

Talking Your Way into Agreement:

“Preference Merge” as “Group Belief
Revision” by Communication

Alexandru Baltag, Oxford University

Based on recent joint work with J. van Benthem and
S. Smets.

Overview

1. **Static Epistemics:** plausibility models, beliefs, revision plans, irrevocable and defeasible knowledge.
2. **Dynamic Belief Revision Operators:** “hard” and “soft” announcements; updates by relativization, and lexicographic upgrades; sincerity, persuasiveness.
3. **Preference Merge and Information Merge:** from Social Choice Theory to Social Epistemology; parallel merge and lexicographic merge.
4. **Realizing Preference Merge Dynamically,** by (public, persuasive, sincere) **communication.**

1. Static Epistemics: Kripke semantics

For any binary accessibility relation $R \subseteq S \times S$ and set $P \subseteq S$, the corresponding *Kripke modality* is:

$$[R]P := \{s \in S : \forall t (sRt \Rightarrow t \in P)\}.$$

When we think of sets $P \subseteq S$ as *propositions* and of elements $s \in S$ as *states*, we write $s \models P$ instead of $s \in P$. Hence, the modalities satisfy:

$$s \models_{\mathbf{S}} [R]P \quad \text{iff} \quad \forall t (sRt \Rightarrow t \models P).$$

Interpretations

If R is interpreted as some kind of epistemic, or doxastic “possibility” relation, then $[R]P$ gives a notion of “knowledge”, or “belief”, of P . In this case, we write KP or BP or $\Box P$, instead of $[R]P$.

If R is interpreted dynamically, as describing a possible “action” or “event”, then $[R]P$ is a *dynamic modality*, describing a kind of “dynamic conditional”: if event R happens then P holds after that.

Knowledge, Belief and Plausibility

The natural *language* to talk about knowledge, belief, conditional belief etc. is *modal logic*. All the operators in this talk (the various “knowledge” operators, belief, conditional belief, the dynamic operators etc.) are special types of “necessity” modalities.

The usual semantics for modal logic is *relational*, given in terms of *Kripke models*. All our formal models for “static” information are Kripke models.

Language: (Multi-)Modal Logic

Modal logic is obtained by adding to the usual propositional logic a *necessity operator*, usually denoted by \Box . This is just a Kripke modality $[R]$ for some given underlying “possibility” relation R . In *multi-modal* logic more than one such operator is considered, and in this case the modalities are distinguished by labels, writing e.g. $\Box_a\varphi$, $\Box_b\varphi$ etc. (or $[a]\varphi$, $[b]\varphi$ etc.). The labels come from a fixed set \mathcal{A} , and they can be given various interpretations: “agents”, “actions”, moments in time etc.

Structures: Kripke Models

A *multi-agent Kripke model* is a structure

$$\mathbf{S} = (S, R_a, \|\cdot\|)_{a \in \mathcal{A}}$$

consisting of a set S of “possible worlds” (or possible “states” of the world), a family of binary accessibility relations $R_a \subseteq S \times S$, indexed by “agents” a from a given group \mathcal{A} , and a “valuation” map $\|\cdot\|$ that maps every “atomic sentence” p from a given set Φ of atomic sentences to a set of worlds $\|p\| \subseteq S$. In practice, we use an arrow notation \xrightarrow{a} whenever we want to write that a particular pair is in the relation R_a .

Semantics

Kripke semantics gives an inductive way to define a *satisfaction relation* \models between possible worlds and sentences. Equivalently, this can be stated as defining, for each sentence φ and Kripke model \mathbf{S} , a *truth set* (or interpretation of φ in \mathbf{S}) $\|\varphi\| \subseteq S$, consisting of all possible worlds at which φ is true. The semantics for the atomic sentences is given by the valuation, the semantics for the propositional connectives is given by the usual Tarskian truth-clauses, while the semantics for necessity \Box_a is given by the Kripke modality $[\xrightarrow{a}]$:

$$s \models_{\mathbf{S}} \Box_a P \quad \text{iff} \quad \forall t (s \xrightarrow{a} t \Rightarrow t \models P).$$

Knowledge as Necessary Truth

Epistemic logic, as usually done, is based on Hintikka's idea (1962) of identifying knowledge with a form of "necessary truth", namely *truth in all epistemically possible worlds*. The epistemic possibilities are given by a binary accessibility relation between possible worlds.

Epistemic Models

An *epistemic model* is a multi-agent Kripke model in which all the accessibility relations are *reflexive*:

$$s \xrightarrow{a} s \text{ for all } s \in S, a \in \mathcal{A}$$

Knowledge is simply defined as the “necessity” operator for these models, as above. Most often, we use a K -notation instead of the \Box -notation above, writing e.g. $K_a\varphi$ for “agent a knows that φ ”. Our *reflexivity postulate* on R_a express the *veracity of knowledge*. It is equivalent to requiring the validity of the axiom **(T)**.

Preordered Models and Partition Models

A *preordered-model*, or **S4-model**, is an epistemic model in which all the accessibility relations are *transitive* (i.e. and so they are *preorders*).

A *partition model*, or **S5-model**, is an epistemic model in which all the accessibility relations are *equivalence relations*.

Forms of Introspection

S4-models validate the axioms of the modal system **S4**, and in particular the principle of *Positive Introspection*:

$$K_a P \Rightarrow K_a K_a P.$$

S5-models validate the axioms of the modal system **S5**, which in addition to Positive Introspection includes the principle of *Negative Introspection*:

$$\neg K_a P \Rightarrow K_a \neg K_a P.$$

Belief

A *doxastic model* (or **KD45**-model) is just a multi-agent Kripke model as above, but for which we require different conditions (instead of reflexivity) on the accessibility relations R_a : namely, we ask them to be *transitive*, *Euclidean* and *serial*. Here, “serial” means that every world has a successor:

$$\forall s \forall a \exists t s \xrightarrow{a} t.$$

Formally, belief is defined exactly like knowledge in terms of the accessibility relations, i.e. as *truth in all doxastically possible worlds*.

Full Introspection of Beliefs

We accept both Introspection postulates (4) and (5) for belief. Belief is a notion that is *purely internal* to the agent. To quote Wittgenstein: “One can mistrust one’s own senses, but not one’s own beliefs”. (*Philosophical Investigations*).

But the same argument does not seem to automatically apply to knowledge: since knowledge is an external notion (having to do with “truth” in the real world), one could argue that agents may be wrong about what constitutes knowledge (since they can be wrong about the truth).

Plausibility (Grove) Models

We now interpret the accessibility relation R_a of a multi-modal Kripke model as a “doxastic preference”, a **plausibility relation**, meant to represent “*soft*” *information*: in this reading, $sR_a t$ means that **world t is at least as plausible for agent a as world s** . For this interpretation, it is customary to use the notation $s \leq_a t$ for the plausibility relation R_a (and \geq_a for its converse), and also to denote the associated “knowledge” modality by \Box_a rather than K_a . It is also customary, though not necessary, to assume that \leq_a is a **connected**, or at least a **locally connected** preorder.

Semantics: Plausibility Models

A **finite plausibility frame** is a Kripke structure (S, \leq_a) consisting of a finite set S of “states” (or “possible worlds”), together with “*locally connected*” *preorder relations* $\leq_a \subseteq S \times S$, one for each agent a , called *plausibility relations*.

Preorder: reflexive and transitive.

Locally connected:

$$s \leq_a t \wedge s \leq_a w \Rightarrow t \leq_a w \vee w \leq_a t,$$

$$t \leq_a s \wedge w \leq_a s \Rightarrow t \leq_a w \vee w \leq_a t.$$

Strict version, Epistemic Indistinguishability etc.

We also consider the “*strict*” *plausibility* relation:

$$s <_a t \text{ iff: } s \leq_a t \text{ but } t \not\leq_a s$$

The *comparability* relation \sim_a gives us a notion of *epistemic indistinguishability*:

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

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

$$s \cong_a t \text{ iff: both } s \leq_a t \text{ and } t \leq_a s$$

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

Reading We read $s \leq_a t$ as “state t is at least for agent a as plausible as state s ”.

Belief in Plausibility Models

A player believes P iff P is *true in all the most plausible worlds*:

$$s \models B_a P \text{ iff } \forall t (t \in \text{Min}_{\leq_a} S \Rightarrow t \models P).$$

It is easy to see that, with this relation, plausibility models are doxastic ($KD45$) models: the belief modality is serial and fully introspective.

Forms of “knowledge”

In a plausibility models, there are some important Kripke modalities:

$$K_a P := [\sim_a]P$$

$$\square_a P := [\geq_a]P$$

We call the first “*irrevocable*” *knowledge*, and the second “*indefeasible*” *knowledge* (or “safe belief”).

“Soft” versus “hard” information

A plausibility relation is in general *transitive*, but *not symmetric*, so “indefeasible knowledge” is *not the S5-type*: it is *positively introspective*, but *not necessarily negatively introspective*. One could say that the fully introspective (*S5-type*) knowledge K_a 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_a captures a more realistic notion of “*soft*” *information*.

“Irrevocable knowledge” embodies “hard” information

In a plausibility model, the comparability relation \sim_a is an equivalence relation, that includes the plausibility relation.

So irrevocable knowledge is *S5*-like (truthful and fully introspective) and stronger than the plausibility-based “knowledge” modality \Box_a . Irrevocable knowledge can thus be said to embody “hard information”.

Their relative strength is captured by the entailment:

$$K_a P \implies \Box_a P,$$

Example 1: Prof Winestein

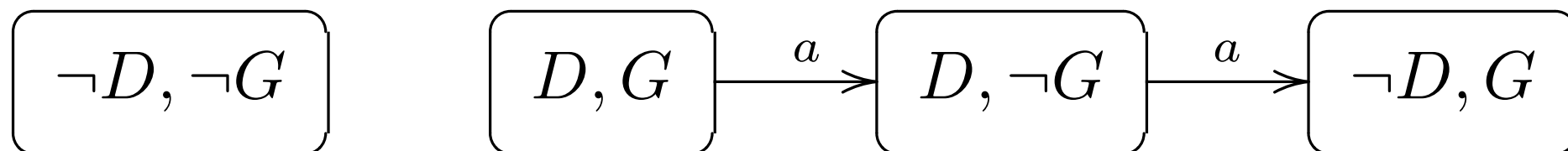
Professor Albert Winestein feels that he is a genius. He **knows** that there are only two possible explanations for this feeling: either he *is* a genius or he's drunk. He doesn't feel drunk, so **he believes that he is a sober genius.**

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

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

A Model for Example 1

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



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

True Belief is not Knowledge

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

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

but **he doesn't "know" he's a genius**, in **any** of the meanings of "knowledge" (irrevocable or indefeasible):

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

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

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

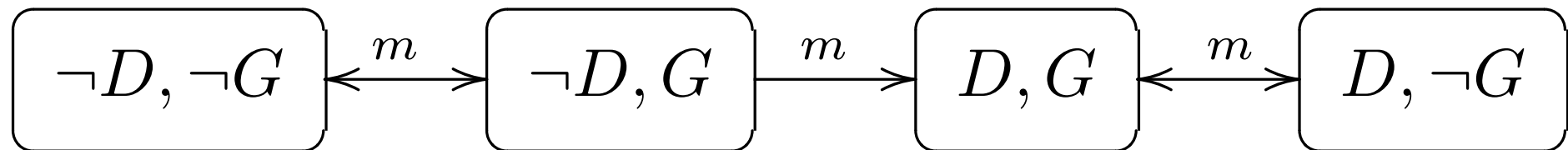
Mary Curry

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

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

She's **completely indifferent with respect to Albert's genius**: being a professor of Creative Cooking, she has no opinion on the matter of Wine Science, so 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.



Marry “knows” though she doesn’t Know

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

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

But *none of these beliefs is irrevocable knowledge*; she doesn’t irrevocably know these things:

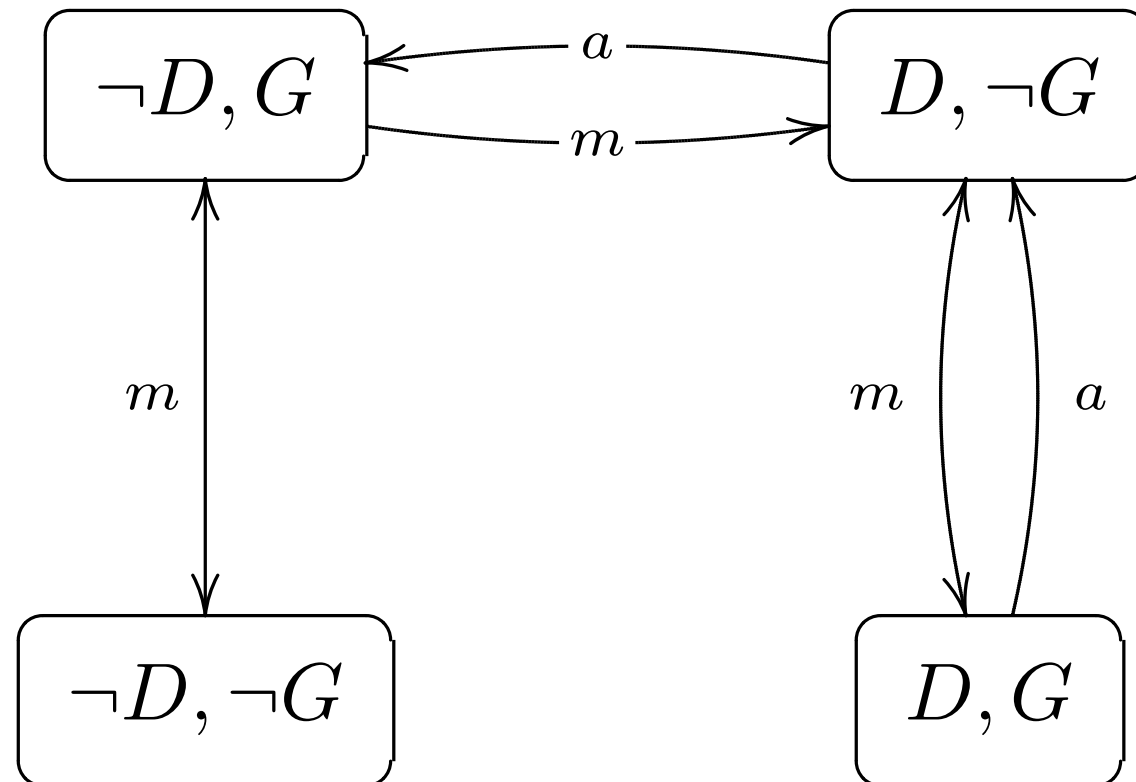
$$(D, G) \models \neg K_m D \wedge \neg K_m G.$$

However, *both her beliefs are “safe”*: so she does “know” them, in the sense of **indefeasible knowledge**

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

A Multi-Agent Model **S**

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



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 “indefeasible knowledge”*:

$$B_a P = \diamond_a \square_a P ,$$

where $\diamond_a P = \neg \square_a \neg P$ is the dual Diamond modality for \square_a (“epistemic possibility”).

The Perfect Believer (Voorbraak's Puzzle)

It seems that there are people who have at least some *false, but "certain" beliefs*: they believe φ so strongly that that they believe they *know* it; but they're wrong: φ is false.

So it is possible to have

$$BK\varphi \wedge \neg\varphi$$

holding at some state.

Puzzle Continued

However, the identity

$$BK\varphi \rightarrow K\varphi$$

can be easily proved from Negative Introspection (for knowledge K) and from the consistency of beliefs with knowledge. Using truthfulness of knowledge, we then get

$$BK\varphi \rightarrow \varphi$$

Solution

The paradox can be solved by noting that it conflates two forms of knowledge.

Irrevocable knowledge is negatively introspective but false sentences cannot be believed to be “known” in this sense. Indeed, *believing that you “irrevocably know” is the same as irrevocably knowing*:

$$BK\varphi = K\varphi$$

Dually, indefeasible knowledge is perfectly compatible with believing that you know: in fact, *all beliefs* are “certain” in this sense, since we have

$$B\Box\varphi = B\varphi$$

But indefeasible knowledge is *not* negatively introspective.

2. Dynamic belief revision operators

From a *semantic* point of view, dynamic belief revision is about “revising” the whole relational structure: **changing the plausibility order** (or the models).

This corresponds to revisions induced by various forms of *communication or observation*, but including the listener’s (observer’s) *belief-revision policy*, her *attitude towards what is announced or observed*, her *dispositions to accept the new information with various degrees of certainty*.

There are many different natural ways in which one can change a plausibility relation, to make the resulting beliefs consistent with some new propositional information φ .

Examples

(1) Update $!P$ with (or *relativization*, to) P : it changes a model by *deleting all the non- P worlds* (or alternatively, deleting all arrows between P -worlds and $\neg P$ -worlds) and *keeping the same plausibility order relations between the remaining worlds*.

(2) Lexicographic upgrade $\uparrow P$: *all P -worlds become “better” (more plausible) than all $\neg P$ -worlds* in the same comparability class, and *within the two zones, the old ordering remains*.

“Hard” and “Soft” Public announcements

The first operation (update) $!P$ corresponds to a *public announcement of a “hard fact”* P : in this case, the announcement comes with an inherent “warranty of truthfulness”, so it is accepted without any reservations.

The second and third operations correspond to various forms of “soft” *public announcements*. In the conservative upgrade, the agents only come to *believe* P , while in the lexicographic upgrade, the agents come to accept P with such a conviction that they consider all P -possibilities more plausible than all non- P ones. But they may still not have “hard” (irrevocable) knowledge of P .

Irrevocable Knowledge is unrevisable

Observe that, for every fact $Q \subseteq S$, we have:

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

This gives a characterization of *irrevocable knowledge* as “*absolute*” belief, invariant under any belief revision: a given belief is “irrevocably known” iff it cannot be revised, i.e. it is believed in any condition.

Knowledge as “stable” belief

Plato: “*permanence*” of belief.

Hintikka: “*robustness*” of belief.

“... by saying “I know that p ”, one makes a commitment stronger than one made by making a simple assertion; one proposes (it is part of one’s proposition) to stick to this statement no matter what further information one expects to receive.” (Hintikka, *Knowledge and Belief*, 1962)

An “absolute” interpretation

If by “further information” we mean *any further evidence* extracted from *any source, however unreliable or deceiving*, then this may include *misinformation*.

“Real knowledge”, in this absolute sense, should be robust *even in the face of false evidence*. This gives us *irrevocable knowledge K*.

(We of course assume here a rational agent, not a fundamentalist: her refusal to revise her belief will then be grounded in an irreproachable justification, not in a blind resistance to belief change!)

“Stability” of belief

The “defeasibility theory” of knowledge (Klein, Lehrer, Stalnaker) takes a more “relative” interpretation: “information” means “*true* information”.

“An agent knows that φ if and only if φ is true, she believes that φ , and she continues to believe φ if any *true* information is received” (Stalnaker 2006).

“A belief α is a piece of knowledge of the subject S iff α is not given up by S on the basis of any *true* information that S might receive” (Rott 2004).

Indefeasible Knowledge

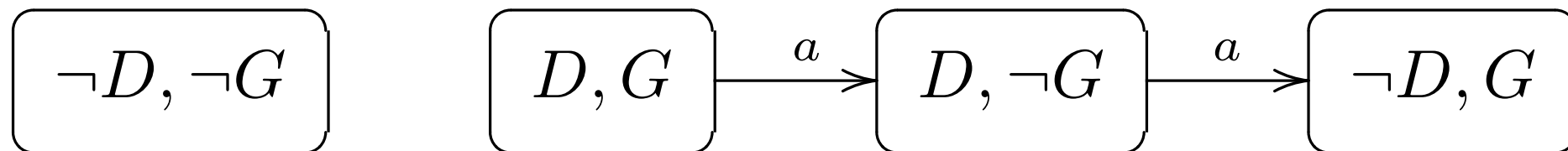
The following equivalence shows that the concept of knowledge described by the defeasibility theory corresponds to our “indefeasible knowledge” \square : for every fact $Q \subseteq S$, we have

$$s \models \square_a Q \quad \text{iff} \quad s \models [!P]B_a Q \quad \text{for all } P \subseteq S.$$

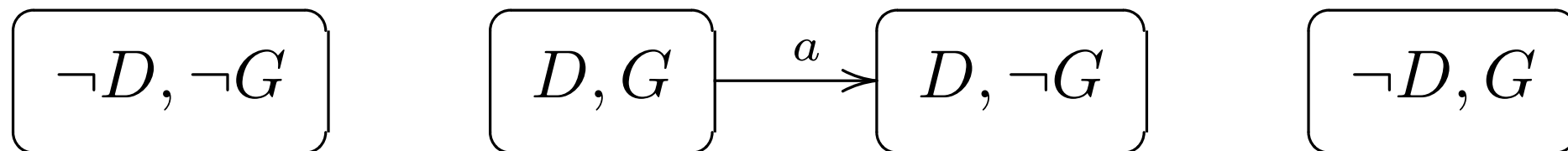
(Observe that a truthful announcement $!P$ can only take place at s iff $s \models P$.)

Example 2: “Showing” Hard Evidence

Suppose we are in the original situation:

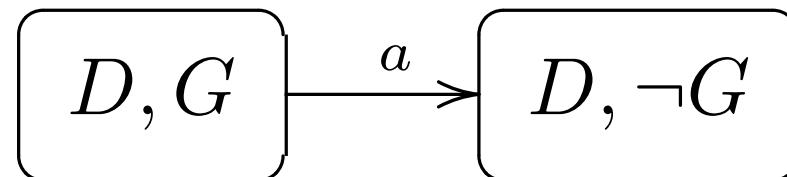


Next, suppose Albert **is shown** the result of a blood test, proving beyond reasonable doubt **that he is drunk**. We take this result to be accepted as “**hard**”, **incontrovertible evidence**. This corresponds to performing an **update** $!D$, resulting in:



Losing Your True Belief

In fact, only the worlds that are connected to the real world (D, G) are relevant, so we can delete the others:



In this way, it is obvious that, after the update is performed on the real world (D, G) , Albert **starts to wrongly believe** that he is **not** a genius!

So (truthful) learning can be **dangerous**: sometimes is better not to learn too much!

The Dangers of Learning

We saw that, if Winestein learnt that he was drunk, he would lose his (true!) belief that he's a genius!

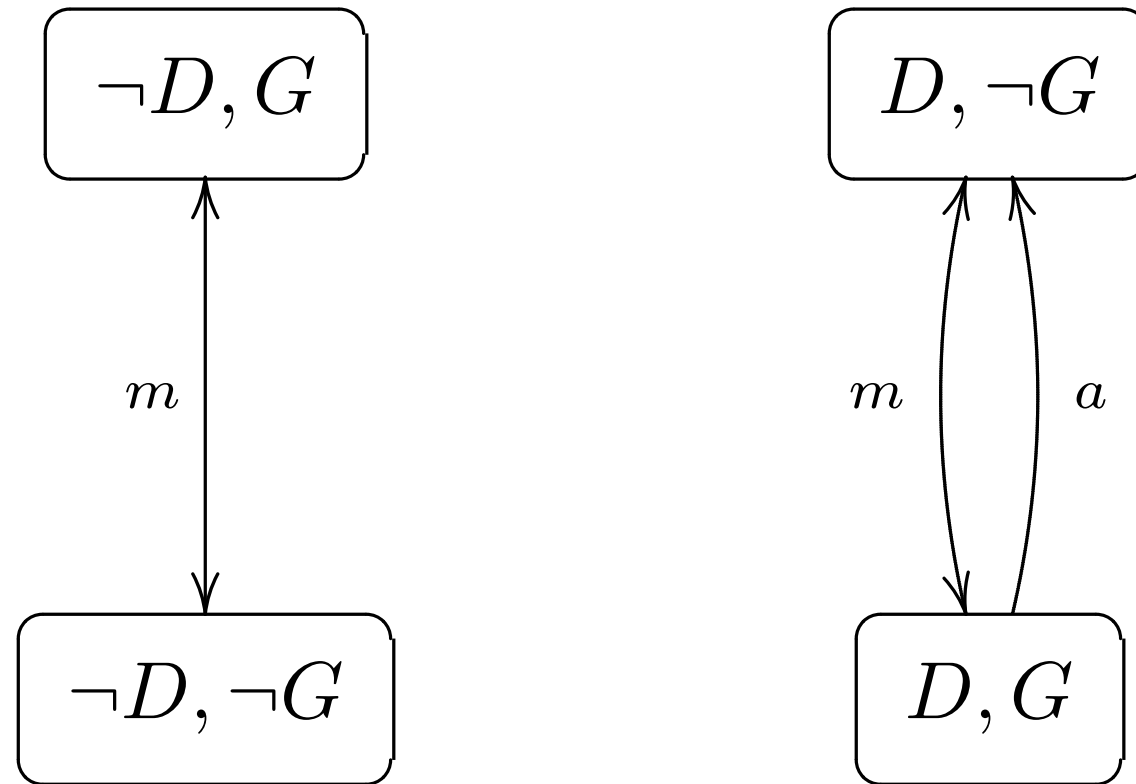
This is an example of a **true, but “un-safe” belief**: it can be lost after acquiring (new) true information.

In Lehrer's terms: **Albert's true belief in his own genius is not (indefeasible) “knowledge”**.

Hard Public Announcement

Let us perform the same update $!D$ on the whole multi-agent model representing the original situation: this corresponds to (a **trusted, infallible** source) **publicly announcing** the result of the test, giving “**hard**” **evidence that Albert is drunk.**

The updated model becomes:



where worlds on the left can be erased, since they are not connected to the actual world (D, G) .

Example 3: “Soft” Public Announcement

Instead of an indisputable drunkness test, let simply Marry announce publicly (to Albert): “*Man, you are drunk!*”. We assume Marry’s announcement is **sincere** and **persuasive**: she tells what she thinks and she convinces Albert.

Since Marry is a fallible human being (and not an infallible source), this announcement is **soft**: in principle, she could be wrong, or she could lie, or she could simply guess and be right only by chance. Albert should also be aware of this fallibility.

Indeed, in the original situation Marry **doesn't know** that Albert is drunk: well, **not in the sense of irrevocable knowledge (K)**.

But... *she does "know" it* in the sense of **indefeasible knowledge**: she correctly believes it, and this belief is safe.

When Can An Agent Make a Sincere Announcement

A general principle is that a *sincere public announcement* by an agent m should **not change the plausibility order of the same agent m** : the announcement represents m 's beliefs or information, so it was supposed to be already “known” by m in a sense, and hence announcing it should not affect m 's beliefs or “knowledge”.

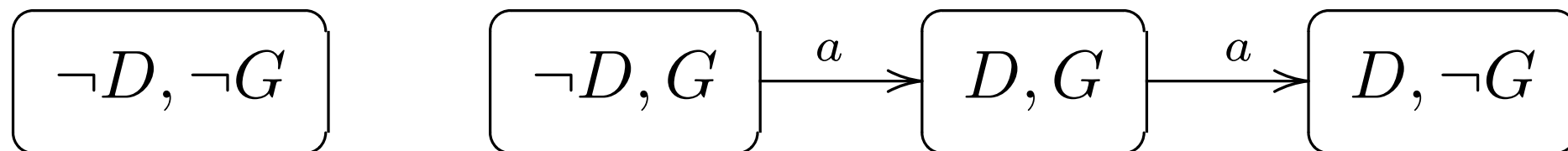
Highly Persuasive Announcements are Lexicographic

So we **cannot** interpret the above announcement as a “hard” update $!D$, since such an update would automatically change Marry’s order (making her irrevocably know D , when she didn’t know it before!).

But, if the announcement is highly persuasive, we **can** model as a “soft” lexicographic announcement $\uparrow D$; i.e. after hearing it, all agents upgrade lexicographically with D : they *start to prefer any D -world to any $\neg D$ -world.*

A Lexicographic Upgrade

The upgrade $\uparrow D$ changes Albert's order to



while *Marry's order is left unchanged!* If we take this invariance as a *commonly accepted feature* of an announcement being *made by Marry*, than Albert implicitly learns **more** than D : he learns that announcing D leaves invariant Marry's order! This means he learns $\square_m D$, i.e. that B was **indefeasibly known by Marry**. So we can think of this as an upgrade of the form $\uparrow \square_m D$.

Sincerity

For the announcement to be “truly sincere”, i.e. non-deceiving, we need to require that this implicit information is “correct” in some sense, i.e. that indeed **Marry believed that she “knew” (indefeasibly) that D .**

But, as we saw, this is the same as simple belief:

$$B_m \square_m D = B_m D.$$

So “sincerity” requires that, before making the announcement, Marry believed that D .

Sincere, Persuasive Soft Announcements

So we came to the conclusion that a **sincere persuasive public announcement** by a fallible agent x has the form

$\uparrow \square_x P$, for some P such that $B_x P$.

3. Preference Merge and Information Merge

In Social Choice Theory, the main issue is how to *merge* the agent's individual preferences in a reasonable way. In the case of *two agents*, a merge operation is a function \odot , taking preference relations R_a, R_b into a “*group preference*” relation $R_a \odot R_b$ (on the same state space).

As usually considered, the problem is to find a “*natural*” *merge operation* (subject to various *fairness conditions*), for merging the agents' preference relations. Depending on the stringency of the required conditions, one can obtain either an Impossibility Theorem or a classification of the possible types of merge operations.

Belief Merge and Information Merge

If we want to *merge the agents' beliefs* B_a, B_b , so that we get a notion of “group belief”, then it is enough to merge the belief relations $\rightarrow_a, \rightarrow_b$.

If we want to merge the agents' *hard information* K_a, K_b , then it is enough to merge the epistemic indistinguishability relations \sim_a, \sim_b .

If we want to merge the agents' *soft information* \square_a, \square_b (or, equivalently, to merge all their *conditional beliefs*), then we have to merge the plausibility relations \leq_a, \leq_b .

Merge by Intersection

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

$$\bigcap_a R_a.$$

In the case of two agents, it takes:

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

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

Distributed Knowledge is Parallel Merge

This form of merge is particularly suited for “hard information” (irrevocable knowledge) K : since this is an absolutely certain, fully reliable, unrevisable and fully introspective form of knowledge, there is no danger of inconsistency. The agents can pool their information in a *completely symmetric manner, accepting the other’s bits without reservations.*

The concept of “distributed knowledge” DK in epistemic logic corresponds to the parallel merge of the agents’ hard information:

$$DK_{a,b}P = [R_a \cap R_b]P.$$

“Dynamic” intuition: pooling information

Another characterization is:

$$s \models DK_{a,b}P$$

iff

$$\exists P_a, P_b \text{ such that } s \models K_a P_a \wedge K_b P_b \text{ and } P_a \cap P_b \subseteq P.$$

The intuition underlying this concept is *dynamic*: distributed knowledge captures the *potential knowledge* obtainable by the group via inter-agent communication: what the agents *could know if they would share all their information*.

But to make this intuition precise, we would need to be able to model “sharing information” dynamically: this is exactly what dynamic-epistemic logic will allow us to do!

Lexicographic Merge

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

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

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

Lexicographic merge of soft information

This form of merge is *particularly suited for “soft information”*, given by either indefeasible knowledge \square or belief B , *in the absence of any hard information*: since soft information is not fully reliable (because of lack of negative introspection for \square , and of potential falsehood for B), some “screening” must be applied (and so some hierarchy must be enforced) to ensure consistency of merge.

$$s \models \Box_{a/b} P$$

iff

$$\exists P_a, P_b \text{ s. t. } s \models \Box_a P_a \wedge \Box_b P_b \wedge \Box_a^{weak} P_b \text{ and } P_a \cap P_b \subseteq P.$$

In other words, all a 's “indefeasible knowledge” is unconditionally accepted by the group, while b 's indefeasible knowledge is “screened” by a using her “weakly indefeasible knowledge”.

Relative Priority Merge

Note that, *in lexicographic merge, the first agent's priority is "absolute"* in the sense that her strong preferences are adopted by the group even when they are incomparable according to the second agent. 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). This leads us to a "more democratic" form of merge: the *(relative) priority merge* $R_{a \otimes b}$, given by $R_{a \otimes b} := (R_a \cap R_b^\sim) \cup (R_a^\sim \cap R_b)$
 $= (R_a^> \cap R_b^{-1}) \cup (R_a \cap R_b) = R_a \cap R_b^\sim \cap (R_a^> \cup R_b)$.

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

Relative Priority Merge can be thought of as a **combination of Merge by Intersection and Lexicographic Merge**: the “hard” information is merged by intersection; then the “soft” information is lexicographically merged; but with the proviso that it still has to be consistent with the group's hard information.

Priority Merge of Soft Information

The corresponding notion of “indefeasible knowledge” of the group is obtained as in the lexicographic merge, *except that both agents’ “irrevocable knowledge” is unconditionally accepted.* Formally:

$$s \models \Box_{a \otimes b} P$$

iff

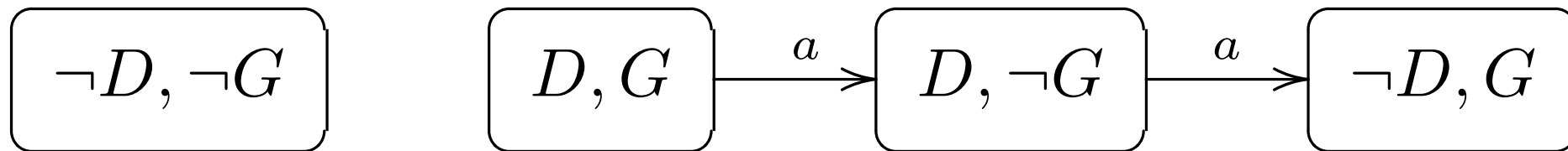
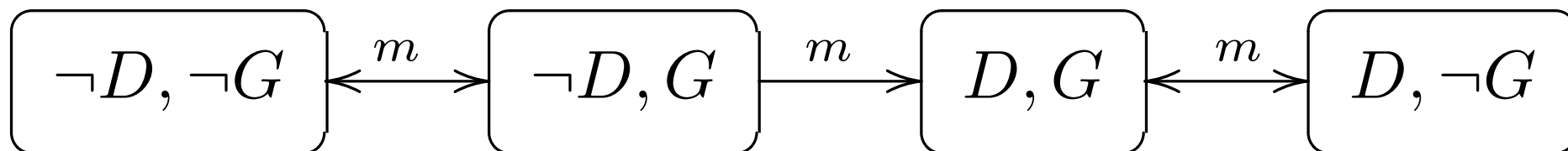
$$\exists P_a, P_b, \varphi'_b \text{ s. t. } s \models \Box_a P_a \wedge K_b P_b \wedge \Box_b P'_b \wedge \Box_a^{weak} P'_b$$

$$\text{and } P_a \cap P_b \cap P'_b \subseteq P.$$

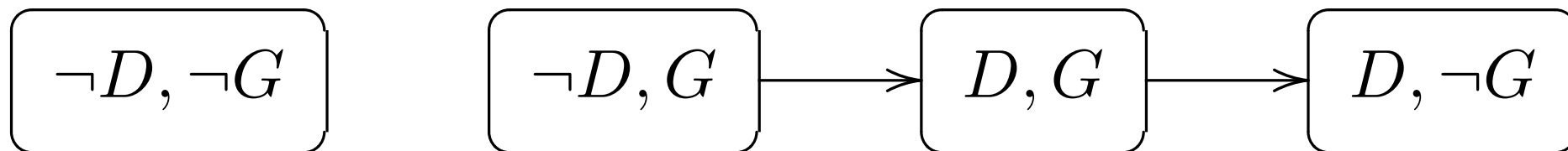
In other words, relative-priority group “knowledge” is obtained by pooling together the following: agent a ’s “indefeasible knowledge”; agent b ’s “irrevocable knowledge”; and the result of screening agent b ’s “indefeasible knowledge” using agent a ’s “weakly indefeasible knowledge”.

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:



Merging Beliefs is Not a Sure Way to the Truth

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

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

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

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

4. “Realizing” Preference Merge Dynamically

Intuitively, the purpose of “preference merge” $R_a \odot R_b$ is to achieve a state in which the two agents’ preference relations are *fully merged* accordingly, i.e. to perform an epistemic action (or sequence of actions) σ transforming the initial model (S, R_a, R_b) to a model (S, R'_a, R'_b) such that

$$R'_a = R'_b = R_a \odot R_b$$

Let us call this “*full realization*” of the merge operation \odot .

Weak Realizations

A weaker form of “realization” is when only one agent (say, b) realizes the merged relation, i.e. we arrive at a situation in which

$$R'_b = R_a \odot R_b$$

(but R'_a may differ).

Let us call this “single-agent realization” of the merge operation \odot .

Merging by Public Communication

For each of the above types of public communication $(!, \uparrow)$, we can ask which merge operations are realizable (in either sense) by a sequence of announcements of that type.

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

Realizing Distributed Knowledge

In the case of *distributed knowledge*, it is easy to design an algorithm to realize this merge operation by a *sequence of truthful public announcements*: in no particular order, the agents have to publicly announce “all that they know” (in the sense of *irrevocable knowledge* K).

More precisely, for each set of states $P \subseteq S$ such that P is known to a given agent a , a public announcement $!(K_a P)$ is made.

The Algorithm

Formally, *the algorithm for single-agent realization* (by agent b) of distributed knowledge requires the other agent (a) to perform the following sequence of announcements:

$$\sigma_a := \prod \{!(K_a P) : P \subseteq S \text{ such that } s \models K_a P\}$$

(where \prod is sequential composition of a sequence of actions).

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

$$R'_b = R_a \cap R_b,$$

and so agent b 's knowledge after this sequence of announcements coincides with distributed knowledge:

$$s \models Dk_{a,b}P \leftrightarrow [\sigma_a]K_bP.$$

Full Realization

The algorithm for *full realization* requires agent b to “answer” by publicly announcing all that *he* knows *after* the previous algorithm has been performed. Formally, this “answer” is the following sequence of announcements:

$$\sigma_b := \prod \{!(K_b P) : P \subseteq S \text{ such that } s \models [\sigma_a]K_b P\}.$$

So the algorithm for full realization is the sequential composition $\sigma := \sigma_a \cdot \sigma_b$. By this algorithm, *distributed knowledge is converted into common knowledge*:

$$s \models DK_{a,b}P \leftrightarrow [\sigma]Ck_{a,b}P.$$

Order-independence

Moreover, the *order* in which the agents make the announcements doesn't actually matter.

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

Realizing Priority Merge

We can realize the priority merge $\square_{a \otimes b}$ of soft information by *lexicographic public updates*, by an algorithm very similar to the one for distributed knowledge.

Essentially, the agents are asked to *publicly announce (via lexicographic upgrades) that they “know” all that they believe they “know”*. Here, “knowledge” means now *indefeasible knowledge* \square_a .

Order-dependence and lack of introspection

The main two differences are that:

- (1) The order matters. The agent that has “priority” in the merge has to be the first to announce all he “knows”.
- (2) Since “knowledge” means now indefeasible knowledge, which is not negatively introspective, the agents don’t know for sure what things they “know” and what not, and the best they can do is to announce all the things they *believe they know*.

Be Persuasive!

But, since believing to (indefeasibly) “know” is the same as believing, they have to **announce that they “know” P , for each proposition P which they believe.**

Note that simply announcing that they believe it, or that they believe they know it, won't do: this will not in general be enough to achieve belief merge. Being informed of another's beliefs is not enough to convince you of their truth. What is needed for belief merge is that the agents try to “convert” the other to their own beliefs by *claiming they “know” what they only believe they “know”*.

Needed: sincere persuasive soft announcements

So we conclude that what we need is upgrades of the form

$$\Box_a P, \text{ for any } P \text{ such that } B_a P,$$

i.e. exactly the kind of upgrades that we earlier used to describe **sincere persuasive public announcements** (of soft knowledge) by a fallible agent.

The Algorithm for Weak Realization

Formally, if a is the agent who has “priority”, then *the algorithm for single-agent realization (by agent b) of priority merge of soft information* requires a to perform the following sequence of soft public announcements:

$$\rho_a := \prod \{ \uparrow (\Box_a P) : P \subseteq S \text{ such that } s \models B_a P \}.$$

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

$$R'_b = R_a \otimes R_b,$$

and so agent b 's indefeasible knowledge after this sequence of announcements coincides with merged knowledge: for all $P \subseteq S$, we have

$$s \models \Box_{a \otimes b} P \leftrightarrow [\rho_a] \Box_b P.$$

Full Realization

As in the case of distributed knowledge, the algorithm for *full realization of priority merge* requires agent b to “answer” by publicly announcing (via lexicographic upgrades) that he “knows” all that *he* believes to “know” *after* the previous algorithm has been performed. Formally, this “answer” is the following sequence of announcements:

$$\rho_b := \prod \{ \uparrow (\Box_b P) : P \subseteq S \text{ such that } s \models [\rho_a] B_b P \}.$$

So the algorithm for full realization is the sequential composition $\rho := \rho_a \cdot \rho_b$. Indeed, it is easy to see that, by this algorithm, *the priority merge of (indefeasible) “knowledges” is converted into common (indefeasible) “knowledge”*: for all $P \subseteq S$, we have

$$s \models \Box_{a \otimes b} P \leftrightarrow [\rho]C \Box_{a,b} P.$$

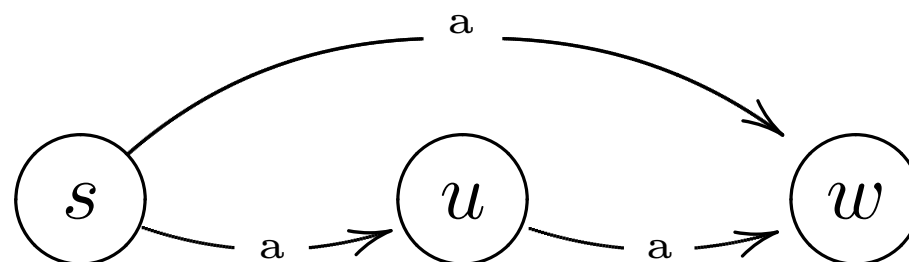
The Rules of the Game

The “rules of the game” in the above algorithm are:

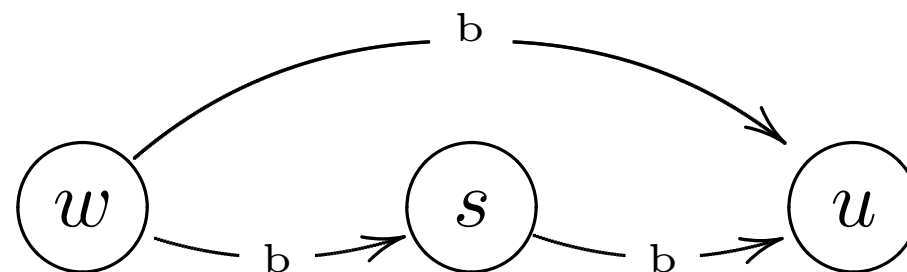
- (1) “*sincerity*”: agents announce that they “know” *only things that they believe they “know”*;
- (2) “*exhaustiveness*”: the algorithm stops only when the agents have announced ‘*all*’ *they (think they) “know”*;
- (3) “*priority order*”, strictly enforced: the agents with higher priority have to *finish announcing all they (think they) “know”* before agents with lower priority can speak.

Order-dependence: counterexample

The priority merge of the ordering



with the ordering

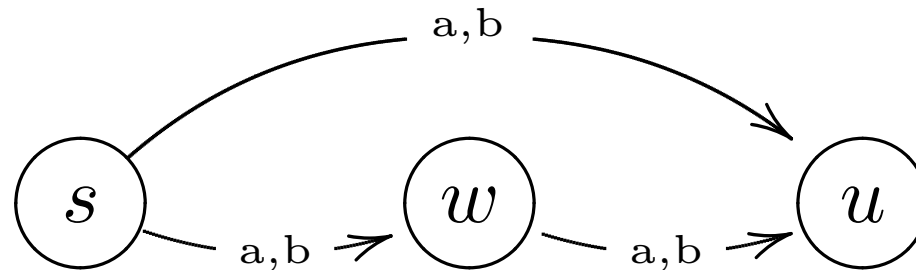


is equal to either of the two orders (depending on which agent has priority). But...

... suppose we have the following public dialogue

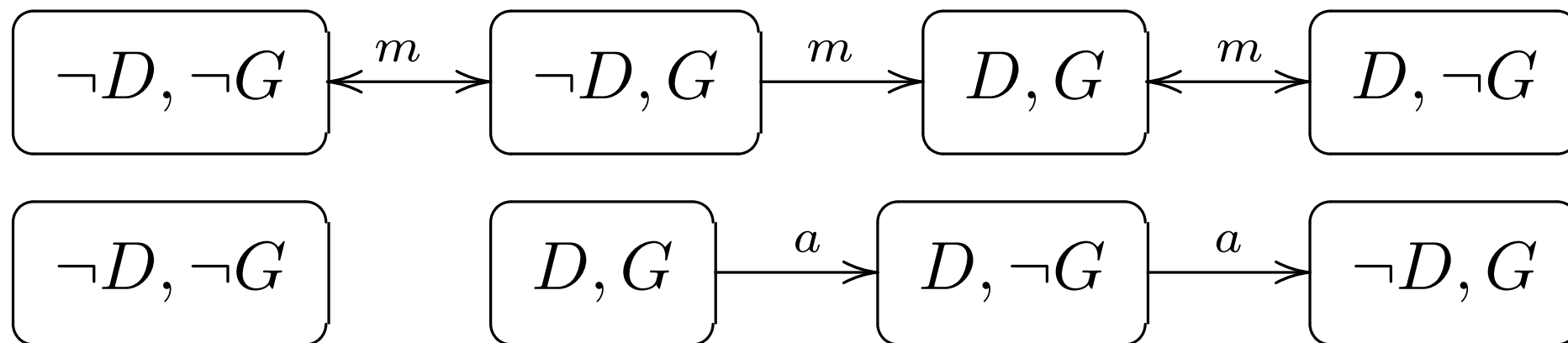
$$\uparrow \Box_b u \cdot \uparrow \Box_a (u \vee w)$$

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



Example

Recall the initial Marry & Albert orders:



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

$$\uparrow K_a(D \vee G); \uparrow \square_m D; \uparrow \square_a \neg G$$

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

$$K_a P = \square_a K_a P$$

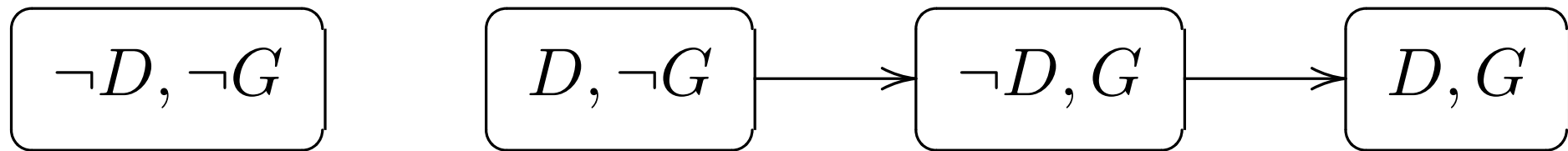
Expertise-guided Priority

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

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

The Way to the Truth

This results in the merged order:

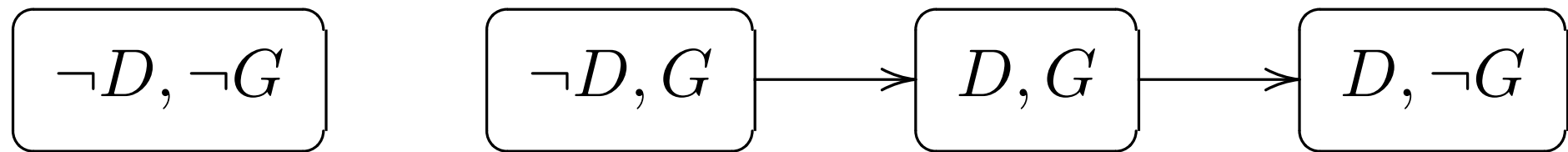


So now *the resulting joint beliefs are all correct!*

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

But... the Order still Matters!

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



leading to the incorrect belief in non-genius!

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

The Power of Agendas

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

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

Open Problem

So, depending on the “agenda”, soft announcements can realize **a whole plethora of merge operations**.

Nevertheless, **NOT everything goes**: the requirements imposed on the plausibility relations generally pose restrictions to which kinds of merge are realizable. E.g. it is easy to see that *neither intersection nor lexicographic merge preserve the “local connectedness” of plausibility relations, and so none of them is realizable in our (locally connected) setting.*

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