Dynamic Logics for Communication

A study of information-updating and belief-changing actions

Alexandru Baltag

(COMLAB, Oxford University)

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

The Problem: Doxastic Changes and Conditionals

Problem: *Understand, explain and compute the dynamics of knowledge and belief* in a *multi-agent* context.

A More general Problem: study and compare notions of *conditional*, such as *doxastic conditionals* $B_a^P Q$, *dynamic conditionals* $[P?]Q$, $[P!]Q$, *counterfactual conditionals* etc.

...In fact, we started from *quantum conditionals*!

Conditional Beliefs and Changes of Belief

In particular, consider *hypothetical* (i.e. *conditional*) *beliefs*, captured by a binary doxastic operator

$$B_a^P Q,$$

as well as actual *changes of belief*, expressed by combining dynamic and doxastic modalities:

$$[P!]B_a Q.$$

"Static" and "Dynamic"

The conditional $B_a^P Q$ is "*static*" w.r.t. the actual state of the system: the state doesn't change, only a's theory about it changes. It is directly related to the standard AGM belief revision theory.

$[P!]B_a Q$ is "*dynamic*": the state changes (the agent learns P), and after that a beliefs P.

In fact, we *can* understand conditional beliefs *dynamically*: $B_a^P Q$ means that, after learning P, agent believes that Q *was* true (in the original state) *before* the learning.

Moore Sentences

To see the difference, consider a Moore sentence

$$P \;:=\; Q \wedge \neg B_a Q.$$

It is rather clear that

$$B_a^P P$$

should always be *true*, while

$$[P!]B_a P$$

will typically be *false* (whenever P is true, i.e. whenever the action $P!$ is a true "learning" action).

Doxastic Logic

The logic of *belief*. Usually accepted axioms: $KD45$.

$$B_a(P \to Q) \to (B_aP \to B_aQ)$$

$$B_a\neg P \to \neg B_aP$$

$$B_aP \to B_aB_aP$$

$$\neg B_aP \to B_a\neg B_aP$$

Doxastic-Epistemic Logic

Adding *knowledge*: satisfying $S5$; knowledge implies belief; agents know their beliefs and non-beliefs.

$$K_a(P \to Q) \to (K_a P \to K_a Q)$$

$$K_a P \to P \wedge B_a P$$

$$K_a P \to K_a K_a P$$

$$\neg K_a P \to K_a \neg K_a P$$

$$B_a P \to K_a B_a P$$

$$\neg B_a P \to K_a \neg B_a P$$

Relational semantics: doxastic models

A *doxastic frame* is a Kripke structure of the form $(S, \rightarrow_a)_a$, such that all the relations $\rightarrow_a \subseteq S \times S$ are transitive, Euclidean and serial:

$$s \rightarrow_a t \text{ and } t \rightarrow_a w \text{ implies } s \rightarrow_a w,$$

$$s \rightarrow_a t \text{ and } s \rightarrow_a w \text{ implies } t \rightarrow_a w,$$

for all $s \in S$ there exists some $t \in S$ such that $s \rightarrow_a t$.

A *doxastic model* is a frame together with a valuation $\| \bullet \| : \Phi \rightarrow \mathcal{P}(\mathcal{S})$, assigning sets of states to "atomic sentences" from a given list $\Phi = \{p, q, \ldots\}$.

Epistemic equivalence in Doxastic models

Two states s, t are *indistinguishable (or epistemically equivalent* for agent a if we have:

$$\forall w \in S \, (\, s \rightarrow_a w \Leftrightarrow t \rightarrow_a w \,)$$

We write $s \sim_a t$ in this case.

Knowledge and Belief in Doxastic Models

Given a doxastic frame S, an *S-proposition* is any subset $P \subseteq S$. Knowledge and belief can be defined as modal operators on S-propositions using the standard Kripke semantics:

$$B_a P = \{s \in S : \forall t \in S : \text{ if } s \to_a t \text{ then } t \in P\}$$

$$K_a P = \{s \in S : \forall t \in S : \text{ if } s \sim_a t \text{ then } t \in P\}$$

Discrete Probabilistic Measures

A *discrete probabilistic space* is a pair (S, μ), where S is a *finite* set of states and $\mu : \mathcal{P}(S) \to [0, 1]$ satisfies the standard axioms of a probability measure.

This is equivalent to having simply a *probability assignment* (on S finite), i.e. a map $\mu : S \to [0, 1]$ such that $\sum_{s \in S} \mu(s) = 1$. This can be uniquely extended to $\mathcal{P}(S)$ by putting

$$\mu(P) := \sum_{s \in P} \mu(s)$$

Note that, to uniquely determine such a probability assignment on a set $S = \{s_1, \ldots, s_n\}$ with n elements, it is enough to specify $n-1$ probabilities: $\{\mu(s_i) : 1 \leq i \leq n-1\}$.

Conversely, any map $\mu : \{s_i : 1 \leq i \leq n\} \to [0,1]$ on $n-1$ states, such that $\sum_{1 \leq i \leq n-1} \mu(s_i) \leq 1$, determines a discrete probabilistic space.

Subjective Probability: "Degrees of Belief"

The *Bayesian*, or "subjective", interpretation of probability: $\mu(P) = \alpha$ means that the agent's "degree of belief" in P, the "intensity" of his belief in P, is given by the number α. "Certainty" corresponds to $\alpha = 1$.

But what about (simple) *belief*? As in $B_a P$.

A "big enough" probability is not enough!

One might be tempted to equate simple "belief" with "a high degree of belief", by putting e.g.

$$BP \text{ iff } \mu(P) \geq \alpha$$

for some big enough α, say $\alpha = 0.99$. Or even $\alpha = 0.5$.

But none of these will make $KD45$ axioms sound. In fact, even K will fail! The K-validity

$$BP \wedge BQ \Rightarrow B\,(P \wedge Q)$$

fails, except if $\alpha = 0$ (i.e. the agent has no non-trivial beliefs) or $\alpha = 1$.

Simple Belief=Certain Belief

So the only natural (and non-trivial) probabilistic interpretation for belief is to take $\alpha = 1$, i.e. to equate (simple) belief with "certain" belief:

$$BP \text{ iff } \mu(P) = 1$$

This is independent on the current state, so it only applies to the case in which the agent has *no knowledge* at all about the current state.

Incorporating Knowledge

If the agent *does* have some information Q about the current state s, i.e. if all he knows is that $s \in Q$, then "certain" belief becomes a conditional probability:

$$s \in BP \text{ iff } \mu(P|Q) = 1\,,$$

where $\mu(P|Q)$ is defined as usual:

$$\mu(P|Q) := \frac{\mu(P \wedge Q)}{\mu(Q)}\,,$$

in the assumption that $\mu(Q) \neq 0$.

Probabilistic Models

So a *(discrete) probabilistic frame* for doxastic-epistemic logic is a structure $(S, \mu_a, \Pi_a)_a$, such that:

- each (S, μ_a) is a (discrete) probabilistic space, and
- each (S, Π_a) is an *Aumann structure*; i.e. each Π_a is an information *partition* of S. (Equivalently, we can use equivalence relations $\sim_a$ instead of partitions Π_a.) For a state s, denote by $s(a)$ *the information cell* of s in the partition Π_a (or the $\sim_a$-equivalence class of s).
- $\mu_a(\, s(a)\,) \neq 0$.

Exercise

Any discrete probabilistic frame is a doxastic frame, if we take

$$s \to_a t \text{ iff } \mu_a(t|s(a)) \neq 0.$$

Moreover, in this doxastic frame, we have:

$$s \sim_a t \text{ iff } s \in t(a) \text{ iff } s(a) = t(a).$$

Conversely, any finite doxastic frame S can be "probabilized", i.e. we can define measures μ_a and thus a discrete probabilistic frame, such that the doxastic frame associated to it (by the above correspondence) is the original doxastic frame S.

Completeness of doxastic-epistemic logic

"Model"= frame together with a valuation $\|\cdot\|$.

The semantics for belief and knowledge is the obvious one:

$$B_a P \ := \ \{ s \in S : \mu_a(P|s(a)) = 1 \}$$

$$K_a P \ := \ \{ s \in S : \ s(a) \ \subseteq P \}$$

Exercise: The above correspondence between finite doxastic frames and discrete probabilistic frames preserves belief and knowledge: the same propositions are believed/known at the corresponding states. The above axioms of doxastic-epistemic logic are *complete* for both discrete probabilistic models, and finite doxastic models.

Knowledge $\neq$ True Certain Belief

Example: $S = \{H, T\}$, $H(a) = \{H\}$, $T(a) = \{T\}$, $H(b) = T(b) = \{H, T\}$;
μ_a is irrelevant; $\mu_b(T) = 1$, and so $\mu_b(H) = 0$.

A coin is on the table, in front of Alice, who sees the upper face of the coin. Bob can't see it, but he believes that it is Tails. Suppose in fact this is true: the real state is T, so Bob's certain belief is true. But he still doesn't know it's Tails!

Common Belief and Common Knowledge

$$CKP = \bigwedge_{a_1,a_2,\ldots,a_n} K_{a_1} B_{a_2} \ldots K_{a_n} P,$$

$$CBP = \bigwedge_{a_1,a_2,\ldots,a_n} B_{a_1} B_{a_2} \ldots B_{a_n} P.$$

Here, interpret $\bigwedge$ as "(infinite) intersection". We cannot define it in the language of doxastic logic, since the language doesn't allow infinite conjunctions.

Learning New Information

Suppose we make a *truthful public announcement*. $P!$ is the action in which a true proposition P is publicly learned by the whole group. For instance, $T!$ is the public announcement that the coin lies Tails up.

"Public Announcement is Conditionalization"

The received wisdom is that *learning new information corresponds to probabilistic conditionalization*: the epistemic-doxastic model after the action $P!$ is obtained by deleting all the non-P states, intersecting the information cell with the new information P, and conditionalizing the probabilistic beliefs with P:

$$S' := P,$$

$$s(a)' := s(a) \cap P,$$

$$\mu'_a(Q) := \mu_a(Q|P).$$

Example continued

This works pretty well *if the agents didn't happen to have believe* $\neg P$.

For instance, after the public announcement $T!$, the model S from the above example becomes: $S' = \{T\}$, $T(a)' = T(b)' = \{T\}$, $\mu'_a(T) = \mu'_b(T) = 1$. *Now*, Bob knows it's Tails up.

The Problem of Belief Revision

But what if the real state in the original model was in fact H, i.e. (against Bob's belief) the coin was lying Heads up?

Then a truthful public announcement must say this, i.e. it must be H!, and so the new model is $S' = \{H\}$. Given Bob's prior belief (that the coin lies Tails up), we get

$$\mu'_b(H) = \mu_b(H|H) = \frac{\mu_b(H)}{\mu_b(H)},$$

which is *undefined*, since $\mu_b(H) = 0$.

Probability theory cannot do belief revision!

In probabilistic applications, e.g. in Game Theory, this problem is sometimes preempted by requiring that $\mu_a(s) \neq 0$ for all states s. But this, in effect, is a way of eluding the problem by simply stipulating that *agents never have any wrong beliefs*!

In fact, this *collapses belief into knowledge*: with our previous definitions for K_a and B_a, the two become equivalent!

Bayesian belief update, based on standard Probability Theory, simply cannot deal with any non-trivial belief revision.

(Discrete) Conditional Probability Spaces

Studied by Popper, Renyi, de Finetti, van Fraasen.
Applied to belief revision by Halpern and others.

A *discrete conditional probability space* is a pair (S, μ), where S is a *finite* set of states and
$\mu : \mathcal{P}(S) \times \mathcal{P}(S) \to [0, 1]$ satisfies the following axioms:

1. $\mu(A|A) = 1$,
2. $\mu(A \cup B|C) = min(1, \mu(A|C) + \mu(B|C))$, if $A \cap B = \emptyset$,
3. $\mu(A \cap B|C) = \mu(A|B \cap C) \cdot \mu(B|C)$.

Binary Conditional Probability Assignments

In fact, all the information about μ is captured by probabilities conditioned only on *pairs of states*, i.e. by the quantities:

$$(s,t)_\mu \;:=\; \mu(\, s|\{s,t\}\,)$$

for all $s,t \in S$.

In other words: discrete conditional probability spaces can alternatively be described as binary probabilistic assignments $(,) : S \times S \to [0,1]$ on a finite set S, satisfying the axioms:

$$(s,s) = 1;$$

$$(t,s) = 1 - (s,t), \text{ for } s \neq t \ ;$$

$$(s,w) = \frac{(s,t) \cdot (t,w)}{(s,t) \cdot (t,w) + (w,t) \cdot (t,s)}$$

for $s \neq w$ and denominator $\neq 0$.

In fact, as in the case of simple probability measures, it is enough to give $n-1$ independent binary probabilities.

(Discrete) Conditional Probability Models

A *discrete conditional probability frame* is a structure $(S, \mu_a, \Pi_a)_a$, such that, for each a, (S, μ_a) is a discrete conditional probability space and (S, Π_a) is an information partition.

K_a and B_a are defined as before, but now we can also define *conditional beliefs*:

$$B_a^P Q := \{ s \in S : \mu_a(Q | P \cap s(a)) = 1 \}.$$

We obtain a *Conditional Doxastic Logic* (CDL), which is a modal-epistemic, multi-agent version of the AGM axioms.

We interpret the conditional belief statement $s \in B_a^P Q$ in the following way: if the actual state is s, then after "learning" that P *is* the case (in this actual state), agent a will believe that Q *was* the case (at the same actual state, before the learning).

In fact, for the logic (i.e. the calculation of K_a, B_a and B_a^P), only the binary probabilities conditioned by states that are in the same information cell, i.e.

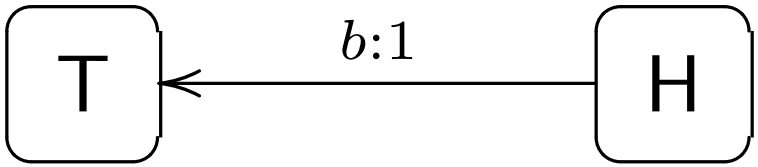
$$(s,t)_{\mu_a} \text{ for } s \sim_a t,$$

are relevant.

So we can ignore the others: even less then $n-1$ probabilities may be enough. Moreover, this provides a way to encode the information partition into the probabilistic information.

Example

We encode the above example as:



Relational Semantics: "Plausibility Models"

The standard semantics for conditional belief, or for modal logics for belief revision, is in terms of "Grove models", "Lewis spheres", "Spohn ordinal plausibility ranking", or (simpler and more modal-logic-like) Kripke models based on a *"plausibility" relation* $s \leq t$, saying that *state* t *is at least as plausible as state* s.

A *finite (epistemic-doxastic) plausibility frame* is a structure $(S, \leq_a, \sim_a)_a$, where S is finite and, for each a, $\leq_a$ is a *total (i.e. "connected", or "complete") preorder* and $\sim_a$ is an equivalence relation.

Knowledge and Conditional Belief

In a plausibility model, K_a is defined in the standard way using $\sim_a$, while

$$B_a^P(Q) := \{ s \in S : Max_{\leq_a} P \cap s(a) \subseteq Q \}$$

where

$$Max_{\leq_a} T = \{s \in T : t \leq_a s \text{ for all } t \in T\}.$$

Local plausibility

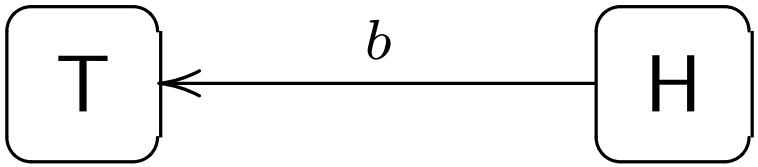
As before, only the plausibility relation between states in the same information cell are relevant; in other words, only the "local" plausibility relation

$$\trianglelefteq_a \ := \leq_a \cap \sim_a \ .$$

In fact, this relation encodes the epistemic relations $\sim_a$ as well. So, as before, we only represent $\trianglelefteq_a$, and for convenience we skip all the loops (since $\trianglelefteq$ is reflexive anyway).

Example revisited

The example above becomes:



Strict Plausibility and Equi-plausibility

For both local and global plausibility relations, we can also consider their "strict" versions:

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

$$s \lhd_a t \text{ iff: } s \trianglelefteq_a t \text{ but } t \not\trianglelefteq_a s.$$

Finally, the relation of "equi-plausibility" is the equivalence relation $\cong_a$ induced by the preorder $\trianglelefteq_a$:

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

A "Qualitative" Interpretation

Our claim is that this is in fact a *qualitative* notion, expressing (not degrees, or intensity, of beliefs) but a type of *"priority" of beliefs*, expressible best in terms of *(binary) conditional beliefs about the current state*: $s \lhd_a t$ means that, when given the (correct) information that the actual state of the system is either s or t, agent a *won't know* which of the two (since they are epistemically indistinguishable) but he'll *believe* that the state was in fact t.

$$s \lhd_a t \text{ "means" } B_a^{\{s,t\}}\{t\}.$$

Conditional beliefs are "firm"

All beliefs captured by a plausibility model are "firm" (though conditional), i.e. *believed with (conditional) probability* 1: given the condition, something is either believed or not. We give the agent some additional information about the state of the system (that it is either s or t) and we ask her a *yes-or-no question* ("Do you believe that the state is t ?"); we write $s \lhd_a t$ iff the agent's answer is "yes".

Firm, but Revisable, Beliefs

This is a firm answer, so it expresses a firm belief.

"Firm" does not imply "un-revisable" though: if later we reveal to the agent that the state in question was in fact t, she should be able to accept this new information; after all, the agent should be introspective enough to realize that her belief, however firm, was just a belief.

Argument: Comparing the two kind of models

Can we extract a (total) plausibility preorder from a (discrete) conditional probabilistic model, *in such a way that the two notions of conditional belief coincide*?

YES. But it is **NOT** given by

$$s \leq_a t \text{ iff } \mu_a(s) \leq \mu_a(t),$$

since this won't give rise to the same conditional belief.

Moreover, it is **NOT** given by

$$s \leq_a t \text{ iff } \mu_a(s|\{s,t\}) \leq \mu_a(t|\{s,t\})$$

either!

The only such notion that does the job is the following:

$$s \leq_a t \text{ iff } (t, s)_{\mu_a} \neq 0.$$

The corresponding strict preorder is:

$$s <_a t \text{ iff } (s, t)_{\mu_a} = 0$$

In other words: *the plausibility order between two states doesn't have anything to do with the relative degrees of belief in the two states, but only with conditional certainty (of one state, given both).*

Converse

So any discrete conditional probability model is a plausibility model (with the same notions of knowledge and conditional belief).

Conversely, every finite plausibility model can be "probabilized": we can define conditional probability measures for each agent, that will give rise to the same conditional beliefs.

This correspondence between the two types of models can be used to prove *completeness of CDL with respect to (both) discrete conditional probabilistic models (and plausibility models).*

Conditional Doxastic Logic (CDL)

The syntax of CDL (without common knowledge and common belief operators) is:

$$\varphi \ := \ p \mid \neg\varphi \mid \varphi \wedge \varphi \mid B_a^{\varphi}\varphi$$

while the semantics is given by the obvious compositional clauses for the interpretation map $|| \bullet ||_{\mathbf{S}} : CDL \to \mathcal{P}(S)$ in a plausibility model $\mathbf{S}$.

In this logic, the *knowledge modality* can be defined as an abbreviation, putting

$$K_a\varphi := B_a^{\neg\varphi}\bot$$

(where $\bot = p \wedge \neg p$ is an inconsistent sentence), or equivalently

$$K_a\varphi := B_a^{\neg\varphi}\varphi$$

Doxastic Propositions

A *doxastic proposition* is a map $\mathbf{P}$ assigning to each plausibility model (or conditional probabilistic model) $\mathbf{S}$ some S-proposition, i.e. a set of states $\mathbf{P_S} \subseteq S$.

The interpretation map for the logic CDL can thus be thought of as associating to each sentence φ of CDL a doxastic proposition $||\varphi||$.

We denote by *Prop* the family of all doxastic propositions. All the above operators (Boolean operators as well as doxastic and epistemic modalities) on S-propositions induce corresponding operators on doxastic propositions, defined pointwise.

A Complete Proof System for CDL

Necessitation Rule:

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

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

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

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

Full Introspection: $\vdash B_a^{\theta}\varphi \to K_a B_a^{\theta}\varphi$

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

Proof System, continued

Hypotheses are (hypothetically) accepted:

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

Minimality of revision:

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

One can add axioms for *common knowledge*, and preserve completeness.

Relations between knowledge and beliefs

Observe that

$$K_a Q = \bigcap_{P \subseteq S} B_a^P Q,$$

or equivalently:

$$s \in K_a Q \quad \text{iff} \quad s \in B_a^P Q \text{ for all } P \subseteq S. \tag{1}$$

Another identity that can be easily checked is:

$$K_a Q = B_a^{\neg Q} Q = B_a^{\neg Q} \bot \tag{2}$$

(where $\bot := \emptyset$).

Discussion of (1)

Identity (1) gives a characterization of *knowledge as "absolute", belief, invariant under any belief revision*: a given belief is "known" iff it cannot be revised, i.e. it is believed in any condition. This resembles Stalnaker's *defeasibility analysis* of knowledge, based on the idea that "if a person has knowledge, than that person's justification must be sufficiently strong that it is not capable of being defeated by evidence that he does not possess" (Pappas and Swain).

Stalnaker's Interpretation

But Stalnaker interprets "evidence" as "true information", saying: "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".

So Stalnaker's concept of defeasible knowledge differs from ours, corresponding in fact to what we will call "safe belief".

(1): Knowledge as Strongly Defeasible Belief

Our "knowledge" is more robust than Stalnaker's: it resists *any* belief revision, i.e. is not capable of being defeated by *any* evidence (including false evidence). So it corresponds to interpreting "evidence" in the above quote as meaning "any information, be it truthful or not". As a consequence, our notion of knowledge (unlike Stalnaker's defeasible knowledge) *is negatively introspective*, and thus fits better with the standard usage of "knowledge" in CS.

Discussion of (2)

Identity (2) says that *something is "known" if conditionalizing our belief with its negation is impossible* (i.e. it would lead to an inconsistent belief).

This corresponds to yet another of Stalnaker's notions of knowledge, defined by him in terms of doxastic conditionals, using precisely the above identity (2).

One of the (trivial but useful) observations arising from our work is that the notion of knowledge defined by (2) does not match Stalnaker's "defeasible" knowledge, but instead it satisfies the closely related identity (1).

Safe Beliefs

We precisely capture Stalnaker's "defeasible knowledge" in our concept of *safe belief*.

This can be defined as the Kripke modality $\Box_a$ associated to the local plausibility relation $\trianglelefteq_a$, i.e. given by

$$\Box_a Q := [\trianglelefteq_a]Q$$

for all **S**-propositions $Q \subseteq S$. We read $s \in \Box_a Q$ as saying that: *at state s, agent a's belief of Q (being the case) is safe*; or at *state s, a safely believes that Q*. Cf. van Benthem and Liu.

Safe Belief as a Weak form of Knowledge

Observe that: $\Box_a$ is an $S4$-modality (since $\trianglelefteq_a$ is reflexive and transitive), but not necessarily $S5$; i.e. *safe beliefs are truthful*

$$\Box_a Q \subseteq Q$$

and positively introspective

$$\Box_a Q = \Box_a \Box_a Q,$$

but not necessarily negatively introspective.

Also, *knowledge implies safe belief*:

$$K_a Q \subseteq \Box_a Q;$$

and *safe belief implies belief*:

$$\Box_a Q \subseteq B_a Q$$

Safe Belief is "Weakly Defeasible"

The last observation can be strengthened to characterize safe belief in a similar way to the above characterization (1) of knowledge: *safe beliefs are precisely the beliefs which are persistent under revision with any true information*; i.e.

$$s \in \Box_a Q \text{ iff:}$$

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

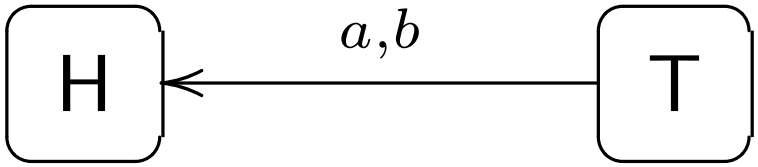
We can thus see that "safe belief" coincides with Stalnaker's non-standard notion of "(weakly defeasible) knowledge". Indeed, safe belief can be understood as a "weak" (non-negatively-introspective) form of "knowledge".

Example 2

Alice and Bob play a game, in which an (anonymous) referee takes a coin and puts it on the table in front of them, lying face up but in such a way that the face is covered (so Alice and Bob cannot see it). The goal of the game is to guess which face is up. Based on previous experience, (it is common knowledge that) Alice and Bob believe that the upper face is Heads (-since e.g. they noticed that the referee had a strong preference for Heads).

And in fact, they're right: the coin lies Heads up.

A Model for Example 2



The actual state s is the one on the left, and we neglected the anonymous referee. The agents *believe Heads is up*; but they *don't know* that Heads is up.

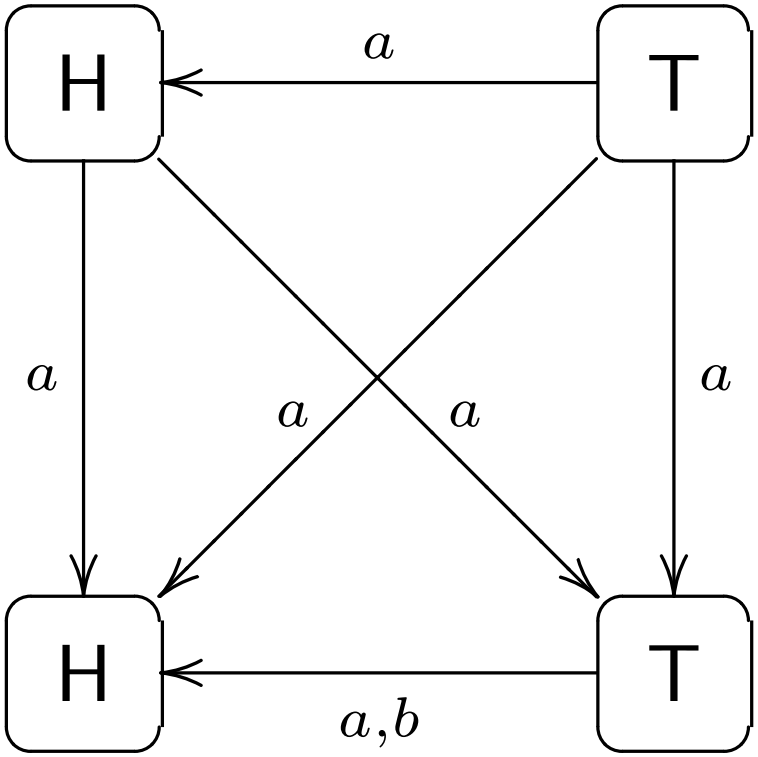
Example 3

Alice has to get out of the room for a minute, which creates an opportunity for Bob to quickly raise the cover in her absence and take a peek at the coin. He does that and so he sees that the coin is Heads up.

After Alice returns, she obviously doesn't know whether or not Bob took a peek at the coin, but she believes he didn't do it: taking a peek is against the rules of the game, and so she trusts Bob not to do that.

The model for Example 3

The situation after this action is described by the following model $\mathbf{S}'$, with actual state s'_1 in the upper left corner:



(We'll later see how this can be computed.)

Examples continued

In both Examples 2 and 3 above, Alice holds a *true belief* (at the real state) that the coin lies Heads up: the actual state satisfies $B_a\mathsf{H}$. In both cases, this true belief is *not knowledge* (since Alice doesn't know the upper face); nevertheless, in Example 2, this belief is *safe* (although it is *not known by the agent to be safe*): no additional truthful information (about the real state s) can force her to revise this belief. To see this, note that any *new* truthful information would reveal to Alice the real state s, thus confirming her belief that Heads is up. So in the first model $\mathbf{S}$ we have $s \models \Box_a\mathsf{H}$.

Examples, continued

In contrast, in Example 3, Alice's belief (that the coin is Heads up), though true, is *not safe*. There is some piece of correct information which, if learned by Alice, would make her change this belief: we can represent this piece of correct information as the doxastic proposition $\mathsf{H} \rightarrow K_b\mathsf{H}$. It is easy to see that the actual state s'_1 of the model $\mathbf{S}'$ satisfies the proposition $B_a^{\mathsf{H}\rightarrow K_b\mathsf{H}}\mathsf{T}$ (since $(\mathsf{H} \rightarrow K_b\mathsf{H})_{\mathbf{S}'} = \{s'_1, t'_1, t'_2\}$ and the maximal state in the set $s'_1(a) \cap \{s'_1, t'_1, t'_2\} = \{s'_1, t'_1, t'_2\}$ is t'_2, which satisfies T.) So, if given this information, Alice would come to wrongly believe that the coin is Tails up!

Other Identities

Believing something is the same as believing that it is safe to believe it:

$$B_a Q = B_a \Box_a Q$$

So *all* beliefs held by an agent "appear safe" to him. Moreover, the only way for an agent to *know* that one of his beliefs is safe is to actually *know it to be truthful*. *Knowing something is the same as knowing that it is safe to believe it*:

$$K_a Q = K_a \Box_a Q$$

Another important observation is that *one can characterize belief and conditional belief in terms of knowledge and safe belief*:

$$B_a Q \;=\; \tilde{K}_a \Box_a Q$$

and more generally

$$B_a^P Q \;=\; \tilde{K}_a P \to \tilde{K}_a (P \wedge \Box_a (P \to Q)).$$

(Here, $\tilde{K}_a P = \neg K_a \neg P$.)

The paradox of the "perfect believer"

Exercise: In standard doxastic-epistemic logic, one can prove that

$$B_a K_a \varphi \to \varphi.$$

For *certain belief* (with probability 1), it seems reasonable to assume the following "axiom":

$$B_a \varphi \to B_a K_a \varphi.$$

But putting these together, belief and knowledge collapse and so we get the "perfect believer's paradox":

$$B_a \varphi \to \varphi$$

Discussion

As usually stated, this just shows that the above "axiom" is wrong, and should be rejected. Various authors proposed a different solution: accepting the axiom, but giving up the principle of "negative introspection" with respect to knowledge; no paradoxical conclusion follows then.

Our solution combines the advantages of both: the above axiom is correct if we interpret "knowledge" as $\Box_a$ (but then negative introspection fails); while negative introspection holds if we interpret "knowledge" as K_a (but then the above "axiom" fails").

Conclusion

The paradox of the perfect believer arises from the conflation of two different notions of "knowledge": "Aumman" (partition-based) knowledge and "Stalnaker" knowledge (i.e. safe belief).

Degrees of Safety

In the probabilistic models, we can define "degrees of safety" of a belief, similarly to the degrees of belief.

The *degree of safety of a's belief in* Q *at state* s is given by:

$$\min_{s \in P \subseteq s(a)} \mu_a(Q|P).$$

"Safe belief" is the same as belief with degree of safety=1.

In certain contexts, it is enough to have "weakly safe" beliefs (i.e. with degree of safety > 0): though they might be lost due to truthful learning, they are never reversed (to believe the opposite).

Common Safe Belief

There is a notion of "common safe belief" that plays an important role in games: the theorems usually stated in terms of "common belief (or common knowledge) of rationality" should be in fact be stated in terms of common safe belief of rationality.

$$C\Box P = \bigwedge_{a_1,a_2,\ldots,a_n} \Box_{a_1}\Box_{a_2}\ldots\Box_{a_n}P.$$

Example: The Centipedes Game

Failure of backward induction, despite common belief in rationality at original node. E.g. The Centipedes Game.

Let us assume that "rationality" is an unchanging "fact", constraining once and forever (if true) a player's behavior. Still, a player's *beliefs* about the other's rationality *may change*. We need common belief in rationality at all (future) nodes. To make it robust, the only general assumption that we can make at the original node is *common safe belief of rationality*.

Theorem. Assuming "rationality of a player" is an unchanging feature of a player, common safe belief in rationality at the initial state of a game is enough to ensure the backwards induction solution.

The Logic of Knowledge and Safe Belief

The following set of axioms is complete for discrete probabilistic models (and plausibility models):

- K-axiom for K_a and $\Box_a$;
- $S5$-axioms for K_a;
- $S4$-axioms for $\Box_a$;
- $K_a P \to \Box_a P$;
- $K_a((P \vee \Box_a Q) \wedge (Q \vee \Box_a P)) \to K_a P \vee K_a Q$.

Dynamics: Action Models

An *(discrete conditional-probabilistic, or finite plausibility) action model* is just a (discrete conditional-probabilistic, or finite plausibility) frame Σ, together with a *precondition map*

$$pre : \Sigma \to Prop$$

associating to each element of Σ some doxastic proposition $pre(\sigma)$.

Cf. G. Aucher, H. van Ditmarsch, J. van Benthem and others.

We call the elements of Σ *(basic) doxastic actions*, and we call pre_σ *the precondition of action* σ. The basic actions $\sigma \in \Sigma$ are taken to represent some *deterministic* actions of a particularly simple nature. We only deal here with pure "belief changes", i.e. actions that do not change the "ontic" facts of the world, but only the agents' beliefs. Intuitively, the precondition defines the *domain of applicability* of σ: this action can be executed on a state s iff s satisfies its precondition.

Interpretation of plausibility ordering

The conditional probabilities μ_a, or the plausibility pre-orderings $\trianglelefteq_a$, give *the agent's (probabilistic, conditional) beliefs about the current action.*

But this should be interpreted as *beliefs about changes*, that *encode changes of beliefs*. In this sense, we use such "beliefs about actions" as a way to represent doxastic changes: the information about how the agent changes her beliefs is captured by our action plausibility relations.

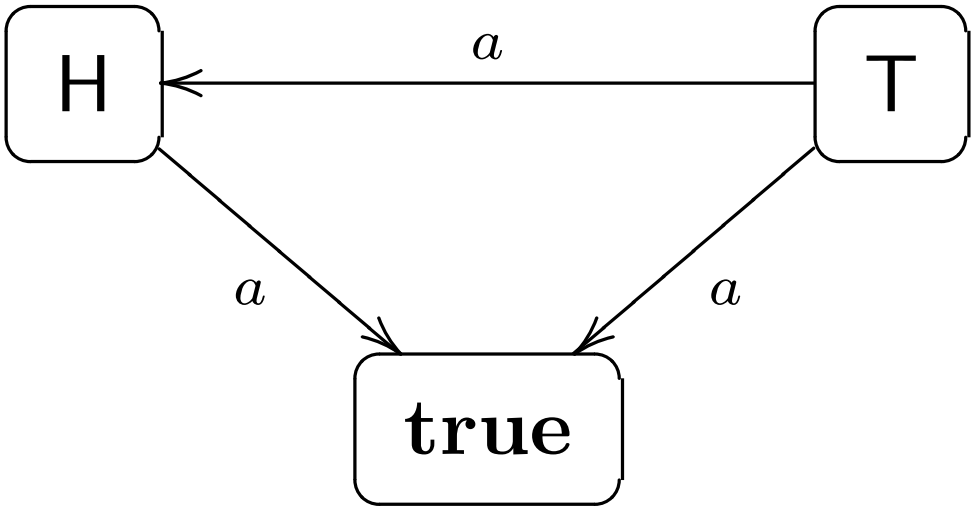
So we read $\sigma \lhd_a \sigma'$ as saying that: if (in addition to his direct observations), agent a is given some information about (the history of) the system (until and including the current action), and if this information is enough to determine that the action is *either* σ *or* σ', but *not enough to determine which* of the two, then she *believes* the action to be σ' .

Ex. 4: Private (Group) Announcements

Let us consider the *action* that produced the situation represented in Example 3 above. This was the action of Bob taking a peek at the coin, when Alice was away. Recall that Alice *believed that nothing was really happening* in her absence, though obviously she *didn't know* this (that nothing was happening). In the literature on dynamic-epistemic logic, this action is usually called *a private announcement to a subgroup*: the "insider" (Bob) learns which face is up, while the outsider Alice believes nothing is happening.

Representation

We represent this as an action model $\mathbf{\Sigma}$:



Needed: An 'Update' Operation

How can we recover the output state model (in Example 3) from the original (input) state model (Example 2) and the model of the action (Example 4)?

We need a binary *'update' operation* $\otimes$, taking any state model $\mathbf{S} = (S, \trianglelefteq_a, \| \bullet \|)_{a \in \mathcal{A}}$ and any action model $\mathbf{\Sigma} = (\Sigma, \trianglelefteq_a, pre)_{a \in \mathcal{A}}$ into an *updated state model* $\mathbf{S} \otimes \mathbf{\Sigma}$, representing the way an action lying in $\mathbf{\Sigma}$ will act on an input-state lying in $\mathbf{S}$. We call this the *update product* of the two models.

The Anti-lexicographic Product Update

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

$$S \otimes \Sigma := \{(s, \sigma) : s \in pre(\sigma)\mathbf{s}\}$$

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

$(s, \sigma) \trianglelefteq_a (s', \sigma')$ iff: either $\sigma \triangleleft_a \sigma', \sigma \sim_a \sigma'$ or $\sigma \cong_a \sigma', s \trianglelefteq_a s'$.

Consequence: the new epistemic uncertainty relation are the *product* of the two uncertainty relations:

$$(s, \sigma) \sim_a (s', \sigma') \quad \text{iff} \quad \sigma \sim_a \sigma', s \sim_a s'.$$

The anti-lexicographic preorder

What this corresponds to, in terms of the "global" (plausibility) preorder relations $\leq_a$, is:

$(s, \sigma) \leq_a (s', \sigma')$ iff either $\sigma <_a \sigma'$ or $\sigma \leq_a \sigma' \leq_a \sigma, s \leq_a s'$.

This corresponds to (one of) the (two) standard way(s) to generate a total preorder on a product from total preorders on its components: the (lexicographic and the) *anti-lexicographic preorder*(s).

Interpretation

The anti-lexicographic preorder gives "priority" to the *action* plausibility relation; this is not an arbitrary choice, but is motivated by our above-mentioned interpretation of "actions" as specific types of *belief changes*. The action plausibility relation captures what agents *really believe is going on at the moment* (while the input-state plausibility relations only capture *past beliefs*). The doxastic action is the one that "changes" the initial doxastic state, and not vice-versa.

Interpretation, continued

Giving priority to action plausibility does not in any way mean that the agent's belief in actions is "stronger" than her belief in states; it just captures the fact that, at the time of updating, *the belief about the action is what is actual, is the current belief about what is going on, while the beliefs about the input-states are in the past.* (Of course, *at a later moment*, the above-mentioned belief about action (*now* belonging to the past) might be itself revised. But this is another, *future update*.) The belief update *induced by a given action* is nothing but *an update with the (presently) believed action.*

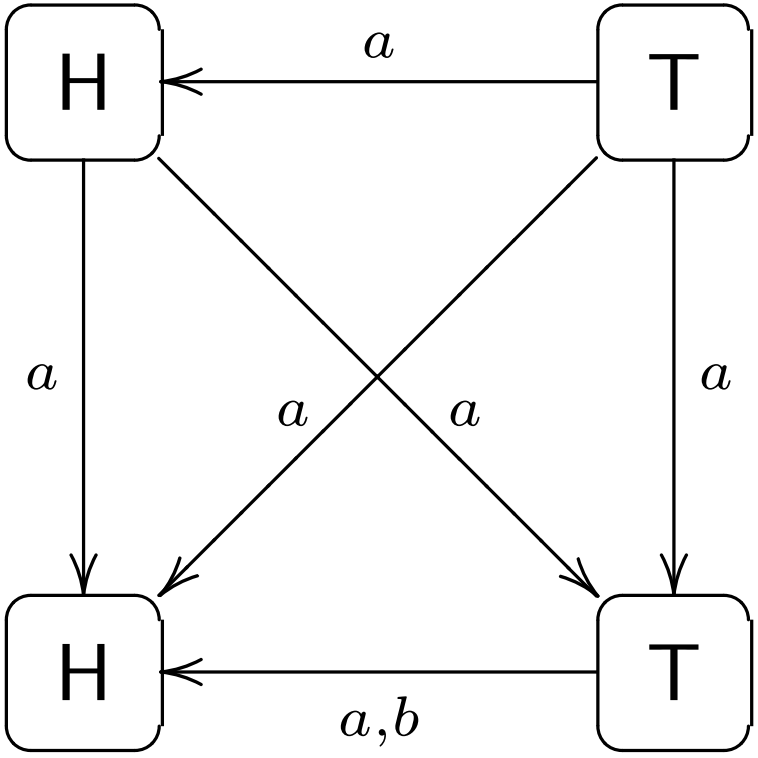
Interpretation, continued

In other words, the anti-lexicographic product update reflects our Motto above: *beliefs about changes* (as formalized in the action plausibility relations) *are nothing but ways to encode changes of belief* (i.e. ways to change the original plausibility order on states).

This simply expresses our *particular interpretation* of the (strong) plausibility ordering on actions, and is thus a matter of *convention*: we decided to introduce the order on actions to encode corresponding *changes of order* on states.

Examples of Update Products

It is easy to see that the update product of the state model in Example 2 and the action model in Example 4 is the state model in Example 3:

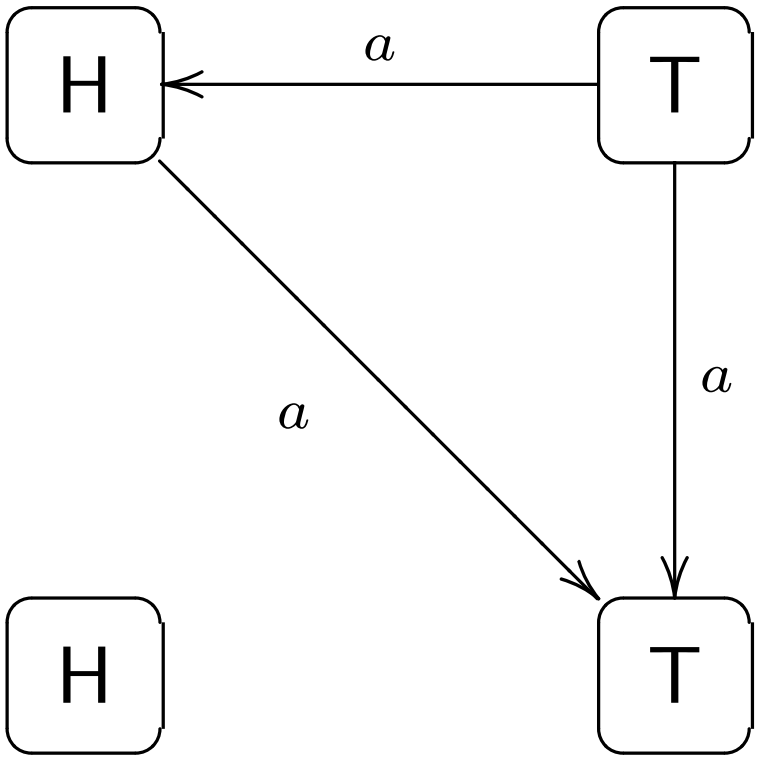


Public Announcements of "Hard Facts"

A special case of private announcements to a subgroup is when the subgroup consists of all the agents: *truthful public announcement* **P**! *of some "hard fact"*, represented by a epistemic proposition **P**. The action model consists of only one node, labeled with **P**. Its effect (via the update product) is to *delete all the non-P states* and *keep the same relations between the remaining states*.

Example

After the previous action, the two are informed that now, *if the coin lies Heads up then Bob knows it.* This corresponds to $(\mathsf{H} \rightarrow K_b\mathsf{H})!$, and the updated state model after that is:



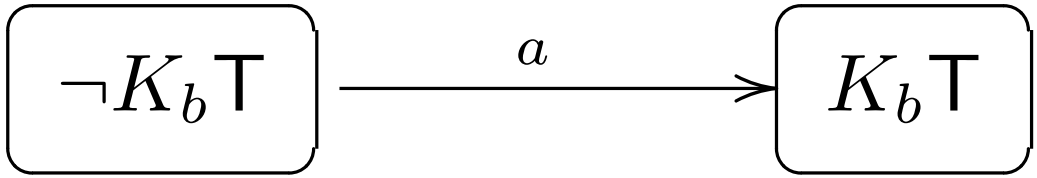
Note that, in this model, Alice came to the (wrong!) belief that T (i.e. the coin lies Tails up): as we saw, this is only possible since her previous true belief that H was *not safe*.

Example 5: Successful Lying

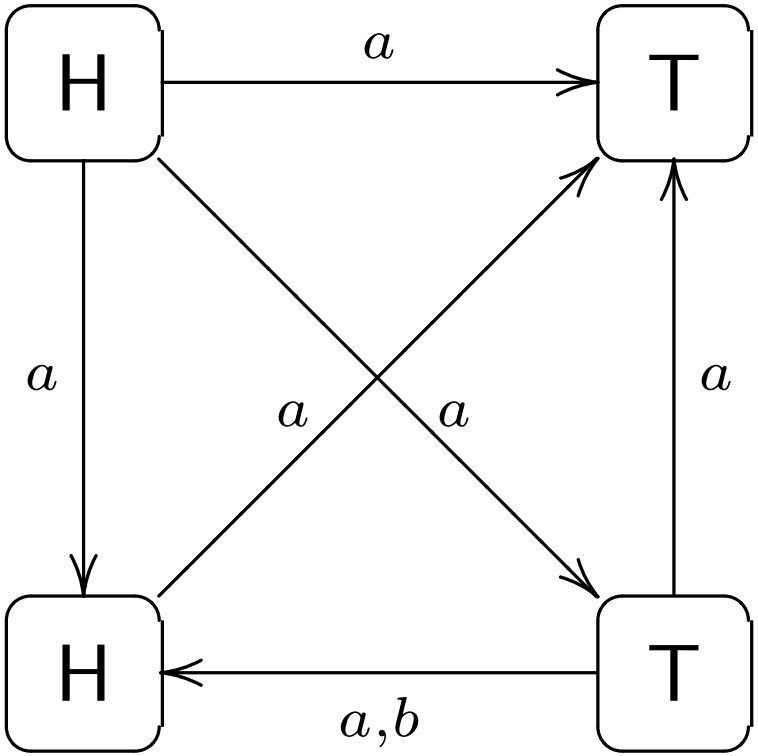
Suppose now that, after Bob took the peek in Alice's absence, Bob sneakily announces: "Look, I took a peek and saw the coin was lying *Tails up*". For our purposes, we can formalize the content of this announcement as $K_b\mathsf{T}$, i.e. saying that "Bob knows the coin is lying Tails up". This is a *public announcement*, but *not a truthful one* (though it does convey some new truthful information): it is a *lie*! We assume that it is in fact a *successful lie*: even after he admitted having taken a peek, Alice still trusts Bob, so she believes him.

Representation

The successful lying action is given by the *left node* in the following model Σ':

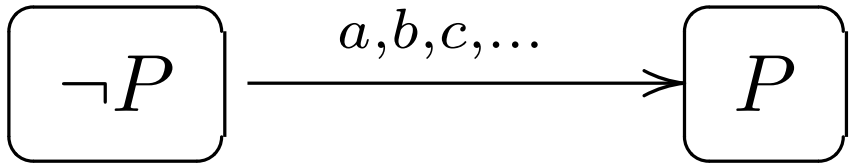


Applying this action model to the state model $\mathbf{S}'$ from Example 2, we obtain the following:


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

Public Announcement of "Soft" Facts

Suppose an announcement $P!?$ is made, in such a way that all the agents *believe* it is truthful, although (unlike in truthful public announcements of "hard" facts) they *don't know* for sure that it is truthful.



Soft Update

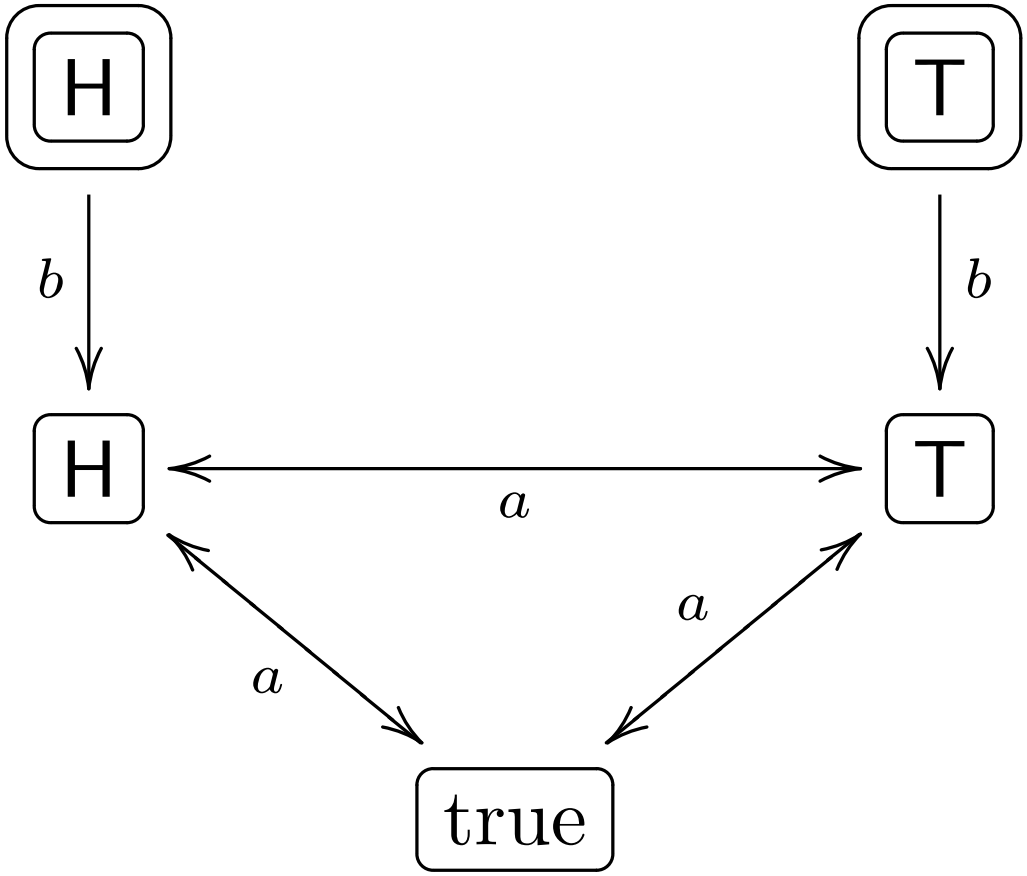
Such an announcement has been considered by other authors, who proposed a notion of "soft update" for it. But it is easy to see that this matches what we get by updating any given state model $\mathbf{S}$ with the action $P!?$ using the anti-lexicographic product update:

The new state model $\mathbf{S} \otimes P!?$ can be thought of as being from $\mathbf{S}$ by keeping the same information cells, and keeping the same plausibility order $s \leq t$ between any two states $s, t \in P$, and similarly between states $s, t \notin P$, while in the same time making all P-states more plausible than all non-P states.

Discovery of Deceit

Suppose that, in fact, when Bob was secretely taking a peek, Alice was taping him (using a hidden camera), so that she was able to *see* Bob taking a peek. Suppose also that it is common knowledge that Bob does not suspect he is being taped: he believes (though he doesn't know for sure) that there is no hidden camera.

Action Model



Here, I used double circles to mark the "real" actions (one of which is actually happening).

Exercise

Compute the updated state model $\mathbf{S} \otimes \mathbf{\Sigma}$.

Interception of messages

If the above "deceit" (Bob secretely looking at the coin) is replaced by a secret communication (say, from Charles to Bob, telling him that the coin lies Heads up), then the action corresponding to "discovery of deceit" by Alice (as above) can also be interpreted as *secret interception (wiretapping) by Alice of the secret message* (between Charles and Bob).

A Probabilistic Version

Given (discrete conditional probabilistic) models S, Σ, we put on $S \otimes \Sigma$:

If $(\sigma, \tau) \neq 0$, and either $s \neq t$ or $\sigma \neq \tau$, then

$$(s\sigma, t\tau) \;=\; \frac{(s,t)\cdot(\sigma,\tau)}{(s,t)\cdot(\sigma,\tau) + (1-(s,t))\cdot(1-(\sigma,\tau))}\,;$$

if $(\sigma, \tau) = 0$, then

$$(s\sigma, t\tau) \;=\; 0\,;$$

if $s = t$ and $\sigma = \tau$, then

$$(s\sigma, t\tau) \;=\; 1.$$

Justification: Jeffrey's Rule

Why is that natural?

Well, this is nothing but

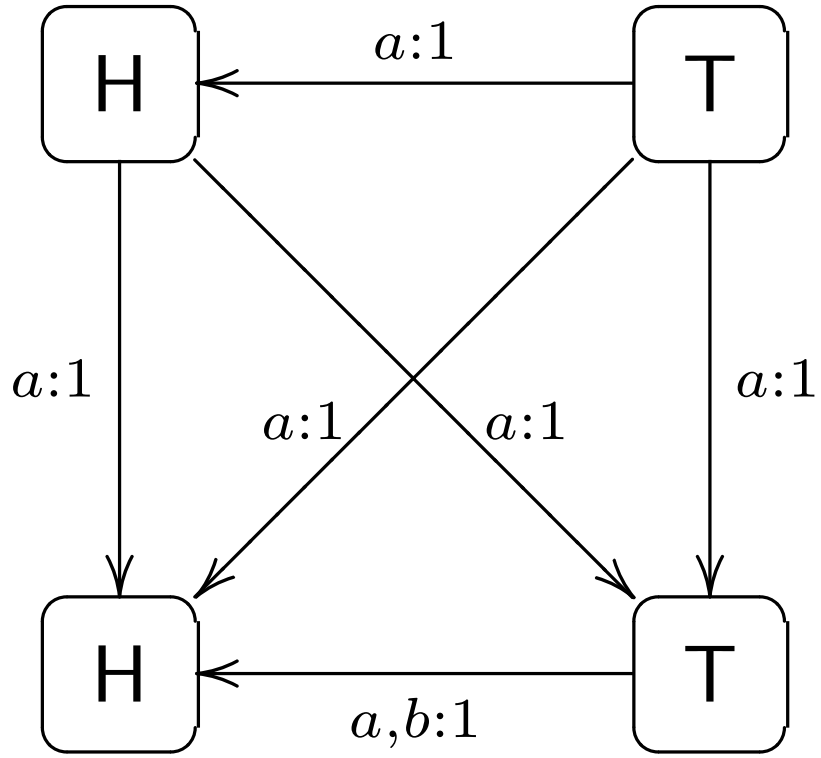
$$(s\sigma, t\tau) = \lim_{x \to (s,t)} \frac{x \cdot (\sigma, \tau)}{x \cdot (\sigma, \tau) + (1 - x) \cdot (\tau, \sigma)}$$

where the limit is taken over x's such that the denominator is $\neq 0$.

This can itself be justified on the basis of a natural generalization to conditional probabilistic spaces of the so-called *Jeffrey's rule* for updating probabilities.

Example

We reinterpret the state models in Examples 2 and 3 (and the action model in Example 4) probabilistically: both initially believed with certainty that H, then Bob took a peek, while Alice was certain he didn't. The resulting model is essentially the same, except that we add probability 1 everywhere:


H
T
a:1
a:1
a:1
a:1
a:1
H
T
a,b:1

Dynamic Modalities

Given a doxastic action σ (living in some action model $\mathbf{\Sigma}$, we can define a corresponding dynamic modality, capturing the *weakest precondition* of σ: for every proposition $\mathbf{P}$, the proposition $[\sigma]\mathbf{P}$ is given by

$$([\sigma]\mathbf{P})_{\mathbf{S}} := \{s \in S : (s, \sigma) \text{ (if defined) } \in \mathbf{P}_{\mathbf{S}\otimes\mathbf{\Sigma}}\}$$

The Laws of Dynamic Belief Revision

$$[\alpha]K_a\mathbf{P} \quad = \quad pre_\alpha \rightarrow \bigwedge_{\beta \sim_a \alpha} K_a[\beta]\mathbf{P}$$

$$[\alpha]\Box_a\mathbf{P} \quad = \quad pre_\alpha \rightarrow \bigwedge_{\alpha \lhd_a \beta} K_a[\beta]\mathbf{P} \wedge \bigwedge_{\alpha \cong_a \gamma} \Box_a[\gamma]\mathbf{P}$$

Here, $=$ is logical equivalence, $\sim_a$ is epistemic indistinguishability between actions, $\lhd_a$ is *strict plausibility* order on actions, while $\cong_a$ is *equi-plausibility* of (indistinguishable) actions:

$$\alpha \cong_a \beta \text{ iff } \alpha \trianglelefteq_a \beta \text{ and } \beta \trianglelefteq_a \alpha.$$

Other Reduction Laws

$$
\begin{aligned}
[\alpha]p &= pre_\alpha \to p \\
[\alpha]\neg\mathbf{P} &= pre_\alpha \to \neg[\alpha]\mathbf{P} \\
[\alpha](\mathbf{P} \wedge \mathbf{Q}) &= pre_\alpha \to [\alpha]\mathbf{P} \wedge [\alpha]\mathbf{Q}
\end{aligned}
$$

Computing Belief Updates

Using above-mentioned characterizations of (conditional) belief, we can deduce reduction laws for B_a and $B_a^{\mathbf{P}}$. We give here the one for simple belief, in the case of a truthful public announcement $\mathbf{P!}$:

$$[\mathbf{P!}]B_a\mathbf{Q} = \mathbf{P} \rightarrow B_a^{\mathbf{P}}[\mathbf{P!}]Q$$

Analysis: *A Cryptographic Attack.*

Two agents, a and b, share a secret key, so that they can send each other encrypted messages over some communication channel.

But the channel is not secure: some outsider c may intercept the messages or prevent them from being delivered (although he cannot read them, or send around instead his own encrypted messages, since he doesn't have the key).

Suppose also the *encryption method* is publicly known (although the key is secret). It is also known that a is the only one who knows some important *secret* (say, whether some *fact P holds or not*). Suppose now that A sends an encrypted message to B, communicating the secret (whether P or $\neg P$). b gets the message, and he's convinced it must be authentic, since it has been encrypted with the secret key.

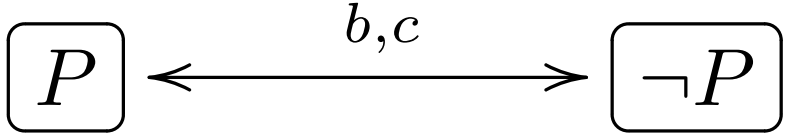
To make sure b got the message, the protocol requires him to *publicly acknowledge its receipt* (i.e. to broadcast over a completely public channel, but impossible to block or falsify, a message saying "Yes, I got the encrypted message"). So both of them will be convinced that they now share the secret, and that c doesn't know the secret, although he may *suspect* they know it. (But they think c can't be sure of that either, since for all he knows the message might have been just junk).

However, suppose that agent c is the only one to notice two features of the specific encryption method: first, that the shape of the already encrypted message can show whether it contains a secret (P or $\neg P$) or it's just junk; second, that without knowing the key or reading the content, he can modify the encrypted message in a trivial way, so that the encoded bit is changed to its opposite: the message will read "P" if it was $\neg P$, and vice-versa. (Encryption methods having similar defects have been already used.)

So the outsider c will secretly intercept the message, change it appropriately and send it to b. Of course, c will never know the secret: he still can't decrypt messages; but instead he has successfully manipulated his opponents' beliefs: a and b will *mistakenly believe that they now share the secret*; while in fact B got the "wrong secret" instead!

Representation of the initial situation

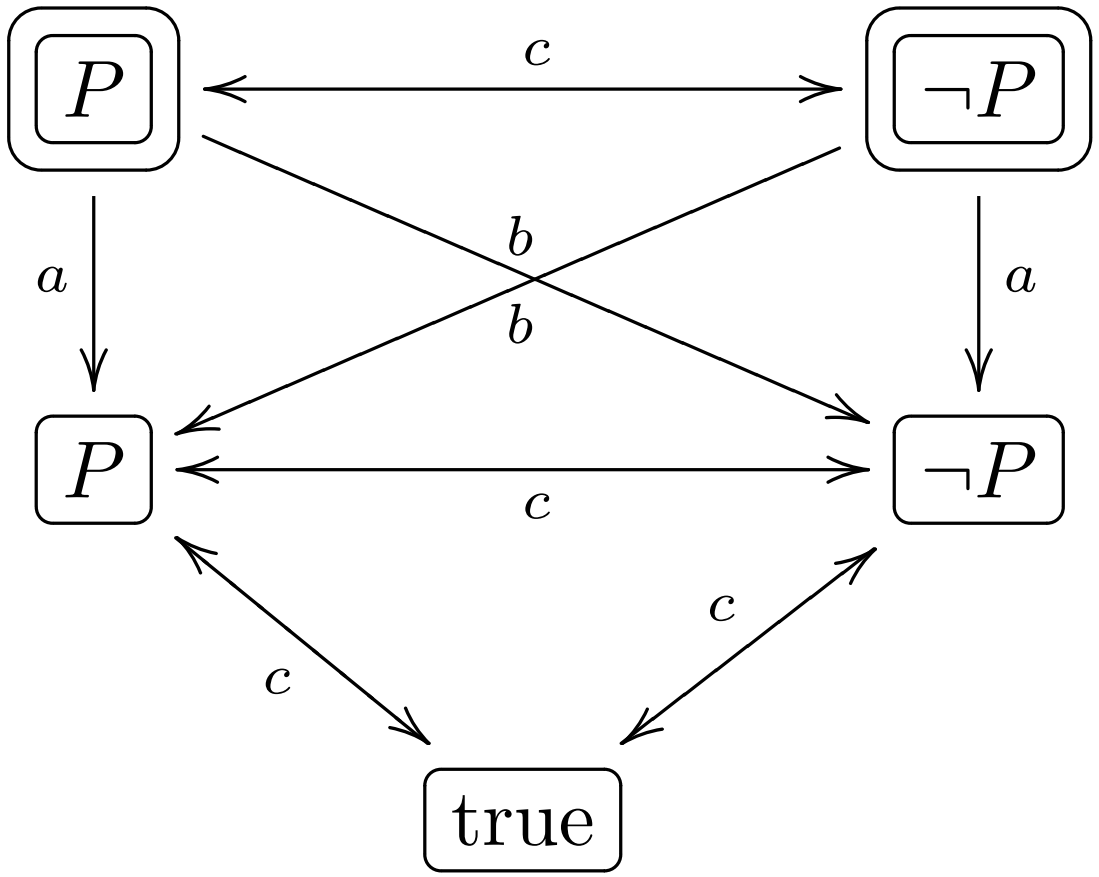
The initial situation in the scenario from the cryptographic attack above is given by:



There are plausibility arrows for b and c between any two states: This reflects the fact that b and c don't know which of these states is the real one (they don't know the secret) and moreover they consider both plausible. In contrast, a knows the state.

Representation of the cryptographic atack

The epistemic program describing the above *cryptographic attack* (including the simultaneous sending of the secret by a, interception and manipulation by c, and receiving and acknowledgement by b) has the following representation:


P
c
¬P
a
b
b
a
P
¬P
c
c
c
true

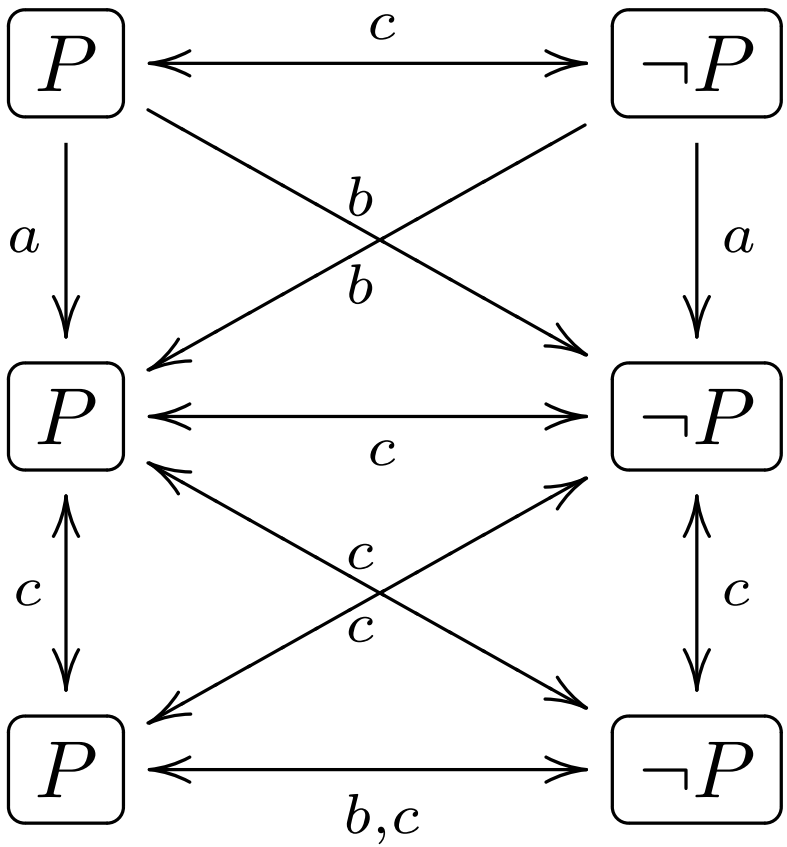
Justification

Only one of the two actions on top (call them α, β) can *really* happen: these are actions in which the "secret" (either P or $\neg P$) is intercepted, modified and resent to b. Only c is aware of the possibility of these actions, but he doesn't know which of them is happening. Moreover, we assume c is a *cautious* player: *he only believes what he knows*. This means he considers α and β equi-plausible: so there are c-arrows between these top nodes.

The two nodes in the middle row (call them α', β') are possible actions that a or b may think to be happening: they represent what *would have* happened if the encryption method was safer. a and b are completely deceived: a knows what message she sent, but she wrongly thinks that b has got it; while b is even wrong about the secret: his arrows point to actions with the wrong preconditions. The bottom node γ corresponds to sending a 'junk' (or empty) message. c cannot distinguish between these three actions, so (being cautious) considers them equi-plausible.

The update Product

Taking the update product of the state model given above for the initial situation *before the attack* with the above action model of the *attack itself*, we obtain a new state model, representing the *final epistemic situation after the attack*:


P
¬P
c
a
b
b
a
P
¬P
c
c
c
c
c
P
¬P
b,c

Reasoning about the Cryptographic Attack

Let $\pi = \alpha \cup \beta$ be the cryptographic attack program. We want to prove that, if the secret was that P was true (as in the initial situation drawn above) then *after the attack, a will believe that b knows this secret* P. This is expressed by the validity of

$$P \to [\pi] B_a K_b P$$

To show this, apply the reduction axioms (using the above notations $\alpha, \beta, \ldots$ for the action nodes in π's graph):

$$[\pi] B_a K_b P = [\alpha] B_a K_b P \wedge [\beta] B_a K_b P$$

For α we prove:

$$\begin{aligned}
[\alpha]B_a K_b P \quad = \quad & [\alpha]\tilde{K}_a \Box_a K_b P \\
& P \to \neg[\alpha]K_a \neg \Box_a K_b P \\
& P \to \neg(P \to K_a[\alpha]\neg \Box_a K_b P \wedge \\
& \wedge P \to K_A[\alpha']\neg \Box_a K_b P)
\end{aligned}$$

So we push dynamic modalities inside, past the epistemic ones (while also changing the actions). Applying again the axiom to programs α' and β', agent b and formula P, we push dynamic modalities further. Finally, we use the Preservation of "Facts" to eliminate dynamic modalities altogether.