragmatic aspects of natural language dialogues**, with a stress on the **multi-agent belief revision and knowledge updates induced by communication**.

Methodology: Extend the **DEL (Dynamic Epistemic Logic)** setting to integrate ideas from **BR (Belief Revision theory)**.

Doxastic Attitudes: "static" and "dynamic"

A **"static" doxastic attitude** $A_i\varphi$ captures an agent's *opinion about some sentence* φ:
e.g. agent i **knows** φ;
agent i **believes** φ etc.

In addition to knowledge or belief, we will consider other such attitudes:
strong belief $Sb_i\varphi$,
"**defeasible knowledge**" $\Box_i\varphi$,
knowledge-to-the-contrary $K_i\neg\varphi$ etc.

But there also exist doxastic attitudes of a **"dynamic" nature**, governing an agent's belief revision.

Dynamic Attitudes

The way an agent **revises her beliefs after receiving some new information** depends on the agent's doxastic **attitude towards the source of information**.

This captures the **agent's opinion about the reliability of information coming from this particular source**.

In a "communication" setting, **the sources of information are other agents**.

We will use the notation

$$\tau_{ji}$$

to denote **listener's j's attitude towards information coming from speaker i**.

Dynamic Attitudes are Doxastic Transformers

While formally τ_{ji} will be atomic sentences, the semantics will associate them with **(single-agent) doxastic transformations**:

maps τ taking any input-sentence φ and any single-agent plausibility relation $(S, \leq)$ (of some anonymous agent), and returning a new plausibility relation $(S, \leq')$ on the same set of states.

Given these transformations, the meaning of τ_{ji} is that:

whenever receiving information φ from source i, agent j will/should revise her beliefs by applying transformation $\tau\varphi$ to her plausibility relation $\leq_j$.

The Primacy of Dynamics

Our strategy will be to take the *dynamic*-doxastic attitudes τ_{ji} as **basic** (and simply call them "doxastic attitudes").

Moreover, we will use them to **define** (most of the) *static* attitudes, as **fixed points of the transformations** τ.

Simplifying Restrictions

For this talk (only), we assume that **all communication is public**: speakers make only public announcements to the whole group.

Stable attitudes

For this talk, I will also assume these mutual doxastic attitudes are **stable**:

they do not change during the conversation
(since τ_{ji} are treated as "ontic facts", that are not affected by communication acts).

Later (but maybe not in this talk?), we will consider the issue of **revising agents' doxastic attitudes**.

2. Multi-Agent Plausibility Models

A **multi-agent plausibility model**:

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

with

- S a **finite** 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,
- a *designated state* (the "**actual world**") $s_* \in S$,

subject to a number of **additional conditions**.

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**:

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

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

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

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

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

We read

$$s \leq_a t$$

as: **agent a considers world t to be at least as plausible as world s (but she cannot epistemically distinguish the two).**

Information Partition

For each agent a, the epistemic indistinguishability relation $\sim_a$ induce a *partition* of the state space, called **agent a's information partition**.

It divides the state space S into *mutually disjoint cells*, called **information cells**:

for any state s, **agent a's information cell at s**

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

consists of **all the worlds that are epistemically possible at s (=indistinguishable from s) for agent a.**

Single-Agent Plausibility Models

For a **single-agent** model $\mathbf{S} = (S, \leq, \|\|, s_*)$, **we can assume** (up to bisimilarity $\simeq$) that the plausibility relation is **connected (total)**: just *restrict the model to s_*'s information cell.*

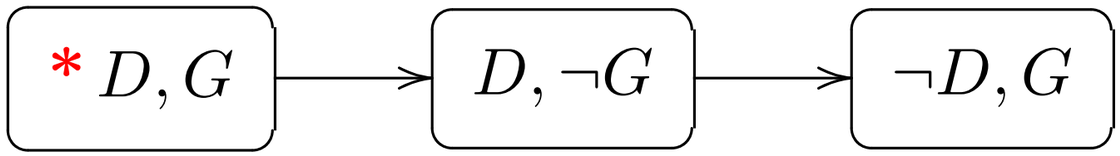
EXAMPLE OF ONE-AGENT MODEL: Prof Winestein

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**.

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) he's drunk or a genius, so we did **NOT** include any world $(\neg D, \neg G)$.

ANOTHER EXAMPLE: Mary Curry

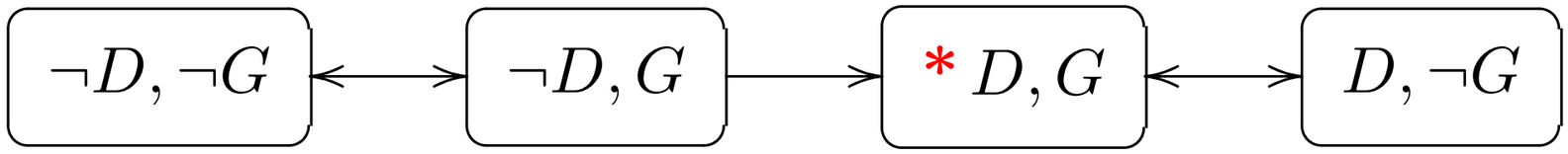
Albert Winestein's best friend is Prof. Mary Curry.

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 single-agent model for Mary alone:



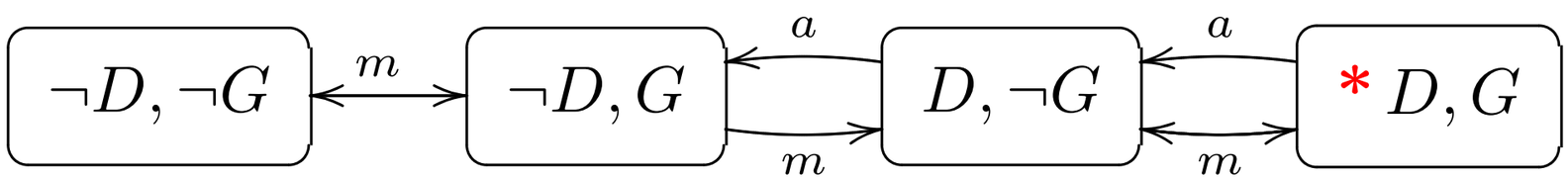
A Multi-Agent Model S

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 following: (1) what is the real world** (i.e. *whether or not Albert is really drunk, and whether or not he is really a genius*), (2) **what are Albert's feelings about being a genius** (i.e. *whether or not Albert feels he is 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 **S**:

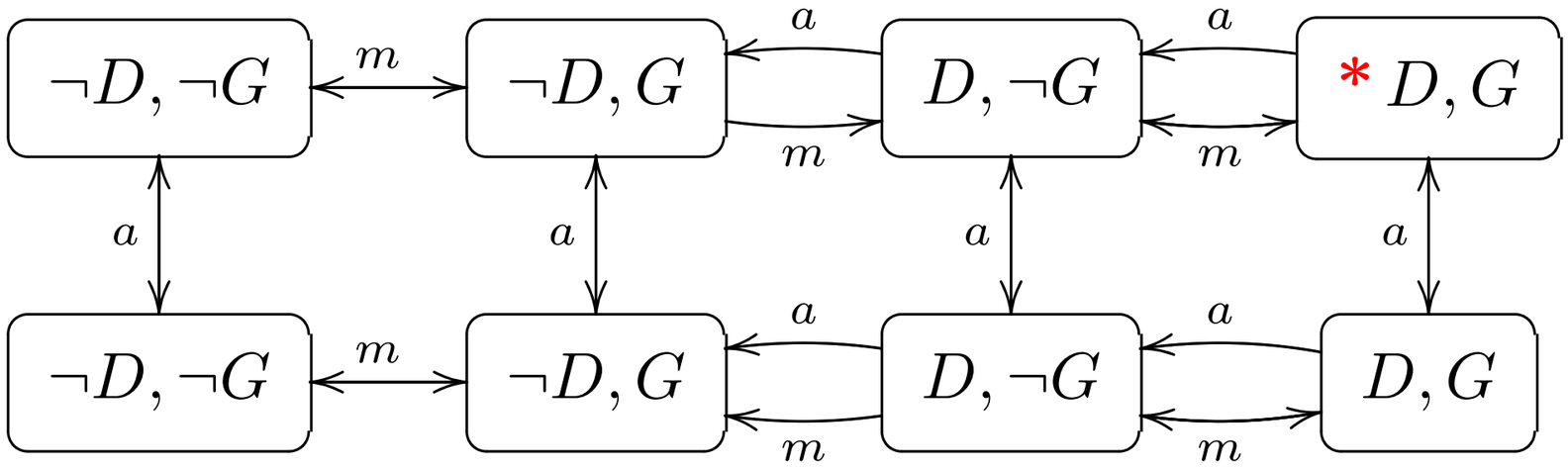


Relaxing the Assumptions: Another Multi-Agent Model

Alternatively, we could of course *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**(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**.

(Irrevocable) Knowledge and (Conditional) Belief

"Irrevocable" Knowledge at a world s is obtained by **quantifying over the worlds that are epistemically possible at** s:

$$s \models K_a\varphi \text{ iff } t \models \varphi \text{ for all } t \in s(a)$$

"Irrevocable Knowledge" is an *absolutely certain, fully introspective and unrevisable* attitude.

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

$$s \models B_a^P\psi \text{ iff } t \models \psi \text{ for all } t \in Max_{\leq_a}(\, P \cap s(a)\,).$$

Epistemology

For logicians and economists, "knowledge" is a very simple notion: the **S5 concept denoted by K, fully introspective** and defined in terms of **equivalence relations** (indistinguishability), or equivalently in terms of information partitions of the state models.

For a **single agent**, it seems even simpler: *we can restrict the state model to the worlds that are indistinguishable from the real one*. So in the single-agent case (as we saw), "knowledge" can be taken to be simply the universal modality: **something is "known" if it is true in all possible worlds**.

Belief

Belief $B_a\varphi$, in the usual logician's and economists' sense, is an equally simple concept: *the special case of conditional belief* $B_a^P\varphi$ in which *the condition* $P = S$ *is true in all worlds (i.e. a tautology).*

B_a is a *normal* modal operator, so it satisfies **Additivity of Belief**

$$B_a\varphi \wedge B_a\psi \;\Rightarrow\; B_a(\varphi \wedge \psi)\,.$$

It also satisfies **Full Introspection**, i.e. both *Positive Introspection*

$$(4)\quad B_a\varphi \Rightarrow B_aB_a\varphi$$

and *Negative Introspection*

$$(5)\quad \neg B_a\varphi \Rightarrow B_a\neg B_a\varphi\,,$$

as well as the axiom of *Consistency of Beliefs*:

$$(D)\quad \neg B_a\bot\,.$$

Unrealistic?

K captures an **absolutely certain** and **fully introspective** type of knowledge.

However, philosophers and linguists argue that this is an **unrealistic** notion, that does NOT match the **common-day usage** of the term "knowledge" in natural language. The intended meaning seems to be **weaker** than our K modality: **less-than-absolutely-certain**.

The Paradox of the Perfect Believer (Voorbraak)

People often **believe they "know" something even when in fact they don't actually know it.**

But this phenomenon *cannot be modeled if we identify "belief" with B, and "knowledge" with K*: **believing you know while not actually knowing is incompatible with the above axioms.**

PROOF: Suppose we'd have $BK\varphi \wedge \neg K\varphi$. Then, by **Negative Introspection**, we have $K\neg K\varphi$. But **knowledge implies belief** (a trivial consequence of our "*Persistence of Knowledge" axiom*), so we have $B\neg K\varphi$. Together with $BK\varphi$ we get, by additivity of Belief, $B(K\varphi \wedge \neg K\varphi)$. But this contradicts **Consistency of Beliefs** (axiom **D**).

(In)Defeasible knowledge

Let us now **define "(in)defeasible knowledge"** $\Box$ by **quantifying over all the worlds that are at least as plausible as (the real world)** s:

$$s \models \Box\varphi \text{ iff } t \models \varphi \text{ for all } t \text{ such that } s \leq t.$$

In other words, **interpret the plausibility order as an "epistemic" relation**: *a world is "epistemically possible" (in the defeasible sense) iff it is at least as plausible as the real world.*

So φ **is "known" in this sense iff it is true in all states that are at least as plausible as** s.

In some of our papers, we called this attitude "**safe belief**".

□ is NOT negatively introspective

Note that this notion of "knowledge" **satisfies Veracity and Positive Introspection** (since $\leq$ is reflexive and transitive), but it **does NOT necessarily satisfy Negative Introspection**.

So this is an *S*4*-type of modality, rather than an S*5 *one*. (More precisely, it satisfies the axioms of the modal system $S4.3$.)

This is OK: it agrees with philosophers' intuition that day-to-day "knowledge" is not always negatively introspective.

"Soft" versus "hard" information

One could say that the fully introspective ($S5$-type) knowledge K captures a notion of '*"hard" information*, that is guaranteed to be truthful beyond any doubt; while the plausibility-based (positively, but negatively, introspective) knowledge $\Box$ captures a more realistic notion of *"soft" information*.

So **irrevocable knowledge embodies "hard information", while (in)defeasible knowledge embodies soft information**.

Their **relative strength** is captured by the entailment:

$$K\varphi \Longrightarrow \Box\varphi,$$

Solving Voorbraak's Puzzle

This allows us to *solve Voorbraak's puzzle*: if we interpret "knowledge" using $\Box$, then the undesirable conclusion that "believing you know is the same as knowing" can no longer be proved, in the absence of Negative Introspection for $\Box$:

$$B\Box\varphi \neq \Box\varphi\,.$$

Of course, this still remains true for K

$$BK\varphi = K\varphi\,.$$

but for $\Box$, one can easily check that we have:

$$B\Box\varphi = B\varphi\,.$$

"Believing you know" in the defeasible sense is the same as simply "believing".

This agrees with our previous conclusion that *"belief", as modeled in plausibility models, is in fact "justified belief"* (or rather, "justifiable" belief). It is "*belief with (relative) certainty*": based on his justification, the agent believes he "knows" φ (in the weak, defeasible sense)

(In)defeasibility versus Irrevocability

Something is "*irrevocable knowledge*" K if it is **a belief that cannot be defeated by any new evidence** (including false information):

$$s \models K_a Q \quad \text{iff} \quad s \models B_a^P Q \text{ for all } P \subseteq S.$$

Something is "*(in)defeasible knowledge*" if it is **a belief that cannot be defeated by any new TRUE evidence**:

$$s \models \Box_a Q \quad \text{iff} \quad s \models B_a^P Q \text{ for all } P \subseteq S \text{ such that } s \in P.$$

Belief, in Terms of "Knowledge"

An important observation, first made by Stalnaker, is that, in a plausibility model, **belief can in fact be defined in terms of "(in)defeasible knowledge"**:

$$BP = \Diamond\Box P\,,$$

where $\Diamond P = \neg\Box\neg P$ is the dual Diamond modality for $\Box$ ("epistemic possibility" in the defeasible sense).

Conditional Belief, in Terms of "Knowledges"

Conditional Belief can also be thus captured in this way, but **ONLY if we use BOTH kinds of "knowledge"**: K and $\Box$.

$$B_a^P Q \quad := \quad \tilde{K}_a P \to \tilde{K}_a (P \wedge \Box_a (P \to Q)) \,,$$

where $\tilde{K}_a P := \neg K_a \neg P$ is the existential dual (Diamond) modality for K.

Strong Belief

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

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

$$\exists w \sim_a s \text{ such that } w \models \varphi$$

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

$$\forall w \sim_a w' \, (w \models \varphi \wedge w' \not\models \varphi \Rightarrow w' <_a w) \, .$$

We write $Sb_a\varphi$ for strong belief. 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**!

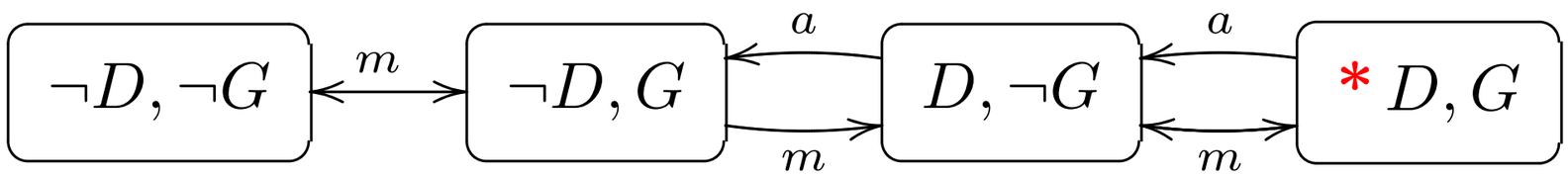
More precisely:
φ is **strongly believed** iff φ **is believed and is also conditionally believed given any new evidence (truthful or not) EXCEPT if the new information is known to contradict** φ; i.e. if:

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

Examples

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

In the Winestein and Mary Curry story



Albert's belief in *genius* is **NOT strong**, and is **NOT "knowledge"**, not even in the (in)defeasible sense $\Box$:

$$(D, G) \models \neg Sb_a G \wedge \neg \Box_a G \,.$$

While *Mary*'s belief that he's *drunk* **IS strong**, and in fact she **(in)defeasibly knows** that he's drunk:

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

The Logic of Knowledge and Safe Belief

The **complete logic of K and $\Box$** is:

- the $S5$-axioms and rules for K;
- $S4$-axioms and rules for $\Box$;
- $KP \to \Box P$;
- $K(P \vee \Box Q) \wedge K(Q \vee \Box P) \ \to \ KP \vee KQ$.

Conditional belief (and thus simple belief) and strong belief can be defined in this system.

3. Doxastic Transformers

Given a (bisimulation-invariant) doxastic language L, a **(single-agent) doxastic transformer** is a *map* τ taking any sentence $\varphi \in L$ and any (single-agent) total plausibility model $\mathbf{S} = (S, \leq, s_*, \|\|)$ into a new total plausibility model $\mathbf{S}' = (S', \leq', s_*, \|\|')$, 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 total preorder $\leq'$ on S'.

We denote by $\tau\varphi$ the map $\tau(\varphi, \bullet)$ induced on single-agent plausibility models.

Examples

(1) Update $!\varphi$ (conditionalization with φ):

executable **only if φ holds** in the real world s_*; in which case,

all the non-φ states are deleted;

and *the same relations are kept between the remaining states.*

(2) Radical Upgrade $\Uparrow \varphi$:

executable **only if there exist φ-worlds in** S; in which case,

all φ-worlds become "better" (more plausible) than all $\neg\varphi$-worlds;

and *within the two zones, the old relations are kept.*

(3) Conservative Upgrade $\uparrow \varphi$:

executable **only if there exist φ-worlds in** S; in which case,

the "best" φ-worlds become better than all other worlds;

all else stays the same.

Different attitudes towards the new information

These correspond to *three different possible attitudes* of the learners towards *the reliability* of the source:

- **Update**: the source is **known to be infallible**.
- **Radical upgrade**: the source is **highly reliable** (or at least **very persuasive**). The source is *strongly believed to be truthful*. This can only happen if the listener doesn't already know that φ is false.
- **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.

More Examples: Negative Attitudes

(4) Negative Update $!^{-}\varphi$: in which case, **all the φ states are deleted** and *the same relations are kept between the remaining states.*

(5) Negative Radical upgrade $\Uparrow^{-} \varphi$:
all $\neg\varphi$-worlds become "better" (more plausible) than all φ-worlds, and *within the two zones, the old ordering remains.* This reflects strong distrust: the listener strongly believes the speaker is lying.

(6) Negative Conservative upgrade $\uparrow^{-} \varphi$:
the "best" $\neg\varphi$-worlds become better than all other worlds, and *in rest the old order remains.* This reflects relative distrust: the listener barely believes the speaker is lying.

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.

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 strongly trust the source to be 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 $\Uparrow_{L_0}\uparrow$.

Example

If the source i is a mathematician, j may accept him as *an infallible source of mathematical statements*, and thus perform an update $!\varphi$ whenever i announces a sentence φ about Mathematics.

In any other case, j might treat the new information more cautiously (say, barely believing it $\uparrow \varphi$, or even ignoring it, and thus applying *id*): indeed, the mathematician i may be utterly unreliable concerning any other area of conversation!

4. Formalizing Doxastic Attitudes

Let us now add to our language two ingredients:

- **dynamic modalities** $[i : \varphi]$
  for **public announcements by agent** i;

- **new atomic sentences**

$$\tau_{ji},$$

for each pair of distinct agents $i \neq j$,
(where τ comes from a given finite set of doxastic attitude symbols, including $!, \Uparrow, \uparrow, id$ etc),
encoding the **agent j's attitude towards agent i's the announcements**.

Semantics

For semantics, we are given a multi-agent plausibility model, with the valuation map extended to the new atomic sentences and satisfying a number of semantic conditions (to follow);
and, in addition, we are also given, for each attitude symbol τ in the syntax, some single-agent *doxastic transformation*, also denoted by τ.

Semantic Constraints

We put some natural semantic constraints (on the valuation of the atomic sentences of the form τ, for any doxastic attitude type τ).

The first says that, in any possible world, **every agent has some unique attitude towards every other agent**:

$$\forall s \exists! \tau \text{ such that } s \models \tau_{ji}$$

The second is an **introspection-type** postulate: the agent **knows her own doxastic attitudes**

$$s \overset{j}{\sim} t \implies (s \models \tau_{ji} \iff t \models \tau_{ji}).$$

A further simplifying assumption

In fact, in most of this talk we will make an even more restrictive assumption, namely:

all the agents' doxastic attitudes towards each other are common knowledge:

$$\forall s, t : \; s \models \tau_{ji} \iff t \models \tau_{ji}$$

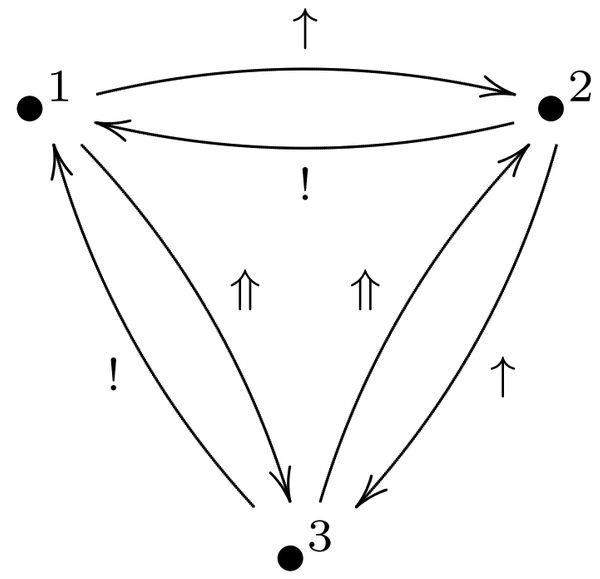
The Mutual Trust Graph

In the presence of this last simplifying assumption, the extra-structure required for the semantics (i.e. the extension of the valuation plus the assignment of a transformation to each attitude symbol) can be summarized as a **graph having agents as nodes and arcs labeled by doxastic transformations**:

the fact that τ_{ji} **holds** in (any state of) the model is captured by **an arc labeled τ from node j to node i.**

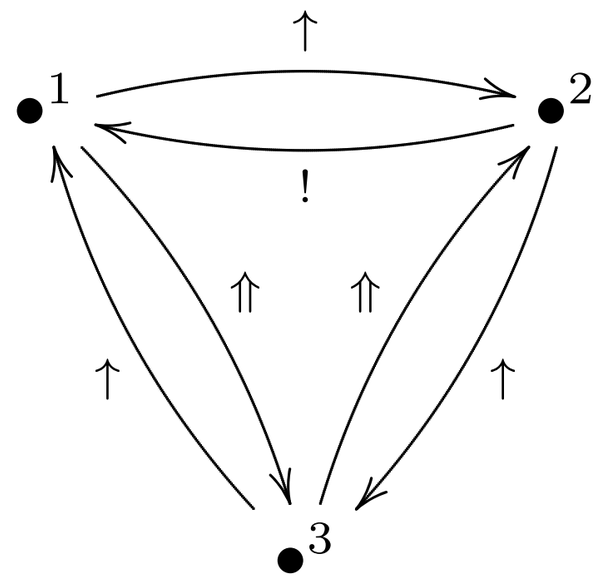
This graph will be called **the "mutual trust graph" of this group of agents**.

Example



Counterexample

Not all graphs are consistent with the assumption that all doxastic attitudes are common knowledge:



is NOT a consistent graph.

Consistency Conditions

$$!_{ji} \Rightarrow !_{ki},$$

$$!^{-}_{ji} \Rightarrow !^{-}_{ki},$$

for all $j, k \neq i$.

Semantics of Communication Acts

The semantics of $i : \varphi$ will be given by the multi-agent doxastic transformation that takes any plausibility model $\mathbf{S} = (S, \leq_j, \|\|)_j$ and returns a new model $(i : \varphi)(\mathbf{S}) := (S', \leq'_j, \|\|')$, where:

the **listeners' new preorders** $\leq_j$ (for $j \neq i$) are given **by applying within each $\sim_j$-information cell the transformer τ** to the order $\leq_j$ within that cell, **where τ is the unique attitude such that τ_{ji} holds** throughout that cell;

while **the speaker's preorder $\leq_i$ is kept the same**;

the **new set of worlds S' is the union of all the new information cells**;

and the **new valuation** $\|p\|' := \|p\| \cap S'$.

Semantics of Dynamic Modalities

The **dynamic modalities** are defined as usual:

$$s \models \mathbf{s}\ [i:\varphi]\psi \text{ iff } s \models_{i:\varphi(\mathbf{s})} \psi.$$

So $[i:\varphi]\psi$ means that:

if i publicly announces φ, then ψ holds after that.

Sincerity of a Communication Act

A communication act $i : \varphi$ is **sincere** if **the speaker believes her own announcement**; i.e. if $B_i\varphi$ holds.

Common Knowledge of Sincerity

In cooperative situations, it is sometimes natural to assume **common knowledge of sincerity**.

This can be done by **modifying the above semantics** of $i : \varphi$, by *restricting the applicability* of the above doxastic transformation (modeling the act $i : \varphi$) to states in which $B_i\varphi$ holds.

We call this an *"inherently sincere" communication act.*

Sincere Lies and (Lack of) Introspection

In general, **sincerity does not imply truthfulness**:
"*I really am the man of your dreams*" is a typical **sincere lie**!

But, when applied to **introspective** properties, sincerity DOES imply truthfulness:
"*I believe I am the man of your dreams*" **is sincere only if it's true.**

Another Mixed Attitude

This observation suggests a natural mixed attitude $!\Uparrow_i$: *the listener strongly trusts the speaker i* to tell the truth, but moreover *she considers the speaker to be infallibly truthful when announcing sentences* (such as "I believe.." or "I know...") *that he can know by introspection.*

This attitude is natural when common knowledge of sincerity is assumed: when applied to introspective properties, sincerity is the same as infallible truthfulness, hence ! is appropriate; for other properties, sincerity can at least be (in the case of a highly reliable source) a warranty of high plausibility.

Super-Strong Trust $!\Uparrow_{ji}$

One can show that all i-introspective sentences are equivalent to ones of the form $K_i\varphi$.

Hence, the attitude $!\Uparrow_{ji}$ can be described as follows: if i announcement φ is (equivalent to) an i-introspective sentence of the form $K_i\varphi$, then j applies an update $!\varphi$; otherwise, he performs an upgrade $\uparrow\varphi$.

Let us call this attitude **super-strong trust**.

Another Description

The attitude $!\Uparrow_{ji}$ can also be described in more semantical terms: after i announces a sentence φ, j will apply an **update** $!\varphi$ iff **she knows that** j **knows** φ**'s truth-value**, i.e. iff we have $K_j(K_i\varphi \vee K_i\neg\varphi)$;
otherwise, j will perform a **radical upgrade** $\Uparrow\varphi$.

5. Static Attitudes as Fixed Points

To each (dynamic) doxastic attitude given by a transformer τ, we can associate a static attitude $\overline{\tau}$.

We write

$$\overline{\tau}_{\mathbf{i}}\varphi$$

and say that **agent i has the attitude $\overline{\tau}$ towards φ**, if i**'s plausibility structure is a fixed point of the doxastic transformation** $\tau\varphi$.

Formally,

$$s \models_{\mathbf{S}} \overline{\tau}_i\varphi \text{ iff } \tau\varphi(S, \leq_i, s) \simeq (S, \leq_i, s)$$

(where $\simeq$ is the bisimilarity relation).

Examples

The fixed point of update is "irrevocable knowledge" K:

$$\bar{!}_j\varphi \;\Leftrightarrow\; K_j\varphi$$

and similarly

$$\overline{!^-}_j\varphi \;\Leftrightarrow\; K_j\neg\varphi\,.$$

The fixed point of radical upgrade is strong belief Sb:

$$\overline{\Uparrow}_j\varphi \;\Leftrightarrow\; Sb_j\varphi$$

and similarly

$$\overline{\Uparrow^-}_j\varphi \;\Leftrightarrow\; Sb_j\neg\varphi\,.$$

The fixed point of conservative upgrade is belief B:

$$\overline{\uparrow}_j\varphi \;\Leftrightarrow\; B_j\varphi$$

and similarly

$$\overline{\uparrow^-}_j\varphi \;\Leftrightarrow\; B_j\neg\varphi\,.$$

The fixed point of identity is tautological:

$$\overline{id}_j\varphi \;\Leftrightarrow\; true\,.$$

Explanation

The **importance of fixed points** $\overline{\tau}$ is that **they capture the attitudes** (towards the sentence φ) **that are induced in an agent after receiving information** (φ) **from a source towards which she has the attitude** τ:

indeed, $\overline{\tau}_j$ is the strongest attitude such that

$$\tau_{ji} \Rightarrow [i : p]\overline{\tau}_j p\,,$$

for all non-doxastic sentences p.

More generally, after an agent with attitude τ towards a source upgrades with information φ from this source, she comes to have attitude $\overline{\tau}$ towards the fact that φ WAS the case (before the upgrade):

$$\tau_{ji} \Rightarrow [i:\varphi]\overline{\tau}_j(BEFORE\varphi)\,,$$

where $BEFORE$ is a one-step past tense operator.

Explanation continued: examples

Conservative upgrades induce simple beliefs:

after $\uparrow \varphi$, the agent only comes to **believe** that φ (was the case).

Radical upgrades induce strong beliefs:

after $\Uparrow \varphi$, the agent comes to **strongly believe** that φ (was the case).

Updates induce (irrevocable) knowledge:

after $!\varphi$, the agent comes to **know** that φ (was the case).

Fixed Points and Redundancy

A fixed point expresses the **"redundancy"**, or **un-informativity**, of a doxastic transformation: in this sense, we can say that the fact that a fixed-point-attitude is induced in the listener by an announcement captures the fact that *repeating the announcement would be redundant*.

This is literally true for non-epistemic sentences:

$$[i:p][i:p]\theta \Leftrightarrow [i:p]\theta\,,$$

but can be generalized to arbitrary sentences in the form:

$$[i:\varphi][i:BEFORE\varphi]\theta \Leftrightarrow [i:\varphi]\theta\,.$$

Honesty

We say that a communication act $i : \varphi$ is **honest** **(with respect) to a listener** j, and write

$$Honest(i : \varphi \to j)\,,$$

if **the speaker** i **has the SAME attitude towards** φ (before the announcement) **as the one (he believes to be) induced in the listener** j **by his announcement** of φ.

By the above results, it seems that, if τ_{ji} holds then honesty should be given by $\overline{\tau}_i\varphi$.

This is indeed the case ONLY if we adopt the simplifying assumption that all doxastic attitudes τ_{ji} are common knowledge.

But, in general, honesty depends only on (having) the attitude that the speaker **believes** to induce in the listener:

$$Honest(i:\varphi \to j) = \bigwedge_{\tau}(B_i\tau_{ji} \Rightarrow \overline{\tau}_i\varphi)\,.$$

General Honesty

A (public) speech act $i : \varphi$ is **honest** iff it is *honest (with respect) to all the listeners*:

$$Honest(i : \varphi) := \bigwedge_{ji\not} Honest(i : \varphi \to j).$$

Example: honesty of an infallible speaker

Assume that it is **common knowledge** that a speaker i is **infallible** (i.e. that $!_{ji}$ holds for all $j \neq i$).

Then *an announcement* $i : \varphi$ *is* ***honest*** *iff the speaker* ***knows*** *it to be true; i.e. iff* $K_i\varphi$ *holds.*

The same condition ensures that the announcement $i : K_i\varphi$ is honest.

Example: honesty of a "barely trusted" speaker

Assume **common knowledge** that a speaker i is only **barely trusted** (i.e. that $\uparrow_{ji}$ holds for all $j \neq i$).

Then *an announcement* $i : \varphi$ *is* ***honest*** *iff the speaker* ***believes*** *it to be true; i.e. iff* $B_i\varphi$ *holds.*

The same condition ensures that *the announcement* $i : B_i\varphi$ *is honest.*

Example: honesty of a strongly trusted speaker

Assume **common knowledge** that a speaker i is **strongly trusted, but not infallible** (i.e. that $\Uparrow_{ji}$ holds for all $j \neq i$).

Then *an announcement* $i : \varphi$ *is* ***honest*** *iff the speaker* ***strongly believes*** *it to be true; i.e. iff* $Sb_i\varphi$ *holds.*

Dis-honest, Truthful Sincerity: Jack Sparrow

Example 1: Suppose an agent i (call him Jack Sparrow) *strongly believes* P, and in fact he *knows* P: so P is actually true. E.g. P is the sentence saying he came to commandeer a ship, raise a crew and then rape, pillage and plunder.

Suppose i knows that he is *strongly distrusted* by his audience j, i.e. we have $\Uparrow^{-}_{ji}$. ("He knows that you won't believe him").

Then the the announcement $i : P$ is **sincere and truthful, but still dis-honest**!

This shows that **sincerity and truth, even taken together, do NOT imply honesty.**

Honest Lies: Brad Pitt

QUESTION: But how can such a strongly distrusted speaker be honest?

ANSWER: By telling lies!

Example 2: This is the same situation as in previous Example 1, except that the speaker i (now call him Brad Pitt) **announces the opposite of what he believes/knows**.

The announcement $i : \neg P$ is an "honest lie" in this situation: insincere and un-true, but conveying truthful information and a correct attitude to the audience!

Honesty of a strongly distrusted speaker

Assume **common knowledge** that a speaker i is **strongly distrusted** (i.e. that $\Uparrow^{-}_{ji}$ holds for all $j \neq i$).

Then *an announcement* $i : \varphi$ *is* ***honest*** *iff the speaker* ***strongly believes*** *it to be* ***false****; i.e. iff* $Sb_i\neg\varphi$ *holds.*

This shows that **honesty does not imply sincerity** either!

Nevertheless, **when (it is common knowledge that) the listeners have a** *"positive"* **attitude towards the speaker (one that implies belief), then honesty does implies sincerity!**

Dis-honest Sincerity, again: George W. Bush

Let us get back to our **strongly trusted** speaker i, who **only believes (but does NOT strongly believe)** φ.

Example 3: φ is the sentence saying that "*There are weapons of mass destruction in Irak*". Agent i is in fact called *George W. Bush*. He has no strong evidence, but only hearsay evidence for φ. He (barely) believes this evidence. But i knows that he's *strongly trusted* by his audience j ("the American People"), i.e. we have $\Uparrow_{ji}$.

Then the announcement $i : \varphi$ ("There are weapons of mass destruction in Irak") would be **sincere but dis-honest**: indeed, this announcement induces in the listeners a **strong belief** in φ, while Bush did **not** have any such a strong belief himself (but only a simple belief)!

What can Bush honestly and sincerely announce?

Well, he might announce that he **believes** φ:

"*We believe there are weapons of mass destruction in Irak*".

The announcement $i : B_i\varphi$ is certainly *sincere*, since $B_i\varphi$ holds.
It is also *honest*, since $Sb_iB_i\varphi$ holds whenever $B_i\varphi$ holds.

But is this **persuasive**: is it enough to convince the American people to go to war?

Letting this issue aside: in fact, he *CAN honestly claim much MORE*! He can claim that he "**(indefeasibly) KNOWS**" (= safely believes) φ:

the act $i : \Box_i\varphi$ is honest (for a strongly trusted speaker) iff $B_i\varphi$ holds.

Honest Exaggerations

"*We KNOW there are weapons of mass destruction in Irak*".

This might sound like a wild exaggeration on Bush's part, but if we interpret it as referring to (in)defeasible knowledge $\Box$, then this is a **sincere** announcement: indeed, by the identity

$$B_i \Box_i \varphi = B_i \varphi$$

we have that, whenever i believes φ, he also believes $\Box_i \varphi$.

But moreover, this is also an **honest** announcement, since **belief implies strong belief that you (defeasibly) "know"**; in fact, **the two are equivalent**:

$$B_i \varphi \;=\; Sb_i(\Box_i \varphi)\,.$$

6. How can we convince the others

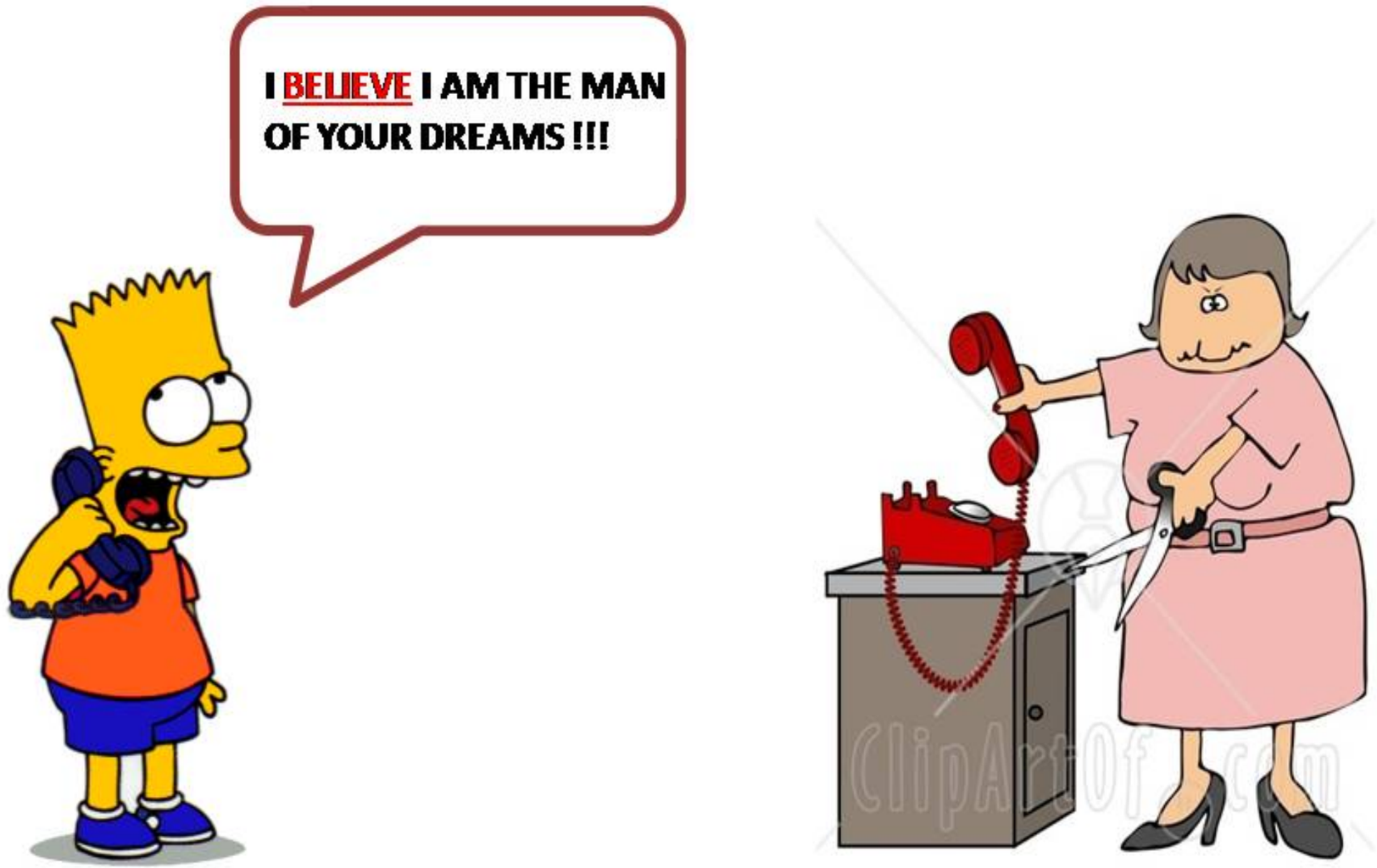
EXAMPLE 4: Bart really **believes that he's the man of Jessica's dreams**, but Jessica ignores him. **What should he say to "convert" her to his belief?**

If true and if **addressed to a listener with a "positive" attitude**, the statement

"*I believe I am the man of your dreams*"

is **guaranteed to be honest and sincere**.

But is it persuasive? Will the girl buy it?!



How can Bart "convert" others to his beliefs, while still being honest and sincere?

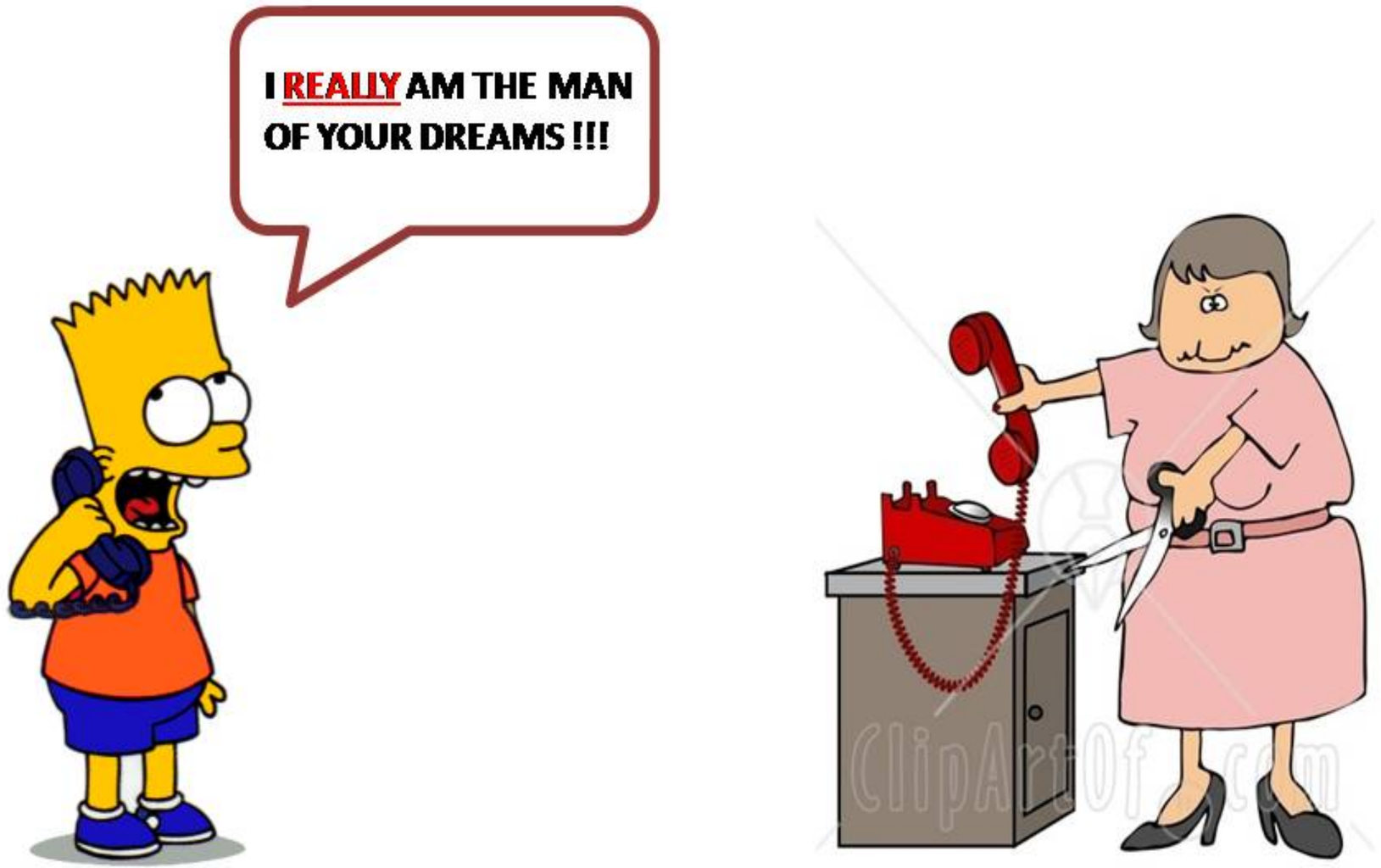
Jessica's natural answer could well be:

> "*It's OK, Bart: I believe that* **you** *believe it.*
> *But* **I** *still* **DON'T** *believe it*!"

This is a **positive** attitude: *Jessica buys into Bart's sincerity.* Let's assume that in fact she *strongly believes* in his sincerity.

But this is NOT the attitude that Bart *wants* to induce in her: namely, the **SAME attitude as his own** attitude **towards the issue itself** (of whether or not he's the man of her dreams)!

And (as for Bush!) **announcing the fact itself won't do**: it'd be **dis-honest**, since Bart *DOESN'T strongly believe* it. He's not THAT sure of himself!



How can Bart "convert" others to his beliefs, while still being honest and sincere?

Persuasiveness

We say that **an announcement $i : \varphi$ is persuasive to a listener j with respect to an issue θ,**
and we write $Persuasive(i : \varphi \rightarrow j; \theta)$,
if
the effect of the announcement is that **the listener is "converted" to the speaker's attitude towards θ.**

Formally, for non-doxastic sentences:

$$Persuasive(i : \varphi \rightarrow_j ; p) := \bigwedge_{\tau} (\overline{\tau}_i \varphi \Rightarrow [i : \varphi] \overline{\tau}_j p) \ .$$

For doxastic sentences, this needs to be modified:

$$Persuasive(i : \varphi \rightarrow_j ; \theta) := \bigwedge_{\tau} (\overline{\tau}_i \varphi \Rightarrow [i : \varphi] \overline{\tau}_j (BEFORE\theta))$$

How to be Honest, (Sincere) and Persuasive

Suppose a **strongly trusted** agent i wants to be **honest, but persuasive with respect to some issue θ that he believes in (although he does not necessarily have a strong belief)**. So we assume $\Uparrow_{ji}$ for all j, and $B_i\theta$ (but NOT necessarily $Sb_i\theta$).

QUESTION: **What can i announce honestly (and thus sincerely), in order to be persuasive?**

This is a very important question:
how can you "convert" others to your (weak) beliefs, while still maintaining your honesty and sincerity?

What NOT to say

$i : \theta$ would be sincere and persuasive, but it's **dishonest** (unless $Sb_i\theta$ holds)!

$i : K_i\theta$ is equally **dishonest**.

$i : B_i\theta$ is honest, but **not persuasive**: it won't change j's beliefs about θ (although it will change her beliefs about i's beliefs).

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

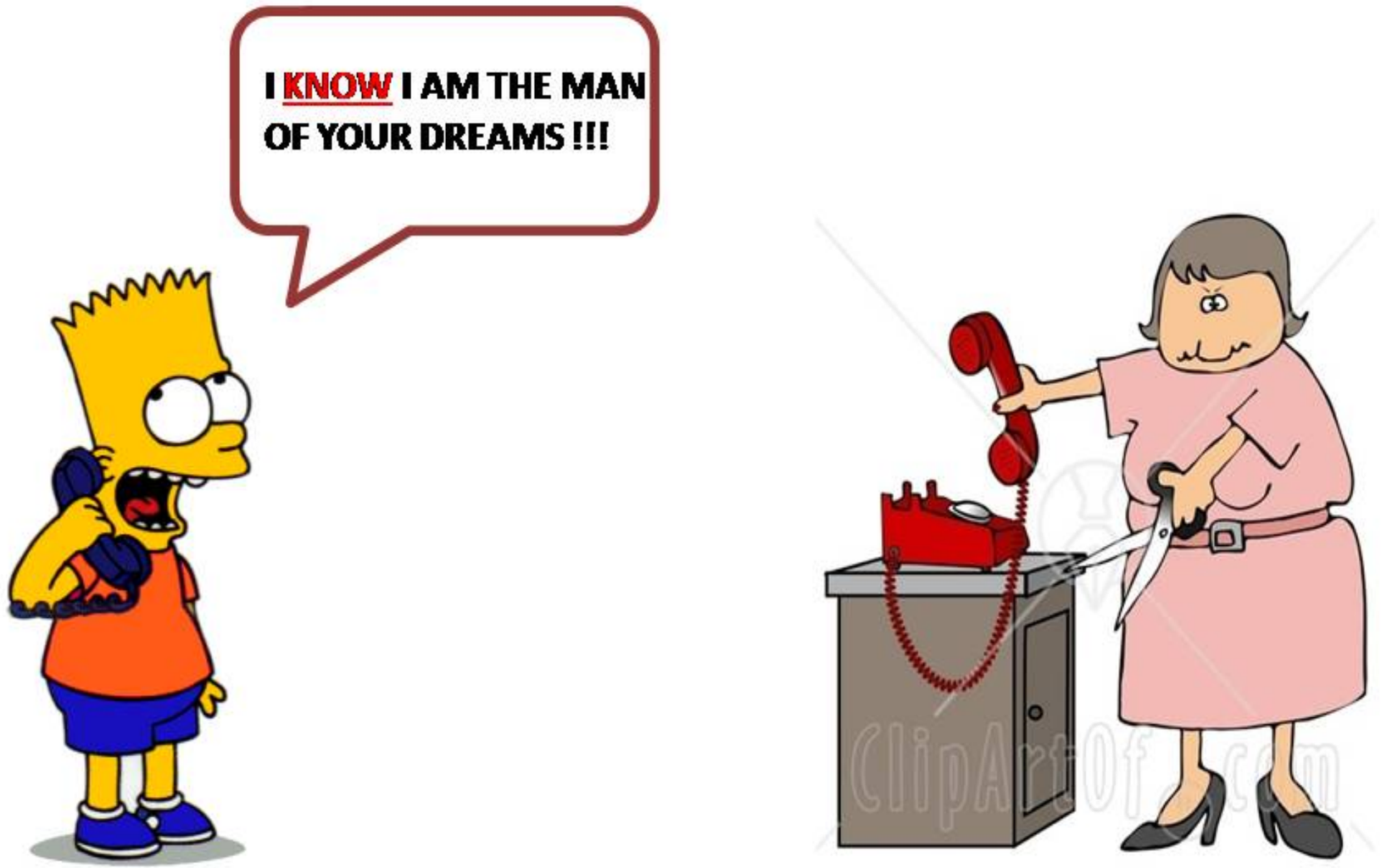
ANSWER: honest exaggerations are persuasive!

It turns out that the only honest and persuasive announcement in this situation is to make an "*honest exaggeration*":

$$i : \Box_i \theta.$$

In other words:

to honestly convert others to your belief in θ, you need to say that you "defeasibly know" θ (when in fact you only believe θ).



How can Bart "convert" others to his beliefs, while still being honest and sincere?

Conclusion

THE LESSON (known by most successful politicians):

If you want to convert people to your beliefs, don't be too scrupulous with the truth:
Announce that **you "know"** things **even if you don't know for sure that you know them**!
History will deem you "honest", as long as you... believed it yourself!

Simple belief, a loud voice and strong guts is all you need to be persuasive!

"*We now know that Saddam Hussein has acquired weapons of mass destruction...*" (G.W. Bush, 2002)

Oh, well..., and you also need suckers to strongly trust you...

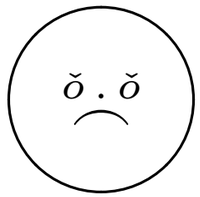
7. Reaching Doxastic Agreement

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

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

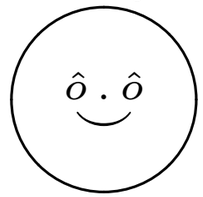
Example of a particular scenario:

Albert



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

Mary



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

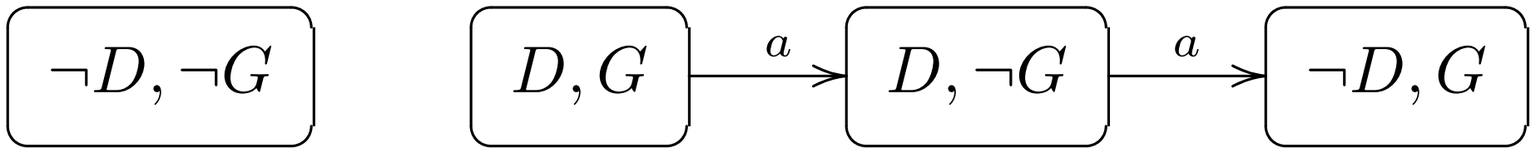
They share their information

$\Downarrow$

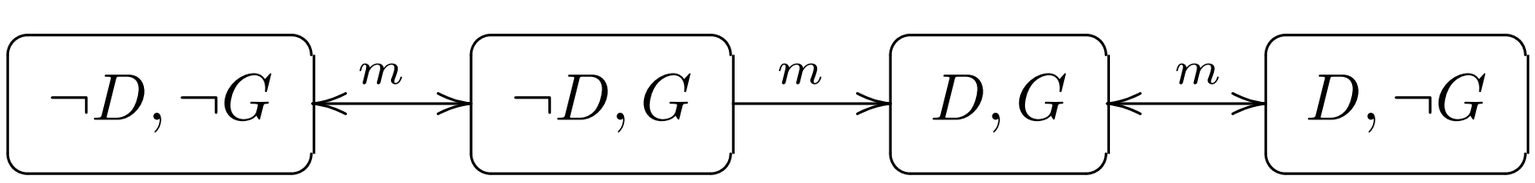
Together they know the same: (D or G) and both believe D and $\neg G$

A Multi-Agent Model S

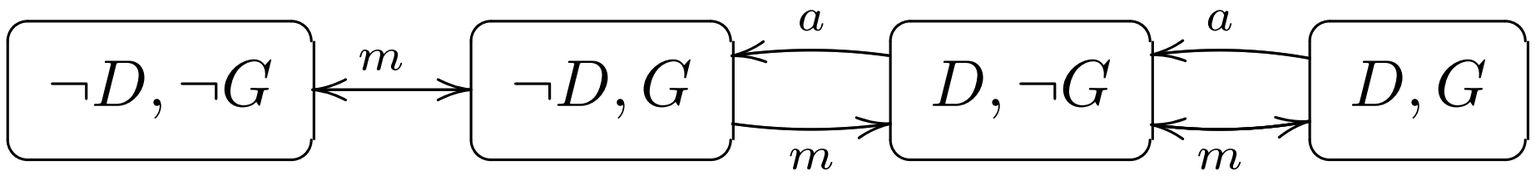
Albert's plausibility order (and information cells):



Mary's plausibility order:



The multi-agent plausibility model **S**:



Main Issues:

• **Agents' goal** = Reach a total doxastic/epistemic agreement ("merge").

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

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

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

Speak loud, be positive, honest (sincere) and persuasive!

Rules of the game for our agents:

- **Fully Public communication:**

- *common knowledge of what* (the message) is announced

- *common knowledge of the fact that all agents adopt the same attitude towards other agents.*

- And we assume that this common attitude is a **"positive" one** (i.e. *it implies belief*), and in particular we'll focus on the attitudes: !, $\Uparrow$ and $!\Uparrow$.

- **Honesty:**

- the speaker should have the same attitude towards the announced information that he expects to induce in the listeners.

- When the group's common attitude is positive, *honesty will imply sincerity.*

- **Persuasiveness**:

- listeners come to share ***the same attitude as the speaker*** *towards the relevant issues.*

• **(Common Knowledge that) All Agents Have the SAME (positive) doxastic attitudes towards all other agents**:

- *all the labels in the trust graph are identical (and positive)*.

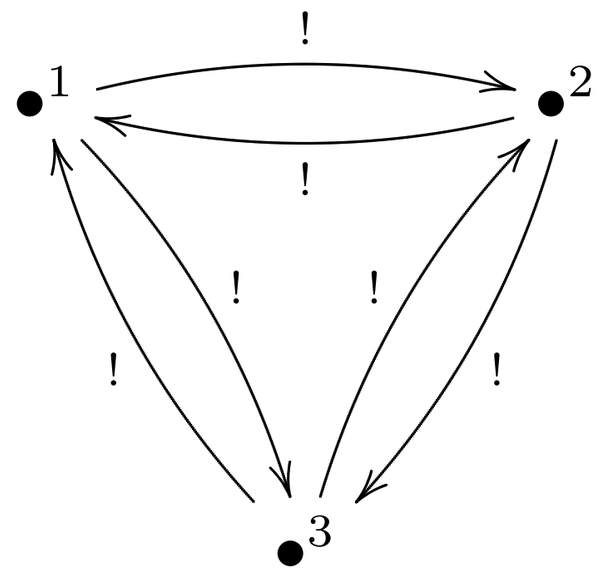
- Let τ be this same doxastic attitude. Then we introduce a sentence

$$\tau := \bigwedge_{i \neq j} \tau_{ji}$$

saying that all agents adopt attitude τ towards all other agents.

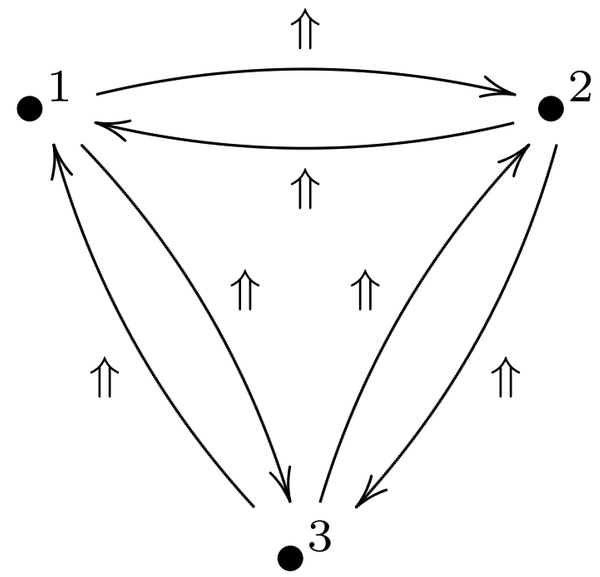
Example 1: (Common Knowledge of) Infallibility

For three agents, ! holds if the graph is



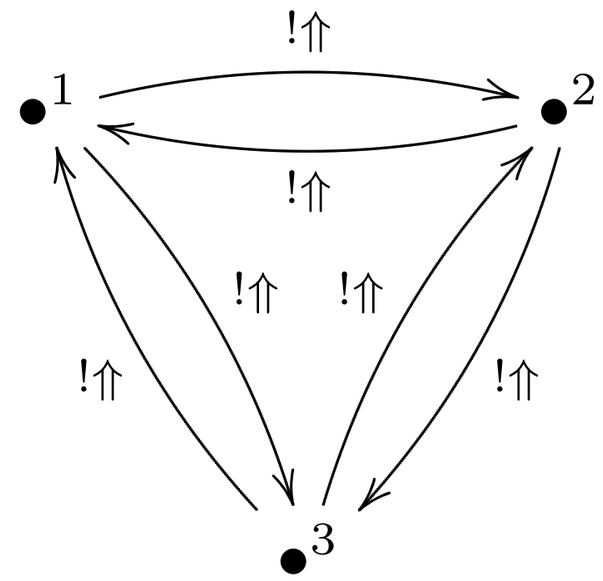
Example 2: (Common Knowledge of) Strong Trust

$\Uparrow$ holds if the graph is



Example 2: (Common Knowledge of) Super-Strong Trust

$!\Uparrow$ holds if the graph is



Goal of Sharing is Total Agreement

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

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

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

Dynamic Merge

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

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

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

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

8. Preference Merge and Belief Aggregation

In Social Choice Theory: the main issue is how to *merge* the agent's individual preferences.

A **merge operation for a group** G is a function $\bigodot$,

taking preference relations $\{R_i\}_{i \in G}$ into
a *"group preference" relation* $\bigodot_{i \in G} R_i$ (on the same state space).

Merge Operations

• So the problem is to find a **"natural" merge operation** (subject to various *fairness conditions*), for merging the agents' preference relations.

• Depending on the conditions, one can obtain either
- an **Impossibility Theorem** (*Arrow* 1950) or
- a **classification of the possible types of merge** operations (*Andreka, Ryan & Schobbens* 2002).

Belief Merge and Information Merge

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

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

• To merge the agents' **soft information** (*all their "strong beliefs"* Sb_i*, or equivalently all their "conditional beliefs"* $B_i^P Q$), we have to **merge the plausibility relations** $\leq_i$.

Merge by Intersection

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

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

In the case of two agents, it takes:

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

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

Distributed Knowledge is Parallel Merge

• **This form of merge is well suited for "knowledge"** K: since this type of knowledge is absolutely certain, there is no danger of inconsistency.

• The agents pool their information in a *completely symmetric manner, accepting the other's bits without reservations.*

Parallel merge of the agents' irrevocable knowledge gives us the standard concept of "distributed knowledge" DK:

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

Lexicographic Merge

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

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

Formally:

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

Lexicographic merge of soft information

Lexicographic merge is particularly suited for "soft information", given by either *strong beliefs* Sb or *conditional beliefs* B^P, **in the absence of any hard information**:

since soft information is not fully reliable, some "screening" must be applied (and so some hierarchy must be enforced) to ensure consistency of merge.

Relative Priority Merge

Note that, **in lexicographic merge, the first agent's priority is "absolute"**.

But **in the presence of hard information, the lexicographic merge of soft information must be modified**:

by **first pooling together all the hard information and then using it to restrict the lexicographic merge of soft information.**

This leads us to a "more democratic" **combination of Merge by Intersection and Lexicographic Merge** , called "**(relative) priority merge**" $R_{a\otimes b}$:

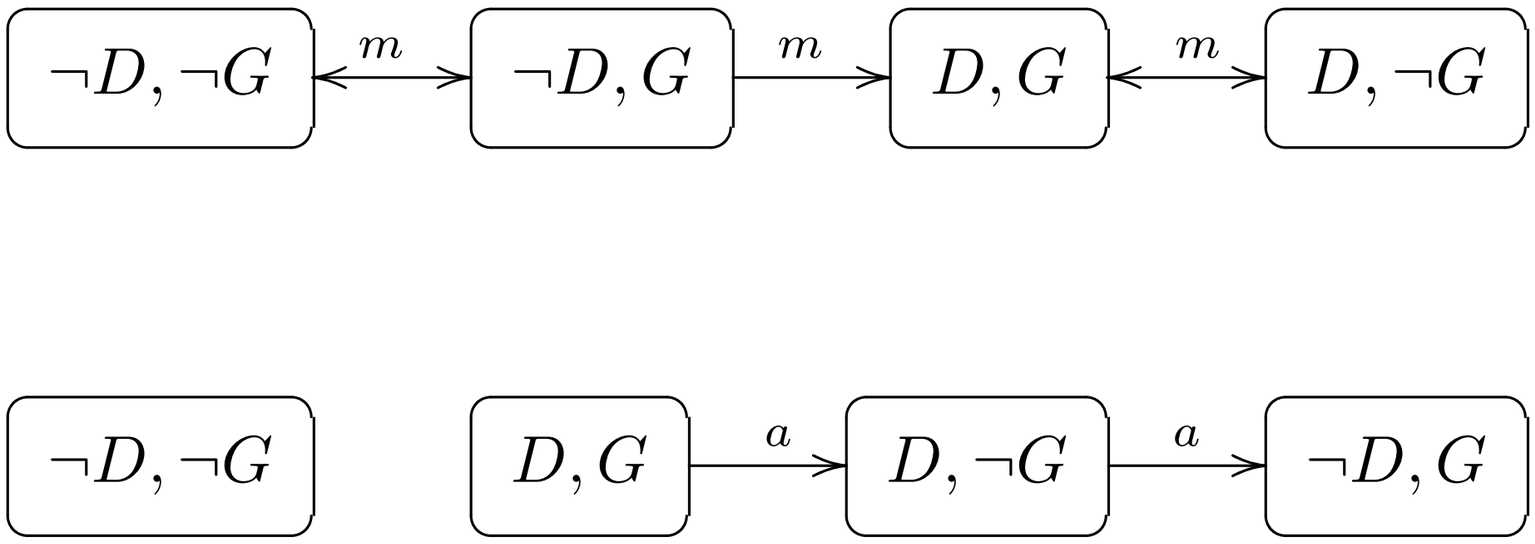
$$R_{a\otimes b} := (R_a^{>} \cap R_b^{\sim}) \cup (R_a^{\cong} \cap R_b) = R_a \cap R_b^{\sim} \cap (R_a^{>} \cup R_b).$$

This means that **both agents have a "veto" with respect to group incomparability**:

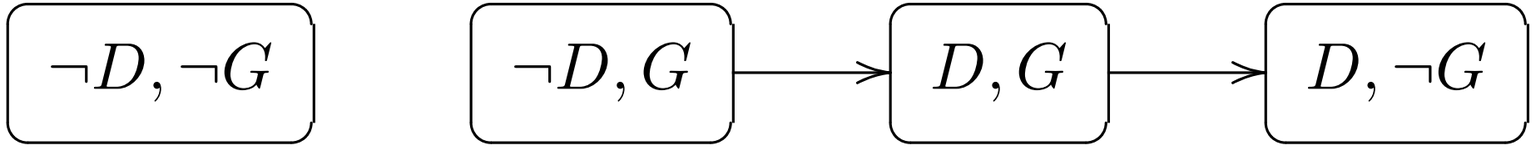
The group can only compare options that **both** agents can compare; **and whenever the group can compare two options, everything goes on as in the lexicographic merge**: agent a's strong preferences are adopted, while b's preferences are adopted only when a is indifferent.

Example: merging Marry's beliefs with Albert's

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



gives us:



9. "Realizing" Information Merge Dynamically

Intuitively, the **purpose** of "preference merge" $\bigodot_{i\in G} R_i$ is to achieve a state in which the G-agents' preference relations are "**merged**" accordingly, i.e.

to perform a sequence π *of communication acts, transforming the initial model* $(S, R_i)_{i\in G}$ into a model $(S, R'_i)_{i\in G}$ such that

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

for all $j \in G$.

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

Realizing Distributed Knowledge

In the case of **knowledge**, it is easy to **design a protocol to realize the parallel merge of agents' knowledge** by a *sequence of joint updates*, IF we assume **(common knowledge of) infallibility** !:

PROTOCOL:

Assume !. Then, in no particular order, the agents have to publicly and sincerely announce "all that they know" .

Realizing Distributed Knowledge

More precisely:

- *a communication act* $\mathbf{a : K_a P}$ *is performed, for each set of states* $P \subseteq S$ *such that* P *is (or comes to be) known to a given agent* a (after the previous communication acts).

- This essentially is the algorithm in Johan van Benthem's paper "One is a Lonely Number".

The Protocol

Formally, if $(a_1, \ldots, a_n)$ is some *arbitrary listing of all agents in* G *without repetitions*, then **a protocol for realizing distributed knowledge within group** G **at state** s**, given attitude** ! is:

$$\pi := \prod_{i=1,n} \rho_i$$

where $\prod$ is sequential composition of a sequence of actions and

$$\rho_i := \prod\{(a_i : K_{a_i} P) : P \subseteq S \text{ such that } s \models [\prod_{j=1,i-1} \rho_j] K_{a_i} P\}$$

NOTE: *The* ***order of the agents*** *in the first* $\prod_i$ *and the* ***order in which the announcements are made*** *by each agent (in the second* $\prod$*) are* ***arbitrary***.

Realizing Priority Merge

If we assume (common knowledge of super-strong trust)

$$!\Uparrow$$

(as the common attitude), then we can **realize the Priority Merge**

$$\bigotimes_i \leq_i$$

of the whole PLAUSIBILITY ORDERS (encoding BOTH SOFT AND HARD INFORMATION) by simply executing *the following protocol*:

The Protocol

PROTOCOL:

Assume ! $\Uparrow$ and some given priority order.
Then, **respecting the priority order**, each agent has to **publicly and sincerely announce that she "defeasibly knows" all that she believes to "know".**

Here, "knowledge" means now **defeasible knowledge** $\Box_a$.

Order-dependence

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

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

No interruptions, please!

The Formalization of the Protocol

Formally, if $(a_1, \ldots, a_n)$ is a **listing of all agents in descending priority order**, the **protocol π' for realizing priority merge** of plausibility relations $\{\leq_a\}_{a \in G}$ **at state s, given common attitude** $\Uparrow$, is the following:

$$\pi' := \prod_{i=1,n} \rho'_i$$

where

$$\rho'_i := \prod \{(a_i : \Box_{a_i} P) : P \subseteq S \text{ such that } s \models [\prod_{j=1,i-1} \rho'_j] B_{a_i} P\}.$$

Here, *the order $(i_1, \ldots, i_k)$ in the first $\prod_i$ is the priority order, while the order of announcements in the second $\prod$ is still arbitrary.*

Be Persuasive!

Note:

- *simply announcing 'what' they believe would be DISHONEST*,

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

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

What we need for belief merge is:

- That each agent tries **to be persuasive**:

- *to "convert" the other to her own beliefs* by **announcing** $\Box_a\varphi$ **when they just believe** φ (and hence they believe that $\Box_a\varphi$).

- This may look like a form of **deceit**, or at least a *"deliberate exaggeration"* (though maybe not an outright lie) ???

Honesty

But, as we already saw:

- such a **communication act** $a : \Box_a P$ is **sincere** (since $B_a \Box_a P$, i.e. $B_a P$, holds),

- and moreover it is **honest**: since we have $B_a P$ implies $Sb_a \Box_a P$, hence a's statement is strongly believed by her.

As a result, this communication act **converts the listeners to the SAME doxastic attitude** towards $\Box_a P$ as the speaker (a) had:

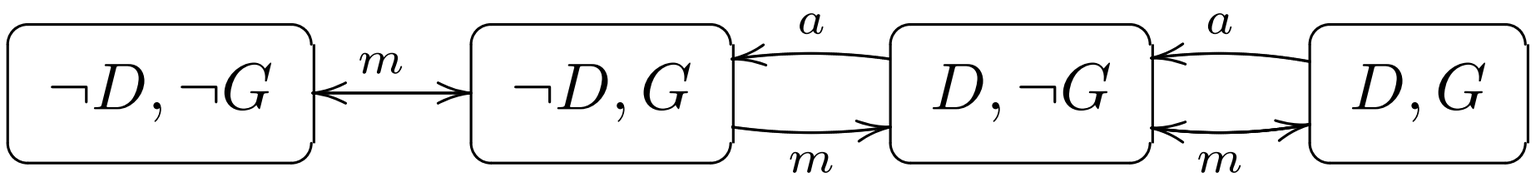
they all will strongly believe that $\Box_a P$ was true.

As a consequence, if the speaker announces all such $\Box_a$'s (that he believes to hold), he will end up **converting all listeners to all his beliefs and strong beliefs**:

this is a perfect act of **"honest persuasion" by a sequence of "sincere exaggerations"**!

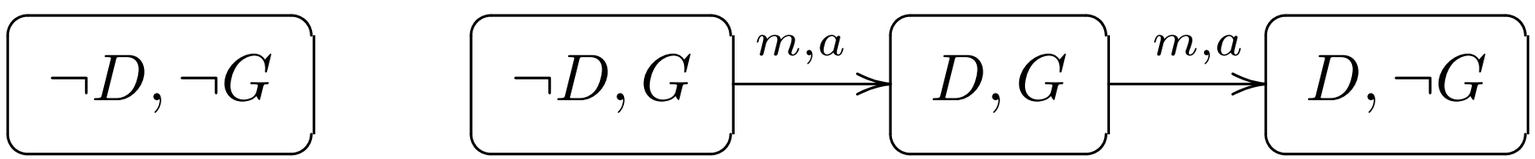
Example : Albert and Mary together

Given:



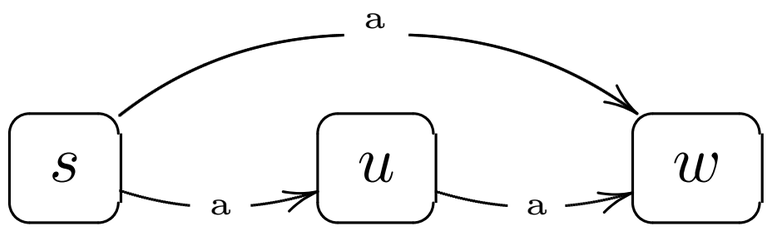
the **protocol to realize the Priority Merge** $R_{m \otimes a}$ consists of: *Mary's sincere* announcement (that she *defeasibly knows* D); then *Albert's infallible* announcement (of *his irrevocable knowledge* of $D \vee G$); followed by *Albert's sincere* announcement (that, *after Mary's announcement, he strongly believes* $\neg G$):

$$\mathbf{m : \Box_m D} \quad ; \quad \mathbf{a : K_a(D \vee G)} \quad ; \quad \mathbf{a : \Box_a \neg G}$$

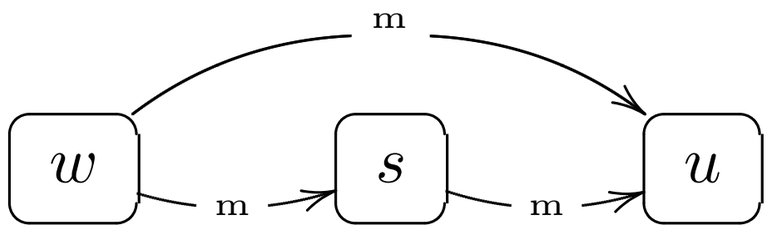


Order-dependence: counterexample

The priority merge of the ordering



with the ordering



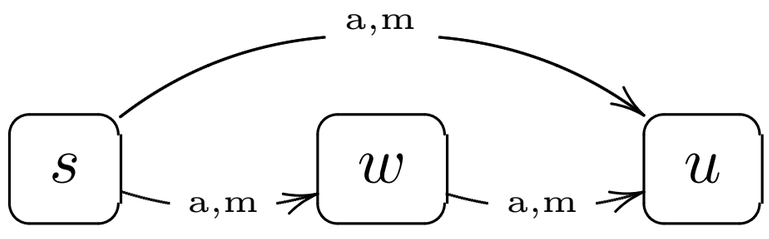
is equal to either of the two orders (depending on which agent has priority). But...

... suppose we have the following public dialogue

$$m : \Box_m u \quad ; \quad a : \Box_a (u \vee w)$$

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

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



The Power of Agendas

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

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

Or the "Speaker of the House", who determines the **order of the speakers** as well as the **the issues** to be discussed and **the relative priority of each issue**.

Different Attitudes

What happens **if we drop the uniformity of attitudes**?

E.g. it is common knowledge that *Mary is infallible* $!_{am}$, but that *Albert is only super-strongly trusted* $! \Uparrow_{ma}$:

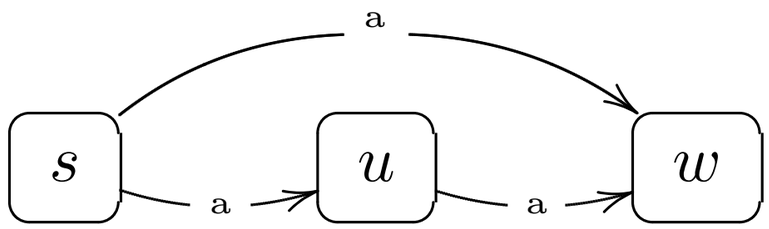
Let us also assume that *we give higher priority to the infallible agent Mary*.

Intuitively, Mary is given **more** persuasive powers, so the merged order should be closer to hers.

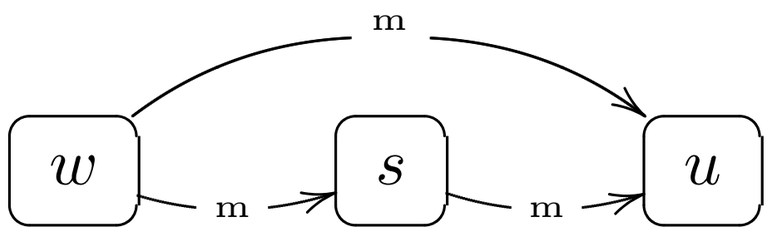
However, **intuition is deceiving**: an infallible agent can honestly announce **less** things than a fallible one!

Counterexample

Given (common knowledge of) $!_{am}\wedge! \Uparrow_{ma}$, the priority merge of Albert's ordering



with Mary's ordering



is *equal to Albert's ordering*, NO MATTER WHAT THE PRIORITY ORDER IS!