Logics for Knowledge and Belief

Chapter 17 in M. Wooldridge "An Introduction to MultiAgent Systems"

1.1. MAIN THEMES: Examples and Jokes

Examples of Multi-agent Systems:

1. **Computation**: a network of communicating computers; the Internet

2. **Games**: players in a game, e.g. chess or poker

3. **AI**: a team of robots exploring their environment and interacting with each other

4. **Cryptographic Communication**: some communicating agents ("principals") following a cryptographic protocol to communicate in a private and secret way

5. **Economics**: economic agents engaged in transactions in a market
6. **Society**: people engaged in social activities
7. **Politics**: "political games", diplomacy, war.
8. **Science**: a community of scientists, engaged in creating theories about Nature, making observations and performing experiments to test their theories.

"Dynamic" and "informational" systems

Such multi-agent systems are *dynamic*: agents "do" some *actions*, changing the system by interacting with each other. E.g. of actions: moves in a game, communicating (sending, receiving or intercepting) messages, buying/selling etc.

On the other hand, these systems are also *informational* systems: agents acquire, store, process and exchange *information* about each other and the environment. This information may be inherently truthful, and then it's typically called *knowledge*. Or the information may be only plausible (or probable), well-justified, but still possibly false; then it's called *(justified) belief*.

Nested Knowledge in Chinese Philosophy

Chuangtze and Hueitse had strolled onto a bridge over the Hao, when the former observed,

"See how the small fish are darting about! That is the happiness of the fish".

"You are not fish yourself", said Hueitse, "so how can you know the happiness of the fish?"

"You are not me", retorted Chuangtse, "so how can you know that I do not know?"

Chuangtse, c. 300 B. C.

Knowledge and Uncertainty

Uncertainty is a corollary of imperfect knowledge (or "imperfect information"):

A *game of imperfect information* is one in which some moves are hidden, so that the players don't know all that was going on: they only have a partial view of the situation.
Example: *poker* (in contrast to chess).

A player may be *uncertain* about the real situation of the game at a given time: e.g. they simply *cannot distinguish* between a situation in which another player has a winning hand and a situation in which this is not the case. For all our player knows, *these situations are both "possible"*.

Evolving Knowledge

The knowledge a player has may *change* in time, due to his or other players' actions.

For instance, he can do some move that allows him to *learn* some of the cards of the other player. As a general rule, players try to minimize their uncertainty and increase their knowledge.

Wrong Beliefs: Cheating

In their drive for more knowledge and less uncertainty, players may be induced to acquire a false "certainty": they will "know" things that are not true.

Example: *bluffing* (in poker) may induce your opponent to believe you have a winning hand, when in fact you don't.

Notice that such a wrong belief, once it becomes "certainty", might look just like knowledge (to the believer):

your opponent may really think he "knows" you have a winning hand.

Distributed Knowledge

Suppose Alice would like to know with whom did Bob go out for dinner. But she only knows he went out with one of his two friends, Charles or Eve (but not both: they can't stand each other). On the other hand, suppose in fact Bob went out with Eve; so Charles must obviously know that Bob didn't go out with him.

If only Alice and Charles could put their knowledge *together*, they would find out that Bob went out with Eve. So Alice gives a phone call to Charles, they chat and find out.

Before the chat, none of them knew that Bob has gone out with Eve, but this fact was *distributed knowledge* between the two of them: putting their knowledge together was enough to ensure the knowledge of this new fact.

“Everybody knows...”

Suppose that, **in fact**, *everybody in the class already knows everything about epistemic logic.*

The teacher, as well as all students, come to class. The teacher gives a whole lecture on epistemic logic, talking loud in front of the students.

Did anybody learned anything new from this?

Answer: YES.

WHY? Think about it!

Common Knowledge

Before the class, each student knows about epistemic logic, but maybe (s)he doesn't know if the others know anything about it.

After the class, everybody (including the teacher) knows that everybody else knows that everybody else knows... etc, everything about epistemic logic.

They all have now **common knowledge** of epistemic logic!

1.2. Puzzles and Paradoxes

I will now illustrate the above-mentioned themes via a number of **epistemic puzzles**:

these are stories, involving multi-agent informational dynamics, that appear to lead to surprising, "puzzling", or even paradoxical conclusions.

Epistemic Puzzle no. 0: The Coordinated Attack

Two divisions of the same army, commanded by general A and general B, are camped on two hiltops overlloking a valley. In the valley awaits the enemy (C). It is clear that if both divisions attack simultaneously they will win, while if only division attacks it will be defeated. So neither general will attack unless he is absolutely sure that the other will attach with him. General A sends a messenger to general B to coodinate a simultaneous attack, by conveying the message "attack at dawn". But it is possible that the messenger would be captured by the enemy. Fortunately, on this particular night, everything goes smooth. How long it will take them to coordinate an attack?

Well, B cannot attack at dawn, after receiving the message, since he's still not sure that A knows he received the message; indeed, A might think it possible the messenger was captured, in which case A will not attack at dawn, since he'll fear B won't attack. So B has to send another messenger to A to confirm the receipt of the first message (an 'acknowledgment'). After receiving it, A knows that B got the first message. But he still cannot attack, since he's not sure B will: for all that B knows, *his messenger* might have been captured (in which case A wouldn't know the first message was received). So A has to send back to B another messenger, confirming the receipt of the previous acknowledgment.

This goes forever, without achieving any coordination: even if no messenger is captured, one can show that no finite number of successful deliveries of "acknowledgments to acknowledgments" can allow the generals to attack!

Epistemic Puzzle no. 1: To learn is to falsify

Our starting example concerns a "*love triangle*": suppose that Alice and Bob are a couple, but Alice has just started an affair with Charles.

At some point, Alice sends to Charles an email, saying:

"**Don't worry, Bob doesn't know about us**".

But suppose now that Bob reads the message (by, say, secretely breaking into Alice's email account).

Then, paradoxically enough, after seeing (and believing) the message which says he doesn't know..., *he will know*!

So, in this case, **learning the message is a way to falsify it**.

As we'll see, this example shows that **standard postulates in a well-known logical Theory of Belief Revision may fail to hold in such complex learning actions**, in which the message to be learned refers to the knowledge of the hearer.

Epistemic Puzzle no. 2: Self-fulfilling falsehoods

Suppose Alice becomes somehow *convinced that Bob knows everything* (about the affair).

This is false (Bob doesn't have a clue), but nevertheless she's so convinced that she makes an attempt to warn Charles by sending him a message:

"**Bob knows everything about the affair!**".

As before, Bob secretely reads (and believes) the message. While false at the moment of its sending, the message becomes true: *now he knows.*

So, *communicating* a false belief (i.e. Alice's action) might be a self-fulfilling prophecy: **Alice's false belief, once communicated, becomes true**.

In the same time, the action of (reading and) *believing* a falsehood (i.e. Bob's action) can be self-fulfilling: **the false message, once believed, becomes true**.

Epistemic Puzzle no. 3: Muddy Children

Suppose there are 4 children, all of them being good logicians, exactly 3 of them having dirty faces. *Each can see the faces of the others, but doesn't see his/her own face.*

The father publicly announces:

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

Then the father does another paradoxical thing: *starts repeating over and over the same question* "**Do you know if you are dirty or not, and if so, which of the two?**"

After each question, the children have to *answer publicly, sincerely and simultaneously, based only on their knowledge, without taking any guesses.* No other communication is allowed and nobody can lie.

One can show that, after 2 rounds of questions and answers, **all the dirty children will come to know they are dirty**! So they give this answer in the 3rd round, after which **the clean child also comes to knows she's clean**, giving the correct answer at the 4th round.

Muddy Children Puzzle continued

First Question: *What's the point of the father's first announcement ("At least one of you is dirty")?*

Apparently, this message is not informative to any of the children: the statement was already known to everybody! But the puzzle wouldn't work without it: in fact this announcement adds information to the system! The children implicitly learn some new fact, namely the fact that what each of them used to know *in private* is now *public knowledge*.

Second Question: *What's the point of the father's repeated questions?*

If the father knows that his children are good logicians, then at each step the father knows already the answer to his question,

before even asking it! However, the puzzle wouldn't work without these questions. In a way, it seems the father's questions are *"abnormal"*, in that they don't actually aim at filling a gap in father's knowledge; but instead they are part of *a Socratic strategy of teaching-through-questions.*

Third Question: *How can the children's statements of ignorance lead them to knowledge?*

Feminist Version: The Amazon Island

On the island of Amazonia, women are dominant and they always obey their Queen, who never lies. On the morning of Day 1, the queen (truthfully) tells the women:

"*At least one of your husbands is a cheater. Each of you already knows whether or not the other women's husbands are cheating, but I forbid you communicate this to the other women. However, if (without any such communication) in any day you get to know before noon that your own husband is cheating on you, I command you to shoot him that same day at noon in the main square (where everybody can hear and see when somebody's shot).*"

For 16 days, nothing happens. Then, in the 17th day, shootings are heard.

Question: How many husbands died?

Puzzle no 4: Sneaky Children

Let us modify the last example a bit.

Suppose the children are somehow rewarded for answering as quickly as possible, but they are punished for incorrect answers; thus they are interested in getting to the correct conclusion as fast as possible.

Suppose also that, **after the first round of questions and answers, two of the dirty children "cheat"** on the others by **secretly announcing each other that they're dirty**, while none of the others suspects this can happen.

As a result, *they both will answer truthfully "I know I am dirty" in the second round.*

One can easily see that the **third dirty child will be totally deceived, coming to the "logical" conclusion that... she is clean**!

So, *after giving this wrong answer in the third round*, she ends up by being punished for her credulity, despite her impeccable logic.

Clean Children Always Go Crazy

What happens to the clean child?

Well, **assuming she doesn't suspect any cheating, she is facing a contradiction**: two of the dirty children answered too quickly, coming to know they're dirty before they were supposed to know!

If the third child simply updates her knowledge monotonically with this new information (and uses classical logic), then she ends up believing everything: **she goes crazy**!

Feminist Version: The Dangers of Mercy

In the Amazonia version of the story, assume that again there are exactly 17 cheating husbands (out of 1 million husbands on the island), while the rest of 999983 husbands are faithful.

Consider what happens now if *all the wives of the 17 cheating husbands secretely decide to break the Queen's rules*, by quietly sparing the lives of their husbands, even when they get to know that they are cheating. We also assume that *all the other wives do not suspect this*: not only that *they strictly obey by the Queen's rules*, but *they believe that it is common knowledge that everybody else obeys by those same rules.*

It's easy to see that, in this case, *17 days will pass without any shooting.* But it's also easy to show that in the 18th day, shots will be heard. **How many husbands will die in this scenario? How many of these are innocent?**

Logics for Knowledge and Belief

Chapter 17 in M. Wooldridge "An Introduction to MultiAgent Systems"

Lecture 1.2. Kripke Models and Logics

Relational models for modal logic in general were introduced in 1959-63 by Kripke, while epistemic Logic was first formalized by Hintikka (1962), who also sketched the first steps in formalizing doxastic logic.

"Episteme" is Greek for *knowledge*, while "doxa" is Greek for *opinion* or *belief*.

The field was further developed and studied by philosophers (Stalnaker, Parikh etc.), economists (Aumann), computer-scientists (Halpern, Vardi, Fagin etc.) and logicians (van Benthem, Moss etc).

Kripke Models

For a given set $Prop = \{p, q, \ldots\}$ of "*atomic propositions*" (also known as basic *facts*) and a given finite set $\mathcal{A} = \{1, 2, \ldots, n\}$ of *agents*, a **multi-agent (Kripke) model** is a triple

$$\mathbf{M} \quad = \quad (W, R_1, \ldots, R_n, \nu\,)$$

consisting of

1. a set W of "*possible worlds*" (or "*states*")
2. a family of binary relations $R_a \subseteq W \times W$, one for each agent $a \in \mathcal{A}$, called "*accessibility relations*"
3. a map $\nu : W \to \mathcal{P}(Prop)$, called *valuation*, that assigns to each world $w \in W$ some set $\nu(w) \subseteq Prop$ of atomic propositions.

The valuation ν expresses the **factual content** of a given world: intuitively, $\nu(w)$ is the set of basic facts (atomic propositions) that are *true* in world w.

The accessibility relations R_a express the **agents' uncertainty** between various possible worlds.

wR_aw' means that, in world w agent a considers world w' to be **possible**:

The set of accessible (=possible) worlds

For $w \in W$ and $a \in \mathcal{A}$, we introduce the notation

$$w(a) = \{w' \in W : wR_a w'\}$$

for the set of worlds that are R_a-accessible from world w.

Intuitively, $w(a)$ gives **the set of worlds that agent a considers possible in world w.**

Modalities

For each agent $a \in Agents$, we introduce two new propositional operators $\Box_a$ and $\Diamond_a$, called *modalities*.

For every sentence φ, the operator $\Box_a\varphi$ universally quantifies over R_a-accessible worlds, while the operator $\Diamond_a\varphi$ existentially quantifies over R_a-accessible worlds:

Intuitively, $\Box_a\varphi$ is true at world w iff φ is true at **all** worlds w' such that wR_aw';

$\Diamond_a\varphi$ is true at world w iff φ is true at **some** worlds w' such that wR_aw'.

The Epistemic/Doxastic Interpretation

$\Box\varphi$ may be interpreted as **knowledge** (in which case we use the notation $K_a\varphi$ instead) or **belief** (in which case we use $B_a\varphi$ instead), depending on the context.

Its *existential dual*

$$\Diamond_a\varphi = \neg\Box_a\neg\varphi$$

denotes a sense of **"epistemic/doxastic possibility"**.

The notation $\Box$ is **neutral between the two interpretations (knowledge or belief)**: it can denote either, depending on context.

Kripke Models for Knowledge and Belief

In a context when we want to stress the **knowledge**, we use the notation $K_a\varphi$ instead of $\Box_a\varphi$, and we denote by $\sim_a$ the underlying binary relations R_a. In this case, we call the Kripke models $(W, \sim_1, \ldots, \sim_n, \nu)$ **epistemic models**.

When we want to the stress the **belief** interpretation, we use the notation $B_a\varphi$ instead of $\Box_a\varphi$, and we denote by $\rightarrow_a$ the underlying binary relation R_a. In this case, we call the models $(W, \rightarrow_1, \ldots, \rightarrow_n, \nu)$ **doxastic models**.

When we want to represent *both knowledge and belief*, we use **epistemic-doxastic models** of the form
$\mathbf{M} = (W, \sim_1, \ldots, \sim_n, \rightarrow_1, \ldots, \rightarrow_n, , \nu)$, with K_a interpreted as the Kripke modality for the knowledge relation $\sim_a$, and B_a as the modality for the belief relation $\rightarrow_a$.

The Language of Modal Logic

The language of Modal Logic is simply obtained by adding to Propositional Logic the modalities $\Box_a\varphi$ and $\Diamond_a\varphi$, for each agent $a \in \mathcal{A}$.

So the sentences of Modal Logic are formed as follows:

$$\text{Atomic Sentence } \varphi \quad = \quad p \text{ (atomic proposition in } Prop \text{)},$$

while complex sentences are formed using connectives

$$\neg\varphi, \quad \varphi_1 \wedge \varphi_2, \quad \varphi_1 \vee \varphi_2, \quad \Box_a\varphi, \quad \Diamond_a\varphi,$$

where we can also also use the standard implication and biconditionals, defined as abbreviations:

$$\varphi \Rightarrow \psi \quad = \quad \neg\varphi \vee \psi,$$

$$\varphi \Leftrightarrow \psi \quad = \quad (\varphi \Rightarrow \psi) \wedge (\psi \Rightarrow \varphi).$$

Semantics: the Truth Map

The **meaning** of a modal sentence φ in a model $\mathbf{M} = (W, R_1, \ldots, R_n, \nu)$ is given by the **truth map**, that maps sentences into sets $\|\varphi\|_{\mathbf{M}} \subseteq W$ of possible worlds (intuitively: the set of worlds in which φ is **true**). When the model is fixed, we *skip the subscript* $\mathbf{M}$, writing $\|\varphi\|$.

The truth map is defined inductively, by putting:

$$\|p\| = \{w \in W : p \in \nu(w)\}$$

$$\|\neg\varphi\| = W \setminus \|\varphi\|,$$

$$\|\varphi \wedge \psi\| = \|\varphi\| \cap \|\psi\|$$

$$\|\varphi \vee \psi\| = \|\varphi\| \cup \|\psi\|$$

$$\|\Box_a\varphi\| = \{w \in W : w(a) \subseteq \|\varphi\|\}$$

$$\|\Diamond_a\varphi\| = \{w \in W : w(a) \cap \|\varphi\| \neq \emptyset\}.$$

Equivalent notation: Satisfaction Relation

For $w \in W$, we write $w \models_{\mathbf{M}} \varphi$, or $(\mathbf{M}, w) \models \varphi$, for the **satisfaction relation**: *φ is true at world w in model* $\mathbf{M}$. Again, we can *skip the subscript* when M is fixed.

This relation is defined inductively by:

$$w \models p \quad \text{iff} \quad p \in \nu(w)$$

$$w \models \neg\varphi \quad \text{iff} \quad w \not\models \varphi$$

$$w \models \varphi \wedge \psi \quad \text{iff} \quad \text{both } w \models \varphi \text{ and } w \models \psi$$

$$w \models \varphi \vee \psi \quad \text{iff} \quad \text{either } w \models \varphi \text{ or } w \models \psi \text{ (or both)}$$

$$w \models \Box_a \varphi \quad \text{iff} \quad \forall w' \in W : \ (\, w R_a w' \text{ implies } w' \models \varphi \,)$$

$$w \models \Diamond_a \varphi \quad \text{iff} \quad \exists w' \in W : \ (\, w R_a w' \text{ and } w' \models \varphi \,)$$

Equivalent Notations

Wooldridge's book uses $(\mathbf{M}, w) \models \varphi$ instead of $w \models_{\mathbf{M}} \varphi$.

NOTE: **The truth map encodes exactly the same information as the satisfaction relation**.

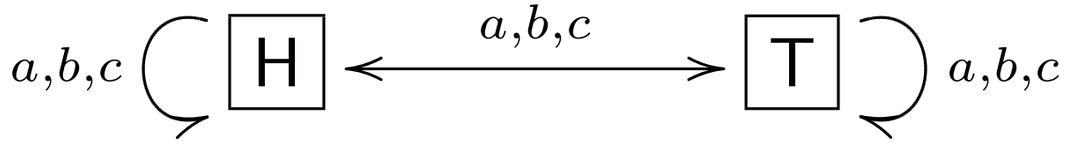
Indeed, the two are **inter-definable**:

$$\|\varphi\|_{\mathbf{M}} \quad = \quad \{w \in W : w \models_{\mathbf{M}} \varphi\},$$

$$w \models_{\mathbf{M}} \varphi \quad \text{iff} \quad w \in \|\varphi\|_{\mathbf{M}}$$

Scenario 1: the concealed coin

Two players a, b and a referee c play a game. In front of everybody, the referee throws a fair coin, catching it in his palm and fully covering it, before anybody (including himself) can see on which side the coin has landed.



This model can be described formally (with no drawings), by giving names to the states, agents and atomic propositions, and specifying the sets $W, Prop, \mathcal{A}$, the valuation ν and the knowledge relations R_a, R_b, R_c:

$W = \{w, w'\}$, $Prop = \{\mathsf{H}, \mathsf{T}\}$, $\nu(w) = \{\mathsf{H}\}$, $\nu(w') = \{\mathsf{T}\}$, $\mathcal{A} = \{a, b, c\}$, $R_a = R_b = R_c = \{(w, w), (w, w'), (w', w), (w', w')\}$.

Examples of true sentences

Check that in this model

$$w \models \neg\Box_a \mathsf{H} \wedge \neg\Box_a \mathsf{T} \wedge \Box_a(\mathsf{H} \vee \mathsf{T})$$

and similarly for all other agents and all other worlds:

no agent knows if the coin is lying Heads up or Tails, but all agents know that it lies either Heads or Tails up.

Epistemic ($S5$-) Models

Note that in all the above models, the relations R_a are equivalence relations!

An **epistemic model** (or $S5$**-model**) is a Kripke model in which all the accessibility relations are **equivalence relations**, i.e. **reflexive**, **transitive** and **symmetric**.

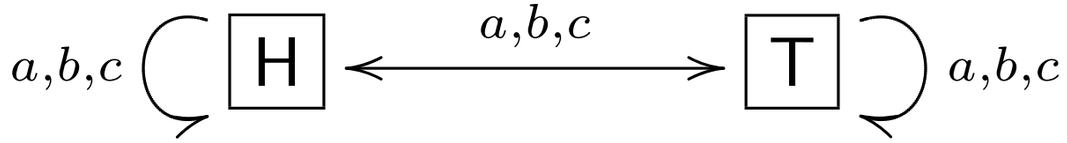
In an epistemic model, $\Box_a$ is interpreted as **knowledge**, and denoted by K_a.

Everybody Knows

We can express the sentence "**everybody knows/believes** φ " in our language by taking the conjunction of knowledge/belief statements for *all agents*:

$$E\Box\varphi \;:=\; \bigwedge_{a\in\mathcal{A}} \Box_a\varphi$$

EXAMPLE: In the model from Scenario 1



everybody knows that the coin lies either Heads up or Tails up:

$$w \models E\Box(\mathsf{H}\vee\mathsf{T})$$

since the set of agents is $\mathcal{A}=\{a,b,c\}$ and

$$w \models \Box_a(\mathsf{H}\vee\mathsf{T})\wedge\Box_b(\mathsf{H}\vee\mathsf{T})\Box_c(\mathsf{H}\vee\mathsf{T}).$$

Knowledge about Knowledge

Moreover, by using concatenations of a-arrows and b-arrows, we can express a**'s knowledge about** b**'s knowledge**,

e .g. **every agent knows that the other agents don't know the upper side of the coin, but that they know it's either Heads or tails:**

$$w \models \Box_b(\neg\Box_a \mathsf{H} \wedge \neg\Box_a \mathsf{T}) \wedge \Box_b\Box_a(\mathsf{H} \vee \mathsf{T})$$

We can also express 3,4,5 levels etc, e.g. *a's knowledge about b's knowledge about c's knowledge about a's (lack of) knowledge*:

$$w \models \Box_a\Box_b\Box_c(\neg\Box_a \mathsf{H} \wedge \neg\Box_a \mathsf{T} \wedge \Box_a(\mathsf{H} \vee \mathsf{T}))$$

In fact, we'll see that in this model the sentence

$$\neg\Box_a \mathsf{H} \wedge \neg\Box_a \mathsf{T} \wedge \Box_a(\mathsf{H} \vee \mathsf{T})$$

is **common knowledge**!

"Common" Modalities

The **language of epistemic/doxastic logic with common knowledge** is obtained by extending the above language of epistemic/doxastic logic with a common knowledge modality

$$C\Box\varphi$$

that quantifies over worlds accessible by *concatenations of any number of arrows*:

$$w \models_{\mathbf{M}} C\Box\varphi \text{ iff } w' \models_{\mathbf{M}} \varphi \text{ for every finite chain (of any legth } n \geq 0)$$

$$\text{of the form } w = w_0 R_{a_1} w_1 R_{a_2} w_2 \cdots R_{a_n} w_n = w'.$$

This includes chains of length 0, i.e. φ has to also be true in the real world w.

This is equivalent to the infinite conjunction "φ *is true*, and *everybody knows/believes* φ*, and everybody knows/believes that everybody knows/believes* φ*, ... etc*":

$$w \models_{\mathbf{M}} C\Box\varphi \text{ iff } w \models \varphi, E\Box\varphi, E\Box(E\Box\varphi), \ldots$$

So $C\Box\varphi$ may be interpreted as **common knowledge** (in which case we use the notation $Ck\varphi$ instead) or **common true belief** (in which case we use $Cb\varphi$ instead), depending on the context.

Examples

Check that in the model above (from Scenario 1) we have that:

$$w \models C\Box\left(\neg\Box_a \mathsf{H} \wedge \neg\Box_a \mathsf{T} \wedge \Box_a(\mathsf{H} \vee \mathsf{T})\right)$$

and similarly for all other agents.

In fact, we have more: *in this particular model, all knowledge is common knowledge*. For every sentence φ and every state s **in THIS model**:

$$s \models \Box_a\varphi \iff s \models \Box_b\varphi \iff s \models \Box_c\varphi \iff s \models C\Box\varphi$$

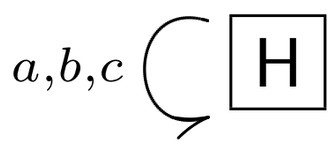
As we'll see, **this is NOT the case in other models**!

Scenario 2: The coin revealed

The referee c opens his palm and shows the face of the coin to everybody (to the public, composed of a and b, but also to himself): they **all see** it's Heads up, and **they all see that the others see it** etc.

What is the epistemic model representing the agents' knowledge after this action?

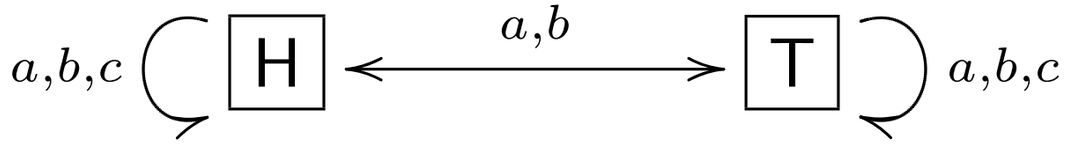
Intuitively, **after the announcement, we have common knowledge of** H, so the model of the new situation is:



EXERCISE: Describe this model formally and check that $C\Box\mathsf{H}$ is true.

Not all Knowledge is common knowledge

In the model



NOT all knowledge is common knowledge.

In particular, agent c **knows something that is not common knowledge**: if we denote by w the real world (on the left, in which the tail is Heads up), we have

$$w \models \Box_c \mathsf{H} \wedge \neg \mathcal{C} \Box \mathsf{H}$$

However, *in this model "everybody knows" is equivalent to "common knowledge"*. **WHY**?

(Truthful) Public Announcements

The action of *showing the face of the coin* (so that everybody can see it, and everybody can see that the others see it) is an example of a **truthful public announcement**.

More generally, suppose that **a sentence φ is publicly announced** by some speaker who **never lies**. We denote this action as $!\varphi$ and we call it *truthful public announcement*.

The *effect* of this action is to **change the model** in the following way:

all the worlds in which φ is false are deleted, and *everything else is left the same.*

Formal Definition: updated model after a public announcement

To define this formally, suppose we are given a model $\mathbf{M} = (W, R_1, \ldots, R_n, \nu)$, and let φ be any sentence.

The result of the action $!\varphi$ performed on the model $\mathbf{M}$ is a *new* ("updated") model

$$\mathbf{M}^{!\varphi} = (W', R'_1, \ldots, R'_n, \nu'),$$

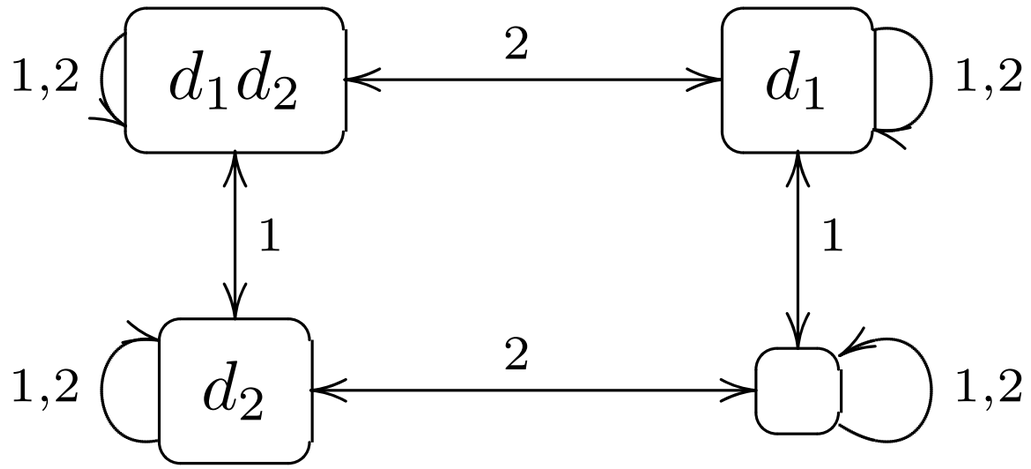
given by

$$\begin{aligned} W' &:= \|\varphi\|_{\mathbf{M}} = \{w \in W : w \models_{\mathbf{M}} \varphi\}, \\ wR'_a s &\quad \text{iff} \quad wR_a s \text{ (for } w, s \in W'), \\ \nu'(w) &:= \nu(w) \text{ (for all surviving worlds } w \in W'). \end{aligned}$$

$\mathbf{M}^{!\varphi}$ is the model that represents the knowledge situation **after** the public announcement.

EXAMPLE 1: Two Muddy Children

Two children 1 and 2 played in the mud and got dirty on their foreheads. Each of them can see the other's forehead, but not his own.



Formally, this model has four possible worlds:

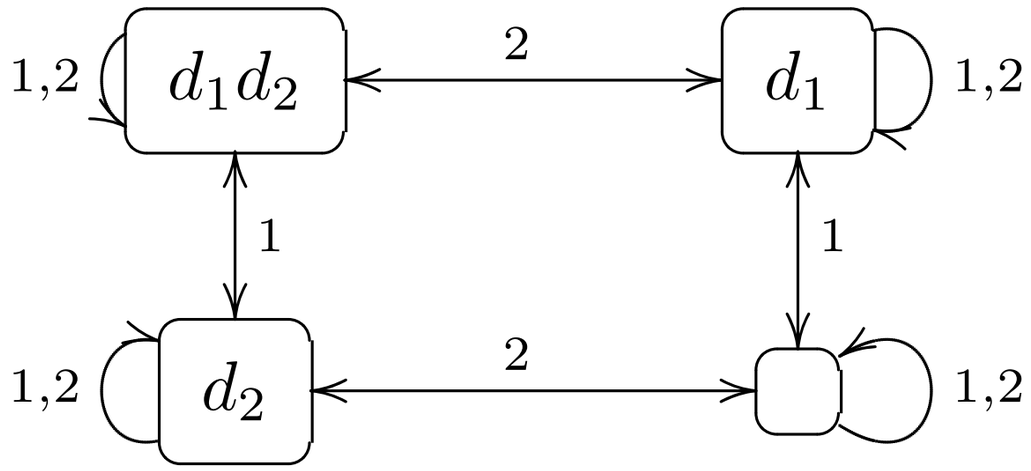
$$W = \{w, v, u, s\}, \;\; \mathcal{A} = \{1, 2\}, \;\; Prop = \{d_1, d_2\}$$

$$\nu(w) = \{d_1, d_2\}, \; \nu(v) = \{d_1\}, \; \nu(u) = \{d_2\}, \; \nu(s) = \emptyset,$$

$$R_1 = \{(w,w), (w,u), (v,v), (v,s), (u,u), (u,w), (s,s), (s,v)\},$$

$$R_2 = \{(w,w), (w,v), (v,v), (v,w), (u,u), (u,s), (s,s), (s,u)\}.$$

In this model



in the real world w (in the upper left corner), *each child knows that the other one is dirty but not that he himself is dirty*:

$$w \models \Box_1 d_2 \wedge \Box_2 d_1 \wedge \neg\Box_1 d_1 \wedge \neg\Box_2 d_2$$

Also, *everybody knows that at least one child is dirty*:

$$w \models E\Box(d_1 \vee d_2),$$

but *it is NOT common knowledge that at least one child is dirty*:

$$w \models \neg C\Box(d_1 \vee d_2).$$

EXAMPLE CONTINUED: a public announcement

Suppose a public announcement is made (by somebody who never lies, like Father) that "at least one of the two children is dirty".

What is the model after this? Well, the sentence that is being announced is

$$d_1 \vee d_2$$

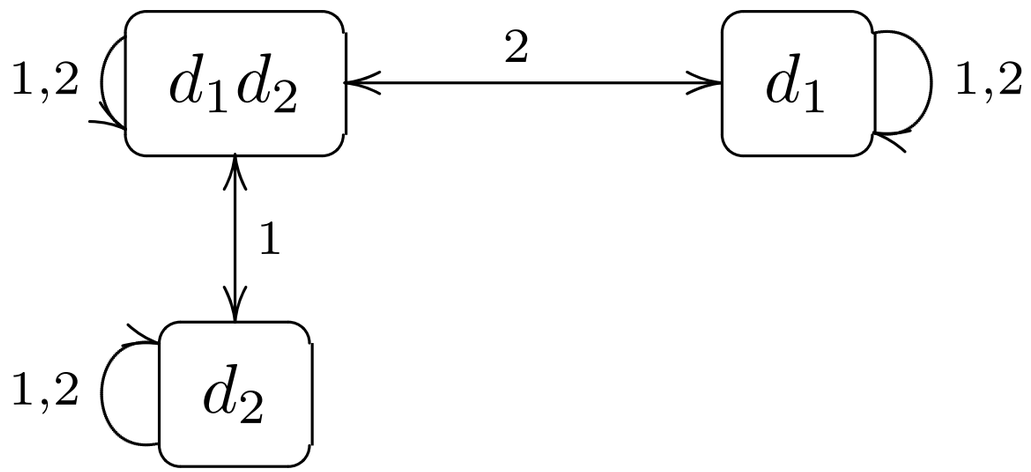
To computed the updated model, we have to check in which worlds is this sentence is true in the old model:

obviously, $d_1 \vee d_2$ is **true** in the worlds

- w, with valuation $\nu(w) = \{d_1, d_2\}$,
- v, with $\nu(v) = \{d_1\}$,
- u, with $\nu(u) = \{d_2\}$,

and it is **false only in world** s (with $\nu(s) = \emptyset$).

Model after Father's Announcement



(I left Father out of the picture, so we still have only two agents $\mathcal{A} = \{1, 2\}$.)

EXERCISE: Check that NOW (in **this** model) *it IS common knowledge that at least one is dirty*: if w is the world in upper left corner then

$$w \models C\Box(d_1 \vee d_2).$$

EXAMPLE CONTINUED: Father's question

Suppose now Father asks all the children:

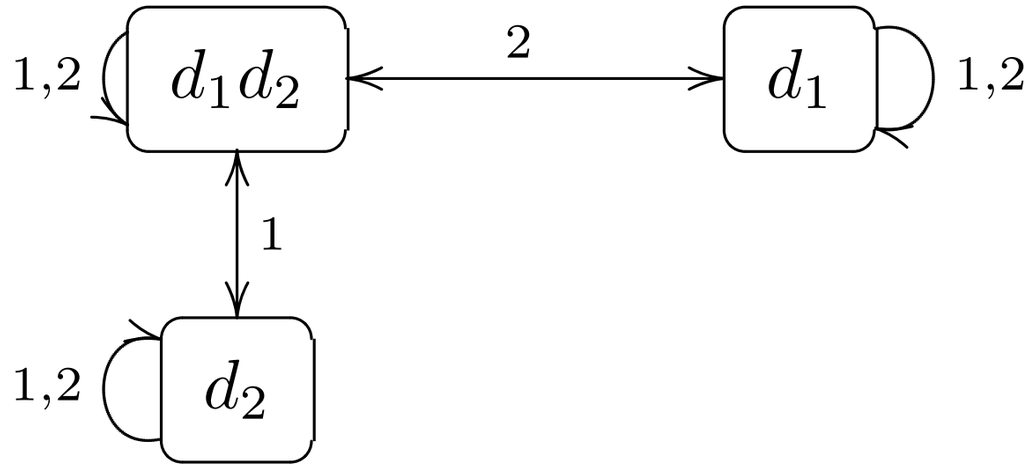
"*do you know if you are dirty or not, and if so, which of the two?*"

Children are required to answer simultaneously, truthfully and publicly (so that instantly the answers become common knowledge).

What will they answer?

Well, we assumed that both children were in fact, so the real world was w.

In our updated model



we have

$$w \models \neg K_1 d_1 \wedge \neg K_1 \neg d_1 \wedge \neg K_2 d_2 \wedge \neg K_2 \neg d_2$$

(**why**??).

So the children will all answer "*I don't know*".

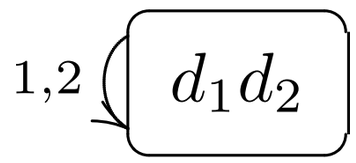
Example continued: another public announcement

We can interpret the children's simultaneous answering (in this first round) as a second public announcement

$$!\,(\neg K_1 d_1 \wedge \neg K_1 \neg d_1 \wedge \neg K_2 d_2 \wedge \neg K_2 \neg d_2)$$

of the fact that *none of them knows whether they are dirty or not.*

But the effect of this new public announcement is to change the model once again:



EXERCISE: Use this newly updated model to show formally that, if Father asks the same question again, both children will answer "*I know I am dirty*".

Validity on a family of models

Given a family $\mathcal{F}$ of Kripke models and a sentence φ in our language, we say that φ is **valid on** $\mathcal{F}$ if φ is true at every world of every model in $\mathcal{F}$.

EXAMPLE: For every two sentence φ and ψ, the sentence

$$\Box_a(\varphi \Rightarrow \psi) \Rightarrow (\Box_a\varphi \Rightarrow \Box_a\psi)$$

is **valid on the family of ALL models** (having agent $a \in \mathcal{A}$).

PROOF

To prove this:

Let $\mathbf{M} = (W, R_a, \ldots, \nu)$ be any model, and let $w \in W$ be any world in it.

To prove our claim, suppose that $\Box_a(\varphi \Rightarrow \psi)$ is true at w, i.e.

$$(1) \qquad w \models_{\mathbf{M}} \Box_a(\varphi \Rightarrow \psi).$$

To show that $\Box_a\varphi \Rightarrow \Box_a\psi$ is also true at w, assume that we have

$$(2) \qquad w \models_{\mathbf{M}} \Box_a\varphi,$$

and we need to **prove** that

$$(?) \qquad w \models_{\mathbf{M}} \Box_a\psi.$$

Let v by an **arbitrary world such that** wR_av.

By the semantics of $\Box_a$, (1) implies

$$(1') \qquad v \models_{\mathbf{M}} \varphi \Rightarrow \psi.$$

and similarly, (2) implies

$$(2') \qquad v \models_{\mathbf{M}} \varphi.$$

Putting (1') and (2') together, and using the properties of implication (the Modus Ponens Rule), it follows that:

$$v \models_{\mathbf{M}} \psi.$$

But this conclusion holds for **EVERY arbitrary world such that** wR_av, i.e. we have

$$w \models_{\mathbf{M}} \Box_a\psi.$$

DONE!

EXERCISE

The following are valid on (the family of) **all epistemic models**:

1. **Veracity of Knowledge:**

$$K_a\varphi \Rightarrow \varphi$$

2. **Positive Introspection of Knowledge:**

$$K_a\varphi \Rightarrow K_aK_a\varphi$$

3. **Negative Introspection of Knowledge:**

$$\neg K_a\varphi \Rightarrow K_a\neg K_a\varphi$$

Validities involving common knowledge

The following are valid on the family of all epistemic models:

1. (**"Fixed-Point Axiom"**, also known as "Mix"):

$$Ck\varphi \;\Leftrightarrow\; (\varphi \wedge K_1 Ck\varphi \wedge \ldots \wedge K_n Ck\varphi)$$

   In words:

   "φ **is common knowledge if and only if: (1)** φ **is true, and (2) everybody knows that** φ **is common knowledge**".

2. (**"Induction Axiom"**):

$$Ck(\varphi \Rightarrow K_1\varphi \wedge \ldots \wedge K_n\varphi) \Rightarrow (\varphi \Rightarrow Ck\varphi)$$

   In words: "**Suppose that we have (1) it is common knowledge that (if** φ **is true than everybody knows it). Then, if (2)** φ **is true, then (3)** φ **is common knowledge.**"

Proofs???

WARM-UP EXERCISE FOR YOUR WERKCOLLEGE!

Lecture 2.1. Public Announcements (Continued)

We continue the investigation of various forms of knowledge and of knowledge-updates.

Distributed Knowledge within a Group

The **distributed knowledge** modality $D\varphi$ is obtained by quantifying over all worlds that are **simultaneously accessible** by **all** arrows (from a given world):

$w \models_{\mathbf{S}} D\varphi$ iff $w' \models_{\mathbf{S}} \varphi$ for all w' such that $(w, w') \in R_1 \cap R_2 \cap \ldots R_n$.

So D is the Kripke modality corresponding to the **intersection of all epistemic relations** $\bigcap_{a \in \mathcal{A}} \sim_a$.

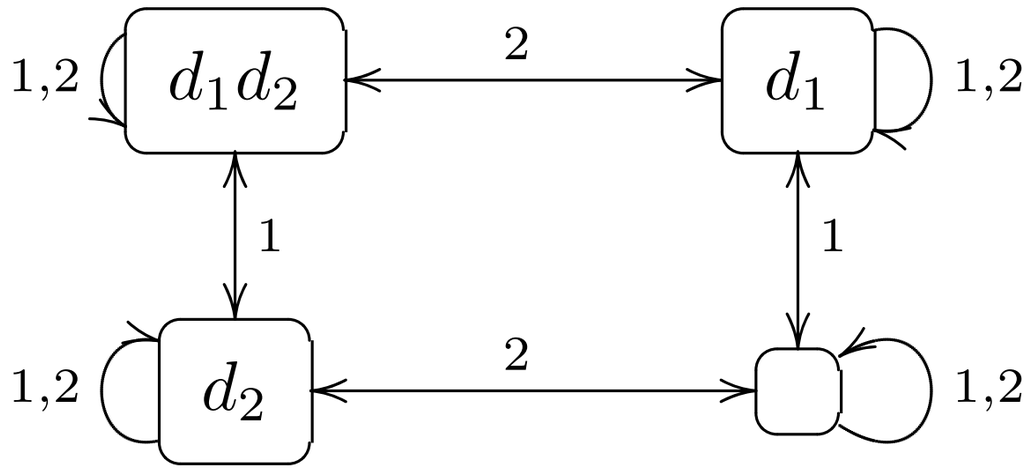
BUT intuitively distributed knowledge corresponds to ***pulling together*** *in one knowledge base all the knowledge possessed by* ***any*** *of the agents* and *closing under logical entailment* (i.e. adding all the sentences that can be proved by applying modus ponens to our knowledge base).

"Pulling together" means **union** of all the knowledge bases (of all agents).

QUESTION: **Why does the above definition use intersection (rather than union)**?

Example

Recall EXAMPLE 1: Two children 1 and 2 played in the mud and got dirty on their foreheads. Each of them can see the other's forehead, but not his own.



Intuitively, the fact that **both of them are dirty** is *distributed knowledge*, though this fact is *not known to any of them.*

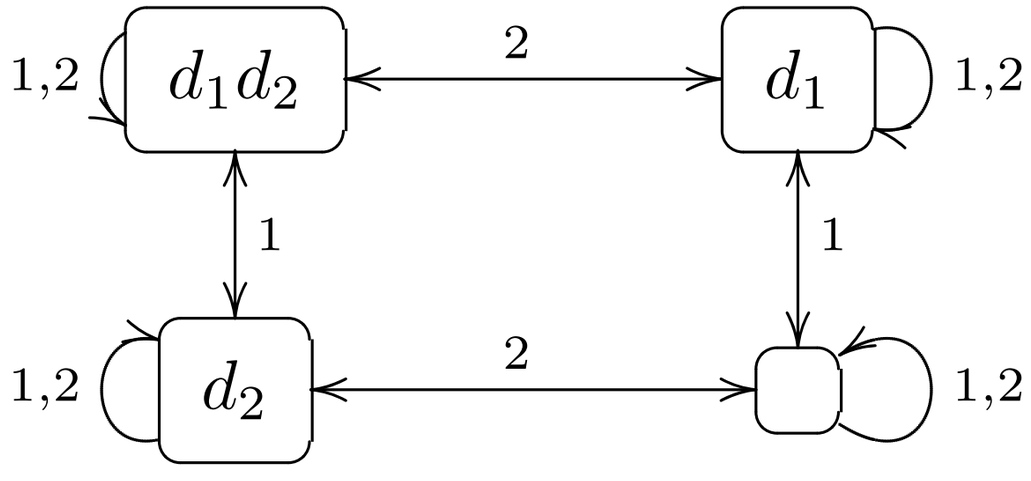
Check that indeed, in the real world $w = (d_1d_2)$, we have:

$$w \models \neg K_1(d_1 \wedge d_2) \wedge \neg K_2(d_1 \wedge d_2) \wedge D(d_1 \wedge d_2)$$

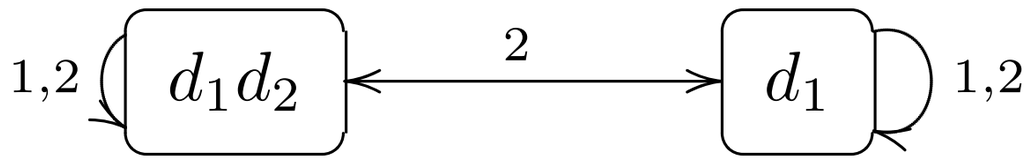
Distributed Knowledge can become (Common) Knowledge

Intuitively, **distributed knowledge can be converted into (common) knowledge by communication within the group**:

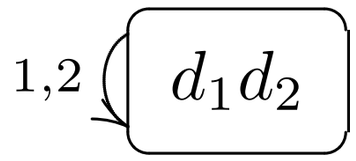
after each agent communicates to all others "*everything (s)he knows*", distributed knowledge will become common knowledge!



child 2 knows that 1 is dirty ($w \models K_2 d_1$) and child 1 knows that 2 is dirty ($w \models K_1 d_2$). To pull together their distributed knowledge $D(d_1 \wedge d_2)$, it's enough for them to communicate: after updating with $!d_1$ the model becomes



and after a second update with $!d_2$ the model becomes



NOW $d_1 \wedge d_2$ became **common knowledge**!

Validities

The following are valid on all epistemic models:

$$K_a\varphi \Rightarrow D\varphi,$$

$$(K_a\varphi \wedge K_b\psi) \Rightarrow D(\varphi \wedge \psi),$$

for all agents $a, b \in \mathcal{A}$.

Knowledge of a Number

In cryptography, agents may possess some **secrets**, which are usually **natural numbers** (e.g. passwords, secret encryption/decryption keys etc.) One can then ask **whether or not an agent** a **knows (the value of) a secret number** n.

To formalize this, we introduce a given (finite) set Var of **numerical variables**, denoted by

$$n, m, \ldots$$

or

$$n_1, n_2, \ldots$$

or (when we one to stress that they are associated with some agents)

$$n_a, n_b, \ldots \quad \text{(where } a, b, \ldots \text{ are agent names)}.$$

In each world w of the model, each variable n will take a specific

natural number $n(w) \in N$ as its value (but the values may differ from world to world).

The Language of Numerical Epistemic Logic

The language of Numerical Epistemic Logic is simply obtained by adding to epistemic logic two new components:

- new atomic sentences of the form

$$n = i \quad \text{(one for each variable } n \in Var \text{ and each number } i \in N\text{)}$$

  These are called *numerical assignments*, e.g. $n_1 = 5$, $m = 3$ etc.

- a *numerical-knowledge operator*, taking any variable $n \in Var$ and any agent $a \in \mathcal{A}$ into a new sentence

$$K_a n \quad \text{(agent } a \text{ knows the value of } n\text{)}$$

Using arithmetical operations on natural numbers, can also add later *more complex numerical sentences*, expressing relationships between variables, e.g. $n = m + 1$, $n \geq m$ etc.

We put

$$At := Prop \cup \{n = i | n \in Var, i \in N\}$$

for the set of *all atomic sentences* (both "factual" atomic propositions $p, q, \ldots$ and numerical assignments $n = 1$, $n = 2$, $m = 3$ etc).

Numerical Epistemic Models

For a given set $Prop = \{p, q, \ldots\}$ of "*atomic propositions*", a set $Var = \{n, m, \ldots\}$ of numerical variables, and a finite set $\mathcal{A}$ of *agents*, a **numerical (epistemic) model** is a triple

$$\mathbf{M} \quad = \quad (W, \sim_a, \nu\,)_{a \in \mathcal{A}} \quad \text{, where}$$

1. W is a set of "*possible worlds*" (or "*states*")

2. each $\sim_a \subseteq W \times W$ is an *equivalence* relation (one for each agent $a \in \mathcal{A}$), called "*accessibility relations*"

3. an *(extended) valuation* map $\nu : W \to \mathcal{P}(At)$, called *valuation*, that assigns to each world $w \in W$ some set $\nu(w) \subseteq At$ of atomic propositions, satisfying

$$\forall n \in Var \forall w \in W \exists! i \in N : \, (n = i) \in \nu(w).$$

So *every variable takes exactly one numerical value in each world.*

This means that our valuation induces now a **(numerical) value map**, taking each numerical variable $n \in Var$ and each world $w \in W$ into a natural number $n(w) \in N$, given by:

$$n(w) := \text{ the unique number } i \in N \text{ such that } (n = i) \in \nu(w) \ .$$

Truth (Satisfaction) Relation

We can extend the **satisfaction relation** to cover the new kind of sentences as well:

$$w \models_{\mathbf{M}} (n = i) \quad \text{iff} \quad (n = i) \in \nu(w) \ , \text{ i.e. } n(w) = i$$

$$w \models_{\mathbf{M}} K_a n \quad \text{iff} \quad \forall w' \in W : \ (w \sim_a w' \Rightarrow n(w) = n(w'))$$

The last clause is the important addition: it gives us the *meaning of "knowing a number"*.

The semantics can be further extended in the obvious way to more complex numerical sentences ($n = m + 1$ etc).

As before, we encode the same information into a *truth map*, giving for each sentence φ some set of worlds

$$\|\varphi\| := \{w \in A | w \models_{\mathbf{M}} \varphi\}.$$

Further Extension: Conditional Knowledge of a Number

In fact, it is useful to *extend* the language further with a *conditional numerical-knowledge operator*

$$K_a^{\varphi} n \quad (\text{ where } \varphi \text{ is another sentence}).$$

We read this as "a knows n given φ". It means that *agent a would know the value of n IF given information φ.*

The **semantics** is extended by putting:

$$w \models_{\mathbf{M}} K_a^{\varphi} n \quad \text{iff} \quad \forall w' \in W : \ (w \sim_a w' \wedge w' \models_{\mathbf{M}} \varphi \ \Rightarrow \ n(w) = n(w'))$$

Obviously, *simple knowledge of a number is a special case of conditional knowledge*:

$$K_a n \Leftrightarrow K_a^{\top} n \quad (\text{where } \top \text{ is any tautology}).$$

An Example

Alice and Bob are each given a card with **some natural number** (from the set $\{0, 1, 2, 3, \ldots\}$):

say n_a is Alice's number and n_b is Bob's number.

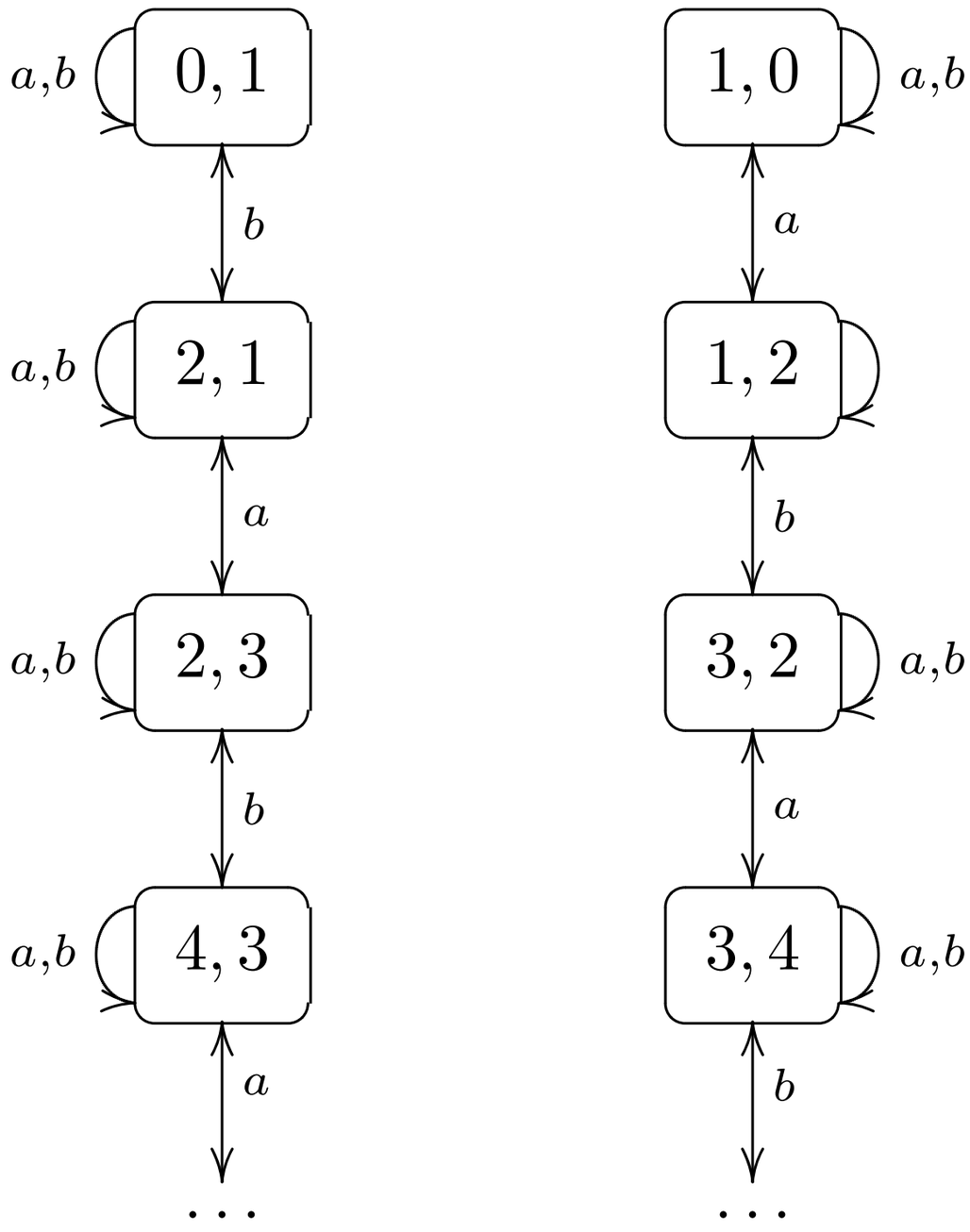
Each of them can only see his/her own number, but not the others' number.

However, they are told (via some truthful public announcement) that **one of the two numbers is the immediate successor of the other**:

so **it is common knowledge that** $n_a = n_b + 1 \vee n_b = n_a + 1$.

MODELLING: we can represent each possible world as a pair of numbers.

The Model is Infinite!



Atomic Sentences and Valuation

We have **two variables** n_a and n_b. As **atomic sentences**, we can consider sentences of the form

$n_a = 0, n_a = 1, n_a = 2, \ldots$ (saying that Alice's number is 0, or 1, etc),

$n_b = 0, n_b = 1, n_a = 2, \ldots$ (saying that Bob's number is 0, or 1, etc),

or even more complex sentences $n_a = f(n_a, n_b)$ or $n_b = f(n_a, n_b)$ (where f is some arithmetical function).

E.g. $n_a = n_b + 1$.

The **valuation** ν is the obvious one:

e.g. $(n_b = n_a + 1), (n_a = 3), (n_b = 4) \in \nu(3, 4)$ etc.

Facts

Suppose that in reality $n_a = 3, n_b = 4$, so the real world is $w = (3, 4)$.

Check that, in **every world (of this model) both agents know their own numbers**, i.e.

$$K_a n_a \wedge K_b n_b \quad \text{is true in all worlds ,}$$

but **in the real world** $(3, 4)$ **none of them knows the other's number**:

$$(3, 4) \models \neg K_a n_b \wedge \neg K_b n_a.$$

However, in this world *agent a knows b's number* n_b *IF GIVEN the information that it is smaller than her own number*:

$$(3, 4) \models K_a^{n_a < n_b} n_b.$$

QUESTION: Are there any worlds in which $K_a n_b$ is true?

Similar question for $K_b n_a$.

Note that **the real world is distributed knowledge though it is not known to any of the two agents**:

$$(3,4) \models D(n_a = 3 \wedge n_b = 4) \wedge \neg K_a(n_a = 3 \wedge n_b = 4) \wedge \neg K_b(n_a = 3 \wedge n_b = 4)$$

Some things are common knowledge, e.g.:

$$(3,4) \models Ck(n_a = n_b + 1 \vee n_b = n_a + 1)$$

$$(3,4) \models Ck\neg(n_a = 0 \wedge n_b = 1)$$

A Validity in Numerical Epistemic Logic

The following is valid on (all) numerical epistemic models:

$$n = i \;\Rightarrow\; (\, K_a n \Leftrightarrow K_a(n = i)\,)$$

"If a variable has value i, then knowing its value is the same as knowing that it has value i. "

The Story Continues

Suppose that, in the Numbers Example above, Alice and Bob are asked (repeatedly): "*Do you know both numbers (yours and the other child's)*?"

Of course, each knows his/her number, so the question is really whether or not they know the *other child's* number.

The first time they both answer (truthfully and simultaneously, no cheating!): "*No, I don't know the other's number*".

When the question is repeated the second time, they both again answer: "*I don't know the other's number*".

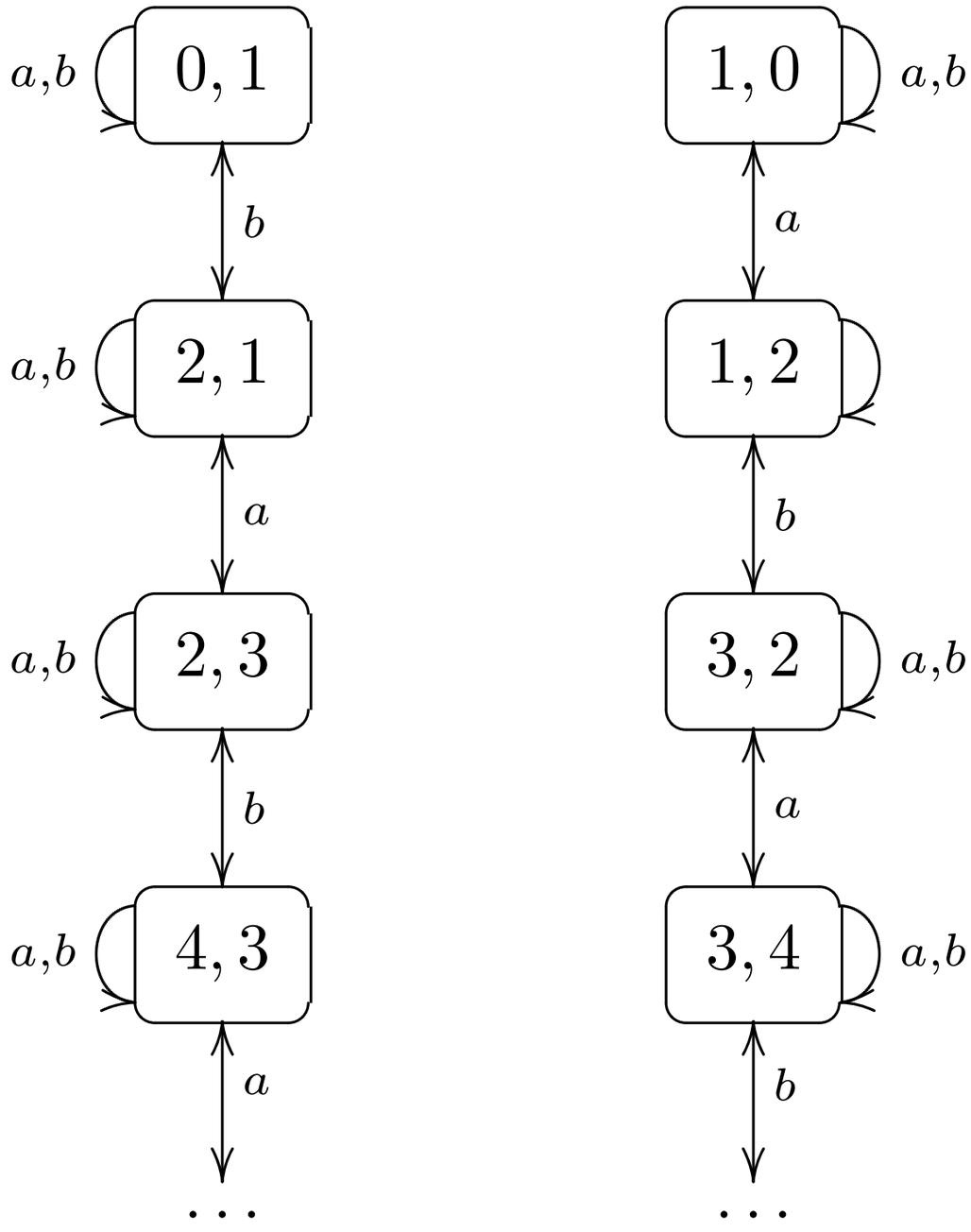
When the question is repeated the third time, Alice answers "*Yes, I know both numbers*", while Bob still answers "*I don't know Alice's number*".

QUESTION 1: **What are the two numbers?**

QUESTION 2: **What will Bob answer if the question is repeated once again?**

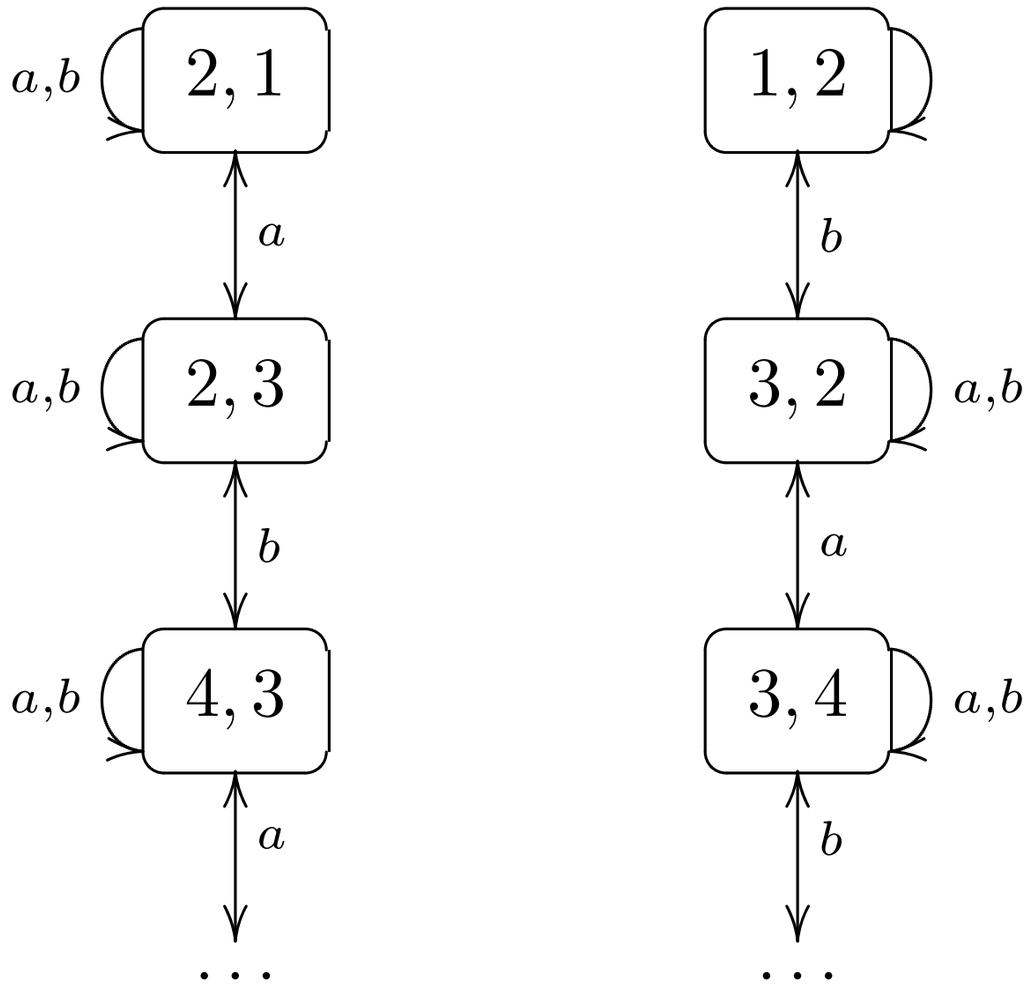
EXERCISE (in class): Solve this problem using public announcements (as updates on epistemic models).

RECALL THE MODEL



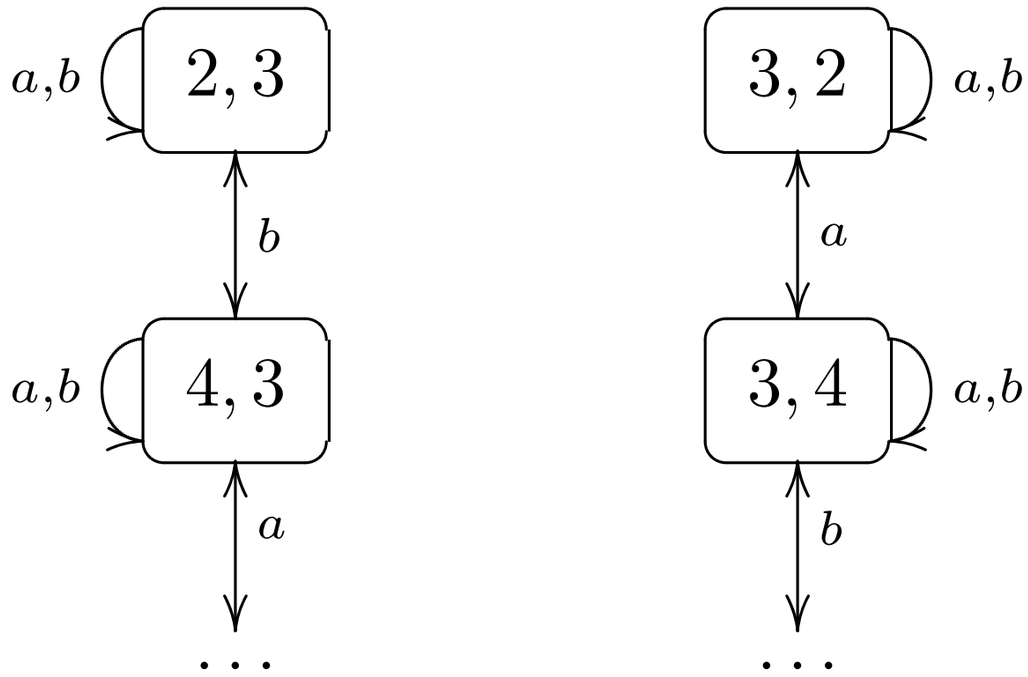
MODEL AFTER FIRST ANNOUNCEMENT

The first announcement $!(\neg K_a n_b \wedge \neg K_b n_a)$ deletes $(1,0)$ and $(0,1)$:



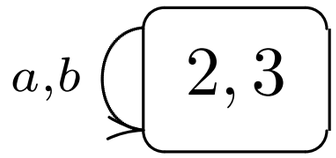
MODEL AFTER SECOND ANNOUNCEMENT

The second announcement $!(\neg K_a n_b \wedge \neg K_b n_a)$ deletes $(2,1)$ and $(1,2)$:



MODEL AFTER THIRD ANNOUNCEMENT

The third announcement $!(K_a n_b \wedge \neg K_b n_a)$ deletes all worlds, EXCEPT $(2, 3)$:



So $n_a = 2$, $n_b = 3$.

Since there is only one world is left after this, **now Bob knows Alice's number** as well: $K_b n_a$.

So, if I asked again, he'll answer "*I know my number*".

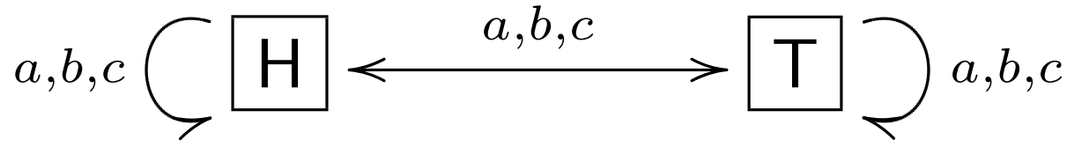
Private Announcements

A **logic for "secret (fully private) announcements"** was first proposed by Gerbrandy (1999).

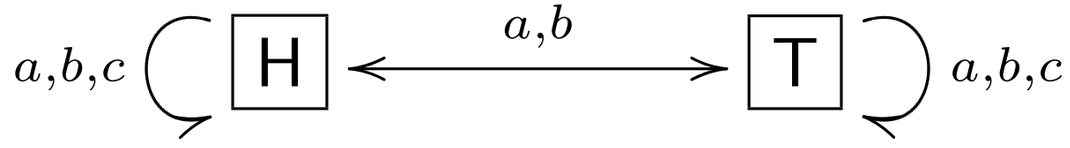
A **logic for "private, but legal, announcements"** (what we will call "*fair-game announcements*") was developed by H. van Ditmarsch (2000).

Scenario 3: 'Legal' Private Viewing

Starts as in our original Scenario 1 (the Concealed Coin), with the model



where nobody knew the coin. In front of everybody, the referee (c) uncovers the coin, so that (they all see that) **he, and ONLY HE, can see the upper face**. This changes the initial model to



Now, c **knows** the real state. E.g. if it's Heads, he knows it, and disregards the possibility of Tails. a and b don't know the real state, but *they know that c knows it.* c's viewing of coin is a "legal", non-deceitful action, although a private one.

Fair-Game Announcements

Equivalently: in front of everybody, an announcement of the upper face of the coin is made, but in such a way that (it is common knowledge that) only c hears it.

This is NOT a public announcement!

Such announcements (first modeled by H. van Ditmarsch) are called **fair-game announcements**, they can be thought of as "legal moves" in a fair game: nobody is cheating, all players are aware of the possibility of this move, but only some of the players (usually the one who makes the move) can see the actual move. The others know the range of possible moves at that moment, and they know that the "insider" knows his move, but they don't necessarily know the move.

Scenario 4: Cheating by secretly "taking a peek"

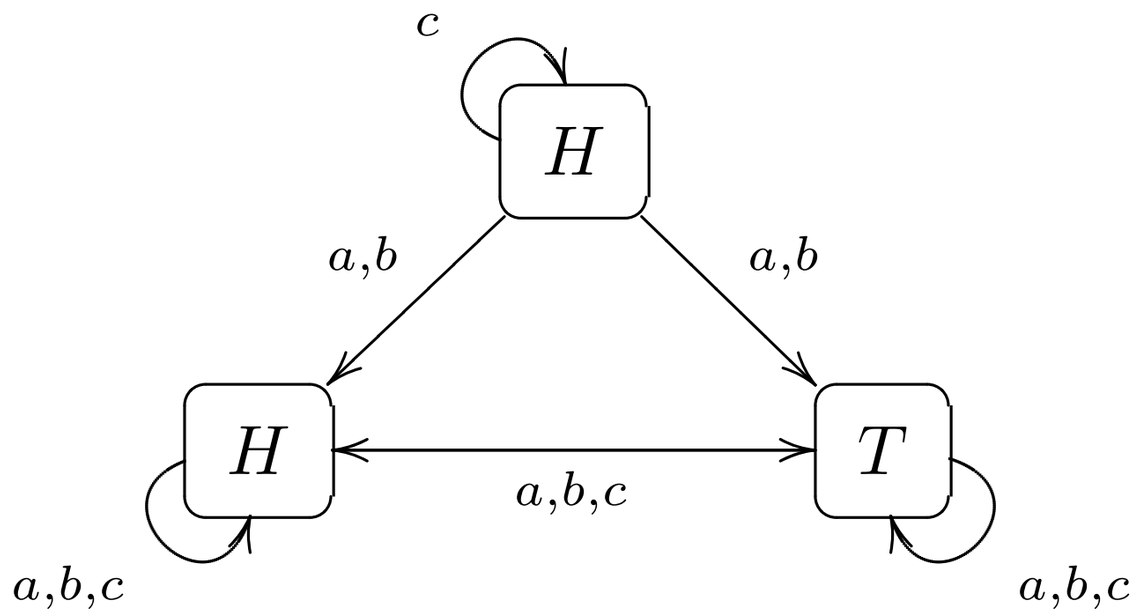
Start again with Scenario 1: the referee c throws a fair coin, catching it in his palm and fully covering it, before anybody (including himself) can see on which side the coin has landed, so that nobody sees the upper face of the coin:

But now let's suppose that the referee c *was cheating*: he has **taken a peek at the coin**, before covering it. **Nobody noticed this**. Indeed, let's assume that c **knows that** a **and** b **did not suspect anything**.

This is an instance of **cheating**: a private viewing which is "illegal", in the sense that it is deceitful for a and b. Now, a and b think that nobody knows on which side the coin is lying. But they are wrong!

What is the model after this "cheating" action?

The Model after Cheating



The *real world* is the top one: a and b think that the only possibilities are the worlds below. That is, they *do not even consider the "real" world as a possibility.*

Note that this is NOT an epistemic model:

not all relations are reflexive, so Veracity does not hold!

Indeed, the sentence $\Box_c \mathsf{H} \wedge \Box_a \neg \Box_c \mathsf{H}$ is true in the top world!

So in this model we CANNOT interpret $\Box$ as knowledge (since knowledge is always true!)

BUT this model is a good model for **belief**.

Something unexpected happened:
as a result, *what used to be (true)* ***knowledge*** *has become (possibly false)* ***belief***!

Doxastic Models

A **doxastic model** (or $KD45$**-model**) is a Φ-Kripke model satisfying the following properties:

- **(D) Seriality**: for every s there exists some t such that sR_at;
- **(4) Transitivity**: If sR_at and tR_aw then sR_aw
- **(5) Euclideaness** : If sR_at and sR_aw then tR_aw

In a doxastic model, $\Box_a$ is interpreted as **belief**, and denoted by B_a.

EXERCISE

The following are valid on (the family of) all doxastic models:

1. **Consistency of Beliefs**:

$$\neg B_a(\varphi \wedge \neg\varphi)$$

2. **Positive Introspection of Beliefs:**

$$B_a\varphi \Rightarrow B_aB_a\varphi$$

3. **Negative Introspection of Beliefs:**

$$\neg B_a\varphi \Rightarrow B_a\neg B_a\varphi$$

Lecture 2.2. Epistemic Actions

We want to generalize public and private announcements, so that we can represent **complex communication scenarios** and other forms of *multi-agent information flow*, that involve **mixtures of private and public features**.

So we want to solve the "Update Problem":

find a **general method to model the information in a multi-agent system after all kinds of updates**: *private and public learning, secret communications* following various communication protocols, *secret wiretapping (interception of messages)* etc.

For this, we need to *generalize* the types of updates we've been studying: deleting worlds, deleting arrows between worlds, adding new worlds with new arrows.

Dynamic Epistemic Logic

- studies the **multi-agent information flow of "hard information"**;
- gives an answer to the Update Problem, based on the BMS (Baltag, Moss and Solecki) approach: **logics of epistemic actions**;
- it arose from *generalizing* previous work on logics for public/private announcements.
- it can model **very complex forms of communication**
- This dynamics has one big limitation. It is **essentially monotonic**: one keeps *adding more information or more beliefs, but never gives up any old beliefs*!
- We will fix this last problem in the second part of the course, by adding "Belief Revision" to our toolbox.

Models for 'Events'

Until now, our Kripke models capture only *epistemic situations*, i.e. they only contain *static* information: they all are *state models*. We can thus represent the *result* of each of our Scenarios, but not what is actually going on. Our scenarios involve various *types of changes* that may affect agents' beliefs or state of knowledge: a public announcement, a 'legal' (non-deceitful) act of private learning, 'illegal' (unsuspected) private learning etc.

We want to use now Kripke models to represent such types of *epistemic events*, in a way that is similar to the representations we have for epistemic states.

Event Models

An **event model** (or "*action model*")

$$\boldsymbol{\Sigma} \quad = \quad (E, R_1, \ldots, R_n, pre)$$

is just like an Kripke model, except that

1. its elements $e \in E$ are now called epistemic **events** (or "*actions*"),
2. the accessibility relations $R_1, \ldots, R_n \subseteq E \times E$ represent each agent's knowledge/beliefs about which event/action that is currently happening,
3. instead of the valuation we have a **precondition map** pre, associating to each event $e \in E$ some sentence pre_e, called the *precondition* of action e.

Epistemic/Doxastic Event Models

As in the case of static models, an event model $\mathbf{\Sigma}$ is called **epistemic** if all the relations $R_1, \ldots, R_n$ are *equivalence relations* (in which case we sometimes denote them by $\sim_i$ instead of R_i).

A **doxastic** event model is one in which all the accessibility relations satisfy the *D45* conditions (i.e. they are *serial, transitive and Euclidean*).

Interpretation

We think of the events $e \in E$ as *deterministic* actions of a particularly simple kind:

they do not change the "facts" of the world, but ONLY the agents' knowledge/beliefs. (E.g. *communication* actions, such as public or private announcements etc.)

(We'll later generalize to other actions.)

For any event $e \in E$, we interpret pre_e as giving the **precondition** of the action e:

this is a sentence that is true in a world w iff e can be executed in world w.

Finally, the accessibility relations express the agents' **knowledge/beliefs about the current action taking place**.

The Product Update

Given a state (Kripke) model $\mathbf{M} = (W, R_1, \ldots, R_n, \nu)$ and an event model $\mathbf{\Sigma} = (E, R_1, \ldots, R_n, pre)$, we define their *update product*

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

to be a new (state) Kripke model, given by:

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

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

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

$$(w, e) R'_a (s, f) \quad \text{iff} \quad w R_a s \text{ and } e R_a f.$$

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

Interpretation

The product arrows encode the idea that: **two output-states are indistinguishable iff they are the result of indistinguishable actions performed on indistinguishable input-states**.

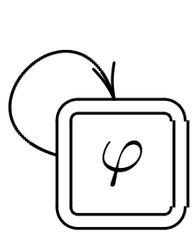
This comprises two intuitions:

1. "*No Miracles*": knowledge can only gained from (the epistemic appearance of) actions;
2. "*Perfect Recall*": once gained, knowledge is never lost.

The fact that the valuation is the same as on the input-state tells us that these actions are **purely informational**: no external facts change, only the agents' information changes.

Examples: Public Announcement

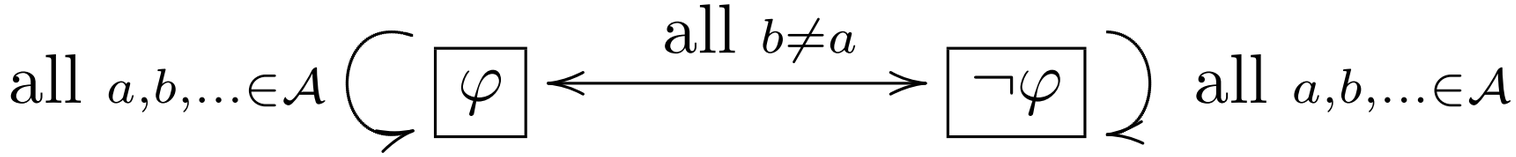
The event model $\mathbf{\Sigma}_{!\varphi}$ for public announcement $!\varphi$ consists of a single action $E = \{e\}$, with precondition $pre_e = \varphi$ and reflexive arrows (loops):



EXERCISE: Check that, for every state model $\mathbf{M}$, $\mathbf{M} \otimes \mathbf{\Sigma}_{!\varphi}$ is indeed the result of deleting all non-φ worlds from $\mathbf{M}$.

More Examples: Fair-Game Announcements

The following event model represents the situation in which *it is common knowledge that an agent a privately learns whether φ or $\neg\varphi$ is the case.* Such actions are called "**fair-game announcements**":



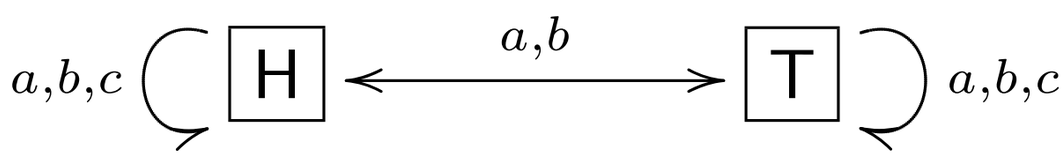
This action on the left is a "fair-game announcement" $Fair_a\varphi$ (by which a privately learns φ), while the action on the right represents the other possibility $Fair_a\neg\varphi$ (by which a privately learns $\neg\varphi$).

The case $\varphi := \mathsf{H}$ (with c instead of a) represents the action in Scenario 3 ("legal viewing" of the card by c).

Example

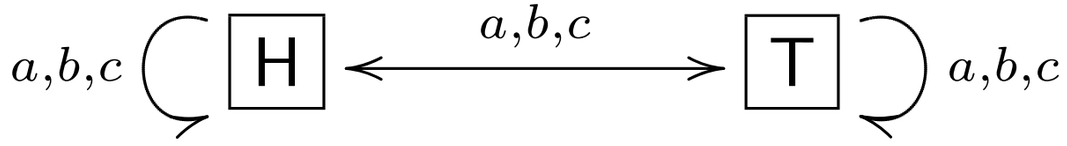
Recall Scenario 3: in front of everybody, the referee (c) uncovers the coin, so that (they all see that) **he, and only he, can see the upper face**.

Here $\mathcal{A} = \{a, b, c\}$, and the action is $Fair_c\mathsf{H}$. The **event model** for this action is:

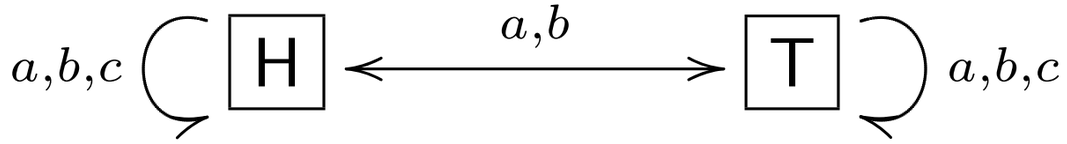


Product Update

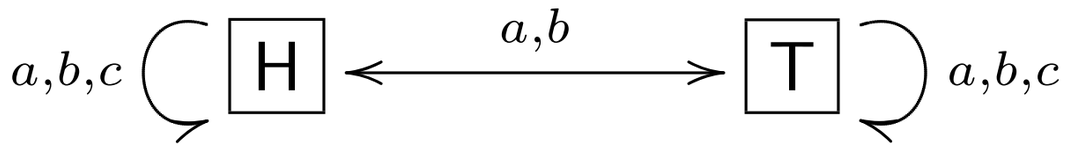
We can now check that the product of the initial state model



and the event model



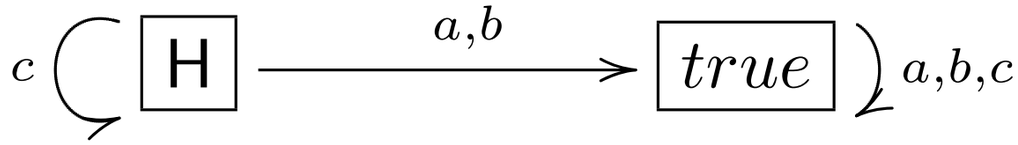
matches the state model we had intuitively used to represent the result of Scenario 3:



More Examples: Taking a Peek

The action in Scenario 4: C takes a peek at the coin and sees the Head is up, without anybody noticing.

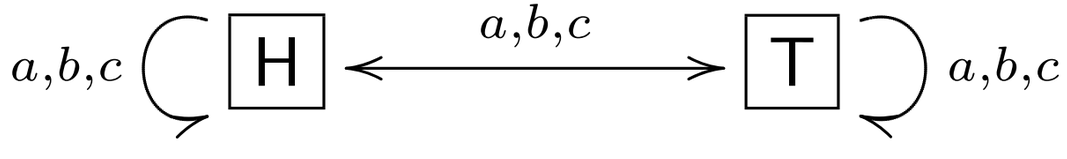
The **event model** for this action is:



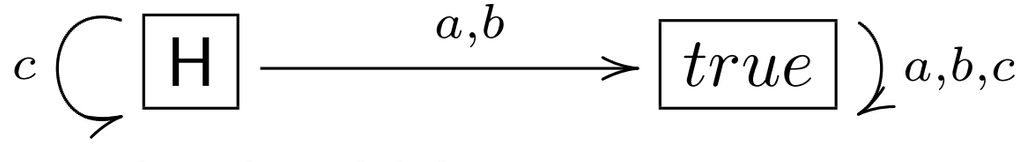
There are two actions in this model: the real event (on the left) is the **cheating action** of c "taking a peek". The action on the right is the apparent action *skip*, having any tautological sentence *true* as its precondition: this is the action in which **nothing happens**. *This is what the outsiders (a and b) think it is going on: nothing, really.*

The Product Update

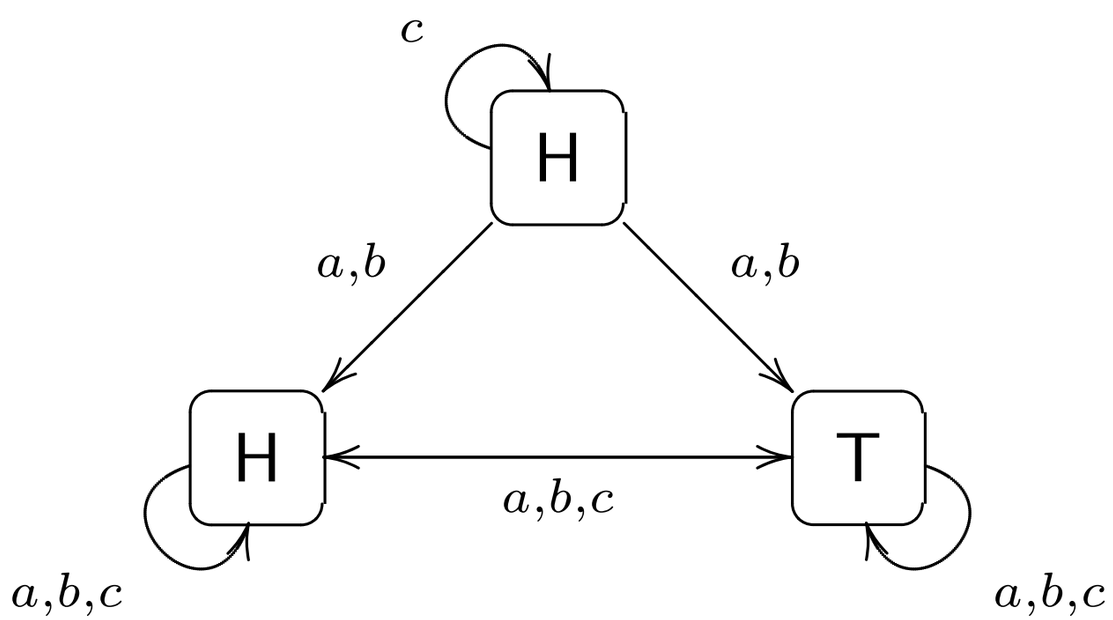
We can now check that the product of



and

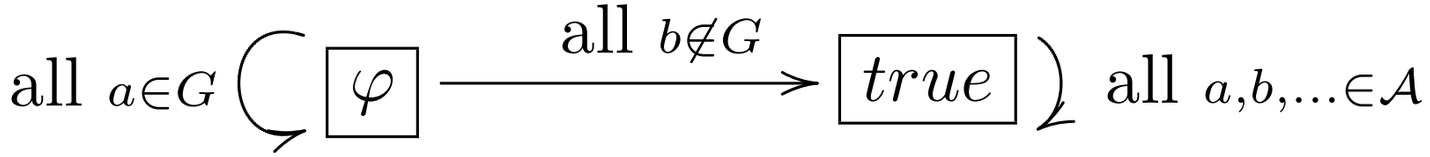


is indeed what intuitively should be:



Fully Private Announcements

More generally, a **fully private announcement** $!_G\varphi$ of φ to a subgroup G is described by *the action on the left* in the following **event model**

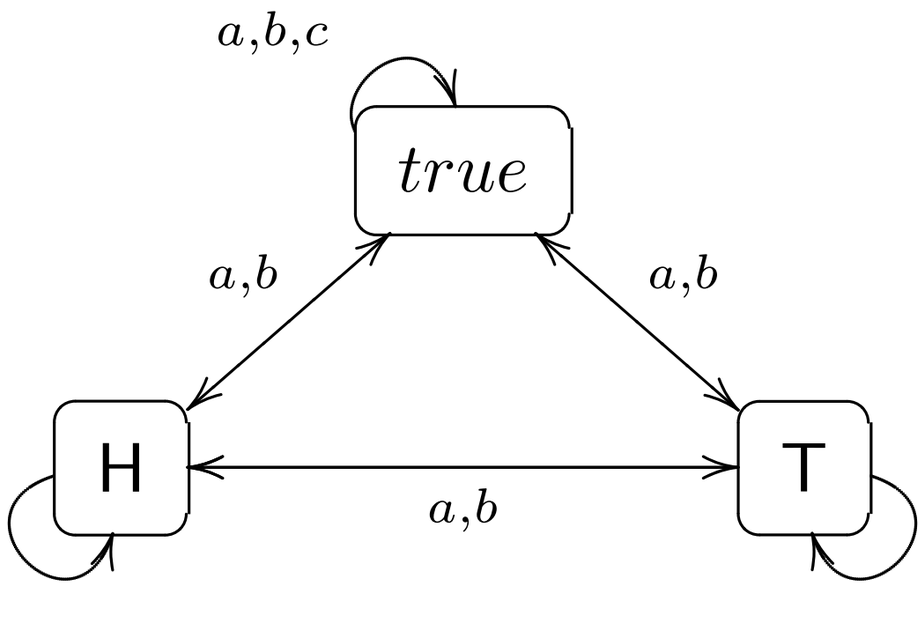


This is a "deceiving" type of action: **the insiders** $a \in G$ **gaining common knowledge that this announcement is made, while outsiders** $b \notin G$ **don't suspect anything**. It subsumes both taking a peek (Scenario 4) and the secret communication in Scenario 5.

More Examples: Scenario 4'

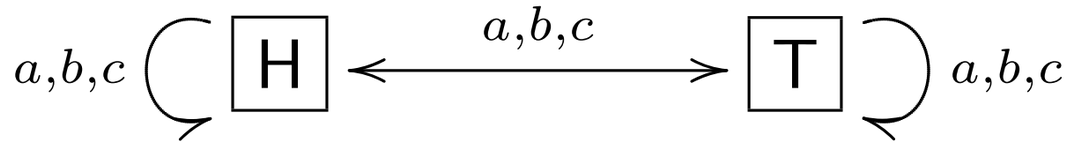
This is exactly as in Scenario 4 (the referee c takes a peek at the coin), except that (it is common knowledge that) when this is happening **the others suspect that he may have taken a peek**. They (a and b) don't know for sure, but *they think it is possible that c may secretly take a peek*. But, even if he did take a peek, they don't know what is the upper face. So they have to consider three possible actions: (1) nothing happens (c doesn't take a peek); (2) c takes a peek and sees the coin lying Heads up; (3) c takes a peek and sees the coin lying Tails up.

The **event model** for this action is

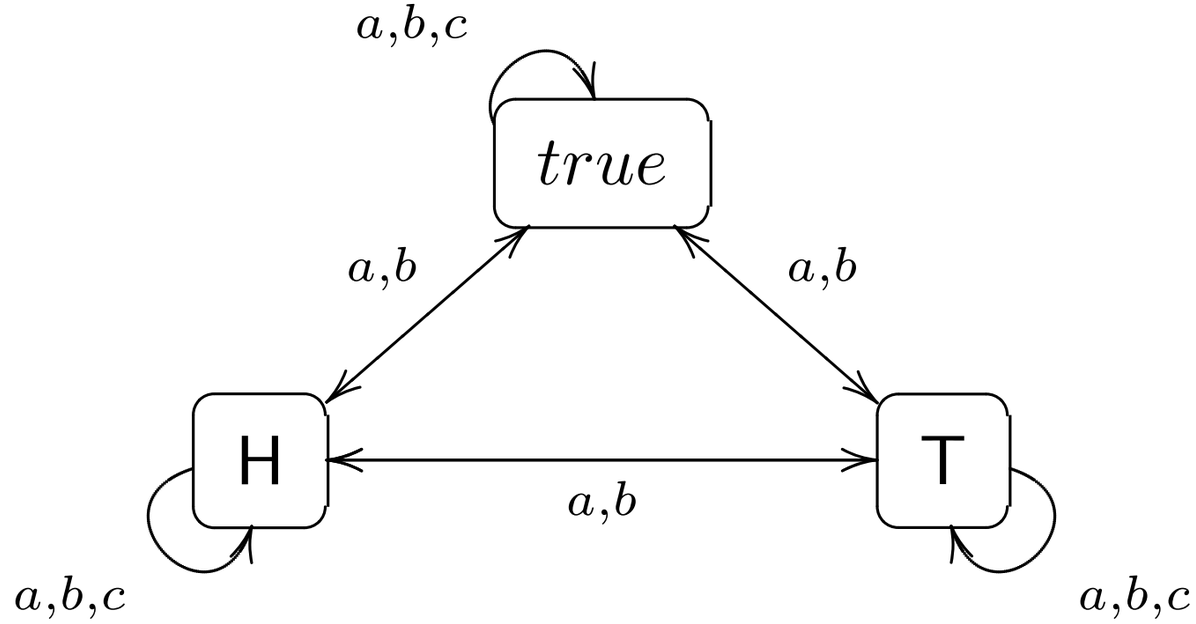


Product Update

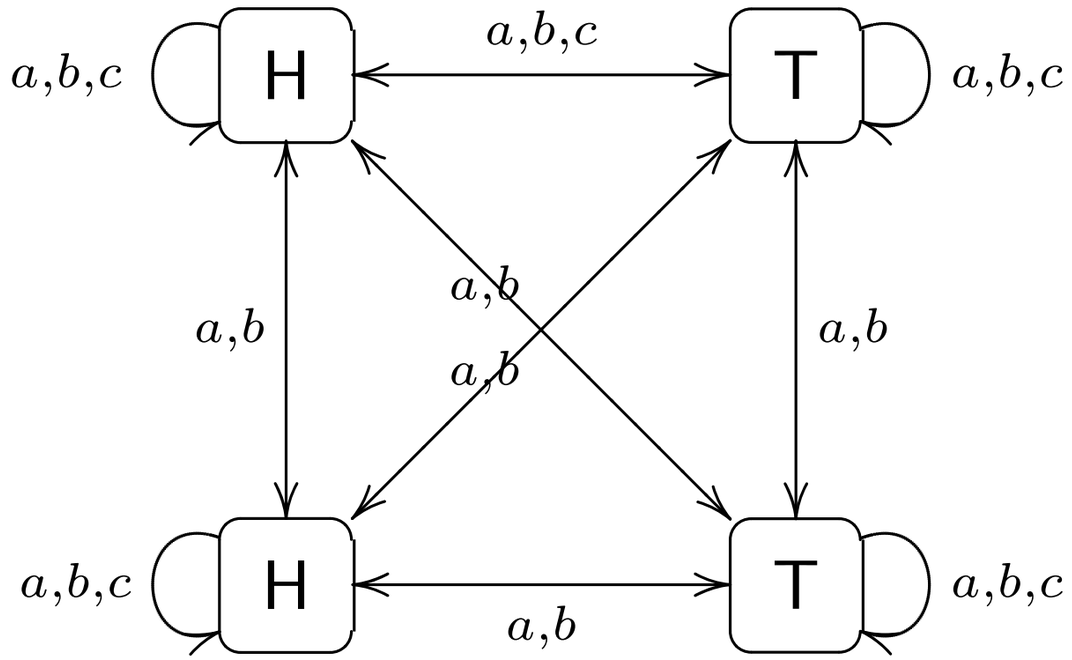
To compute the new epistemic situation after Scenario 4', we take the update product of the original state model



with the event model



obtaining the new state model


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

Scenario 5: Secret Communication

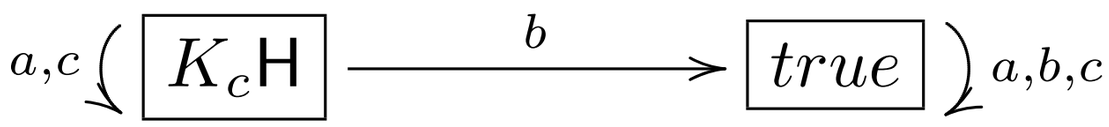
After Scenario 4' happened (the referee c secretly took a peek, with the others suspecting this, but not knowing for sure), c engages in another "illegal" action: **he secretly sends an email to his friend a, informing her that (he knows that) the coin is Heads up**. Suppose the delivery and the secrecy of the message are guaranteed: so a and c have common knowledge that c knew H, and that b doesn't know they know this.

Indeed, b is completely fooled: he doesn't suspect that c could have been engaged in such secret communication.

Example: Scenario 5

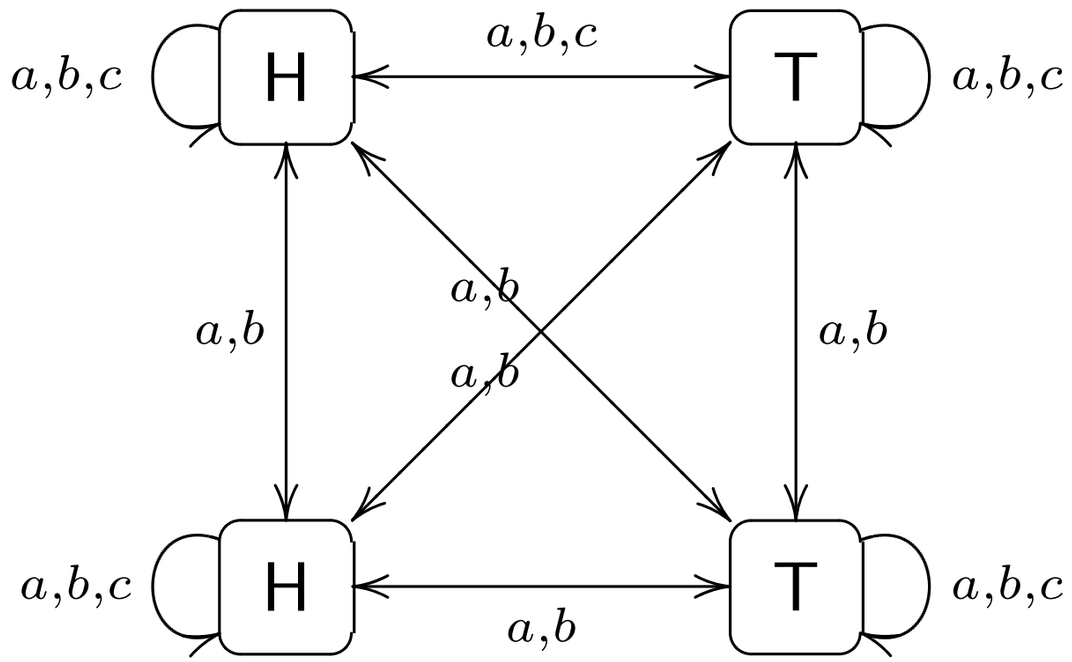
Let us represent the secret communication action: c sends a secret message to a, telling her that he (c) knows that the coin lies Heads up. This message is completely secret: the other agent b doesn't even suspect it that this is happening.

This is yet another instance of a fully **private announcement** $!_{a,c}(K_c\mathsf{H})$, having the **event model**

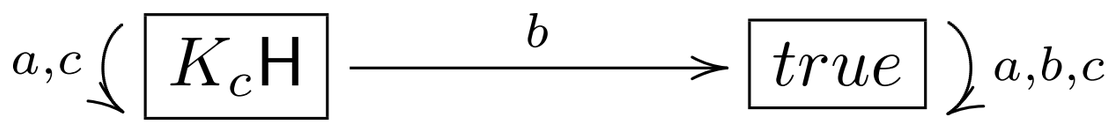


Update Product

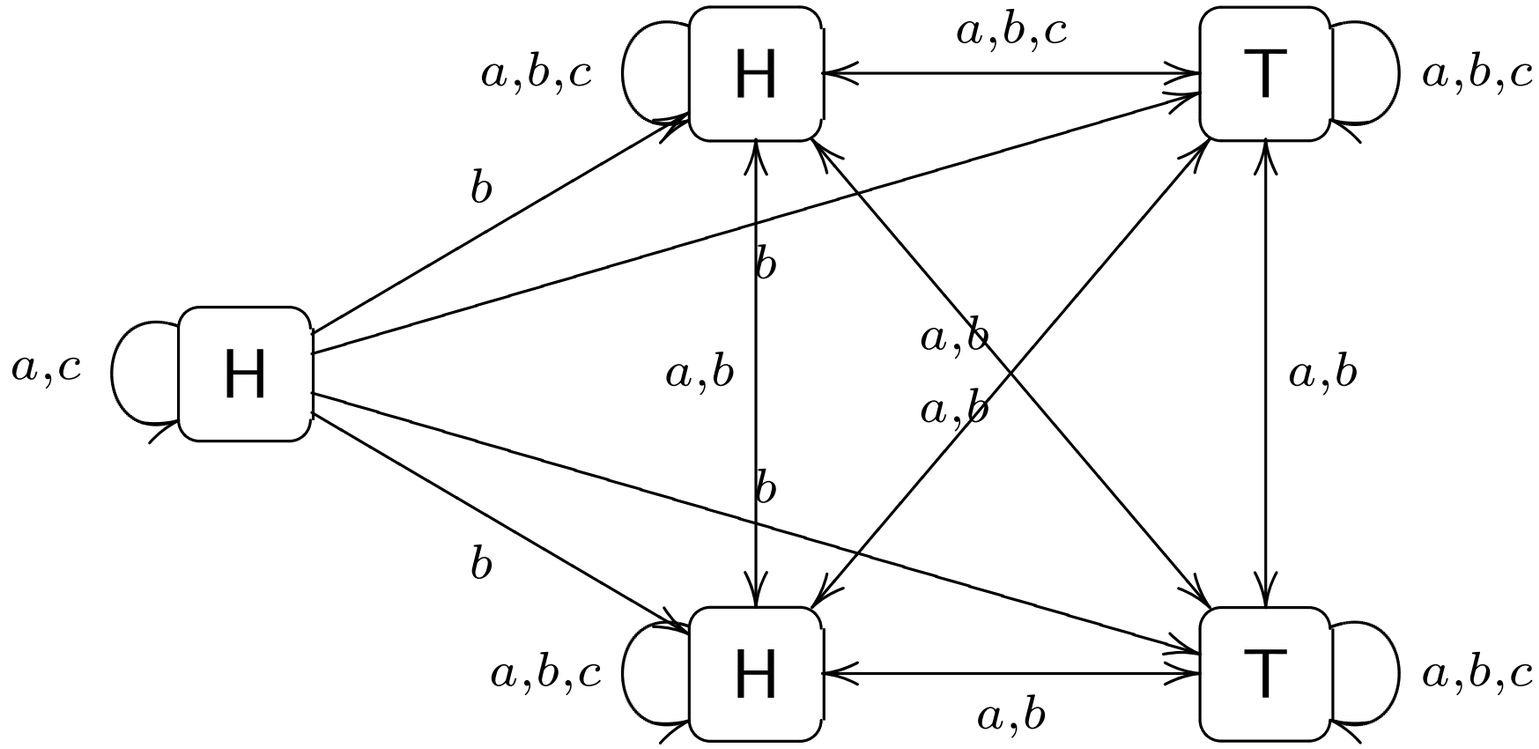
Take now the update product of the state model after Scenario 4'



with the event model for secret communication



obtaining as a result the state model after Scenario 5:



Scenario 5': Wiretapping?

Scenario 5′: after Scenario 4' happened (the referee c secretly took a peek, with the others suspecting this, but not knowing for sure), everything goes on as in Scenario 5, except that in the meantime b is **secretly breaking** into c's email account (or **wiretapping** his phone) and reading c's secret message. Nobody suspects this illegal attack on c's privacy.

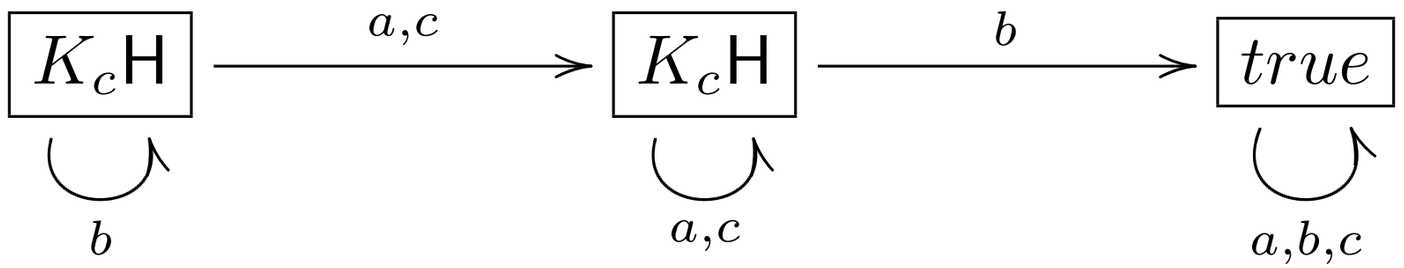
So both c and a think their secret communication is really secret and unsuspected by b: **the deceivers are deceived**.

What is the model of the situation after this action?! Things are getting rather complicated!

Solving Scenario 5': Wiretapping

The action in Scenario 5' is just like the one in Scenario 5, except that the supposedly secret message from c to a is in fact secretly intercepted by b (without a and c suspecting this: "the deceivers are deceived").

This is an instance of a *private announcement with (secret) interception by a group of outsiders*, with **event model**:



EXERCISE: Compute the result of this action, by taking the product update of the state model after Scenario 4' with this event model.

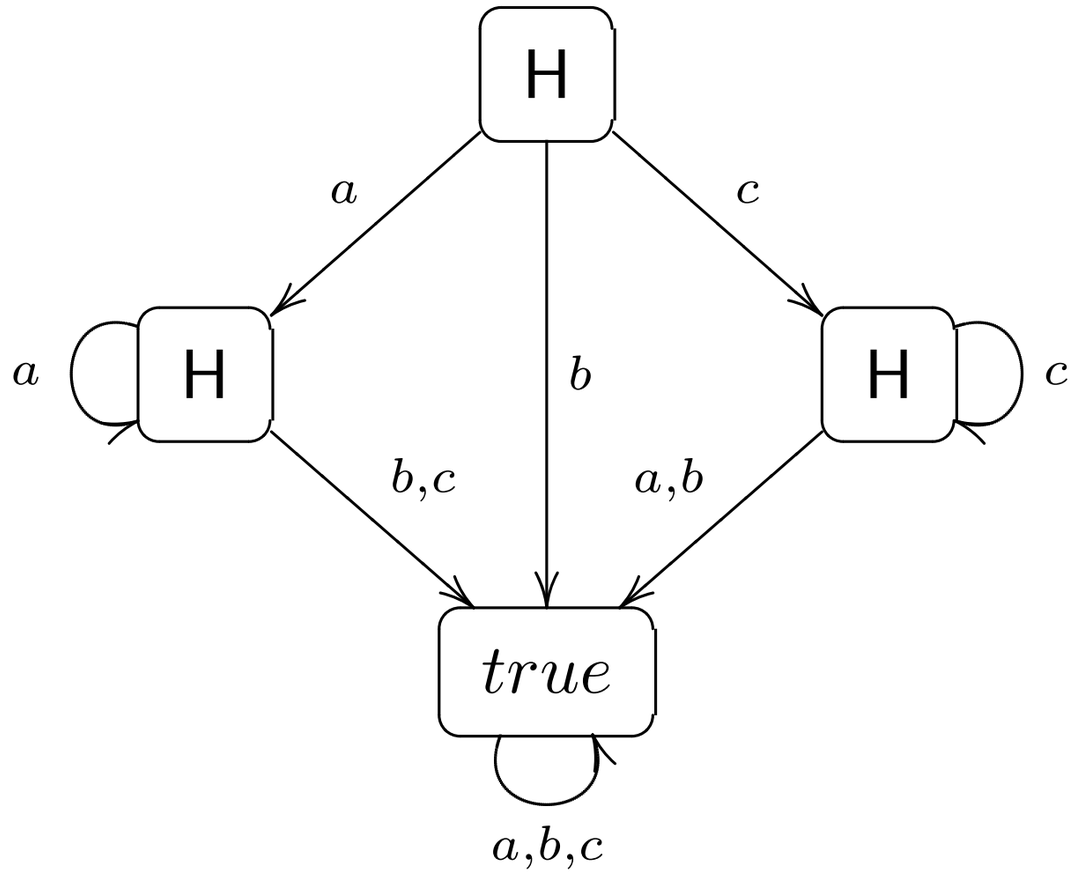
Events involving Simultaneous Actions

Scenario 5' is an example of an event that involves *two simultaneous actions*: (1) the supposedly secret communication between c and a, and (2) the interception of this communication by b.

Another example is **simultaneous peeking**: suppose that, just before the face of the coin was covered, each of agents a and c takes a peek at the coin, secretly and separately from each other. So each of them sees that the coin lies Heads up, but in such a way that nobody else suspects this. Moreover, each of them knows that the others don't suspect what he's doing.

Let us draw the event model for this action (next slide).

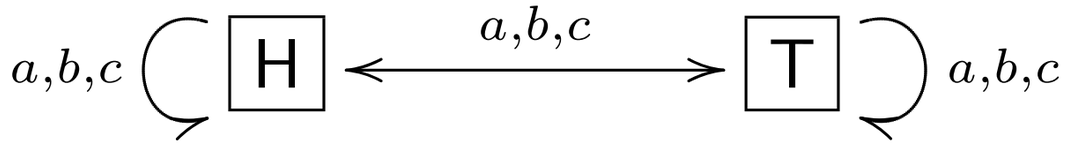
Event Model for Simultaneous Peeking



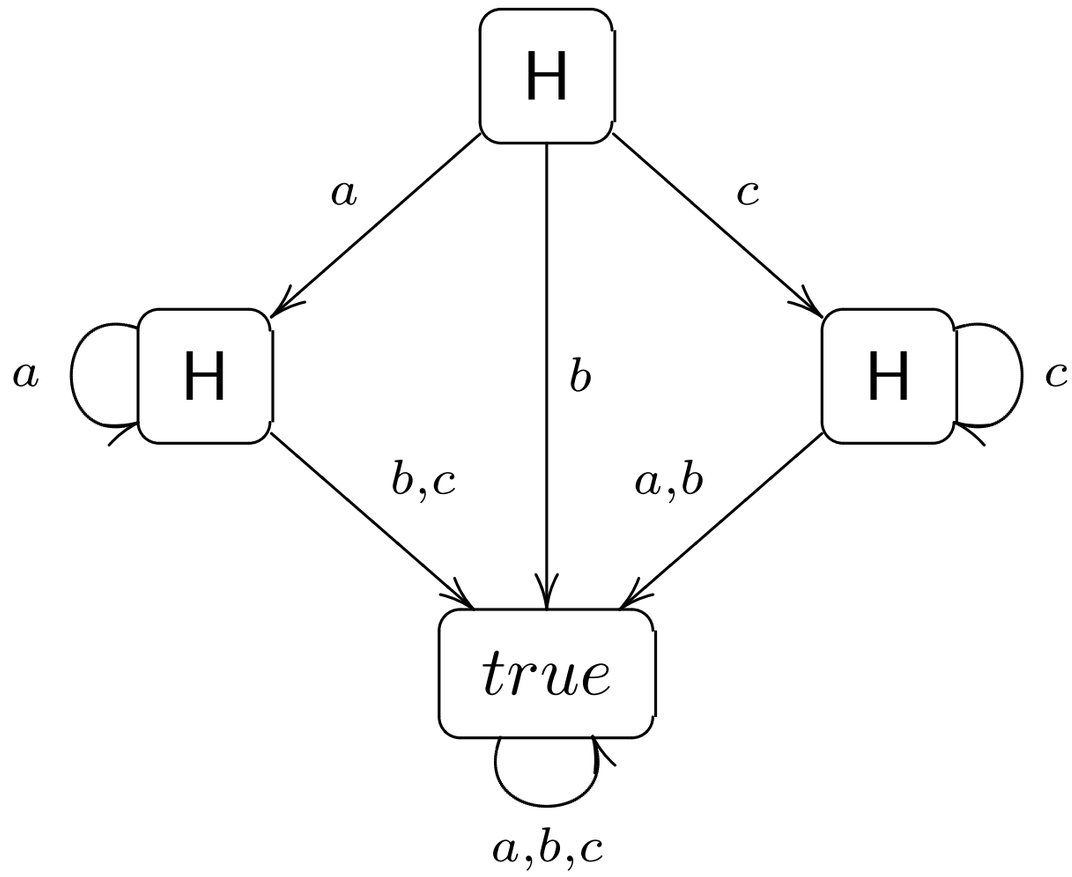
Nobody considers possible the real event that is actually happening! This is because the real event involves two simultaneous secret actions (a secretly taking a peek, and b secretly taking a peek), and nobody suspects that this simultaneous pair of actions happens!

Product Update

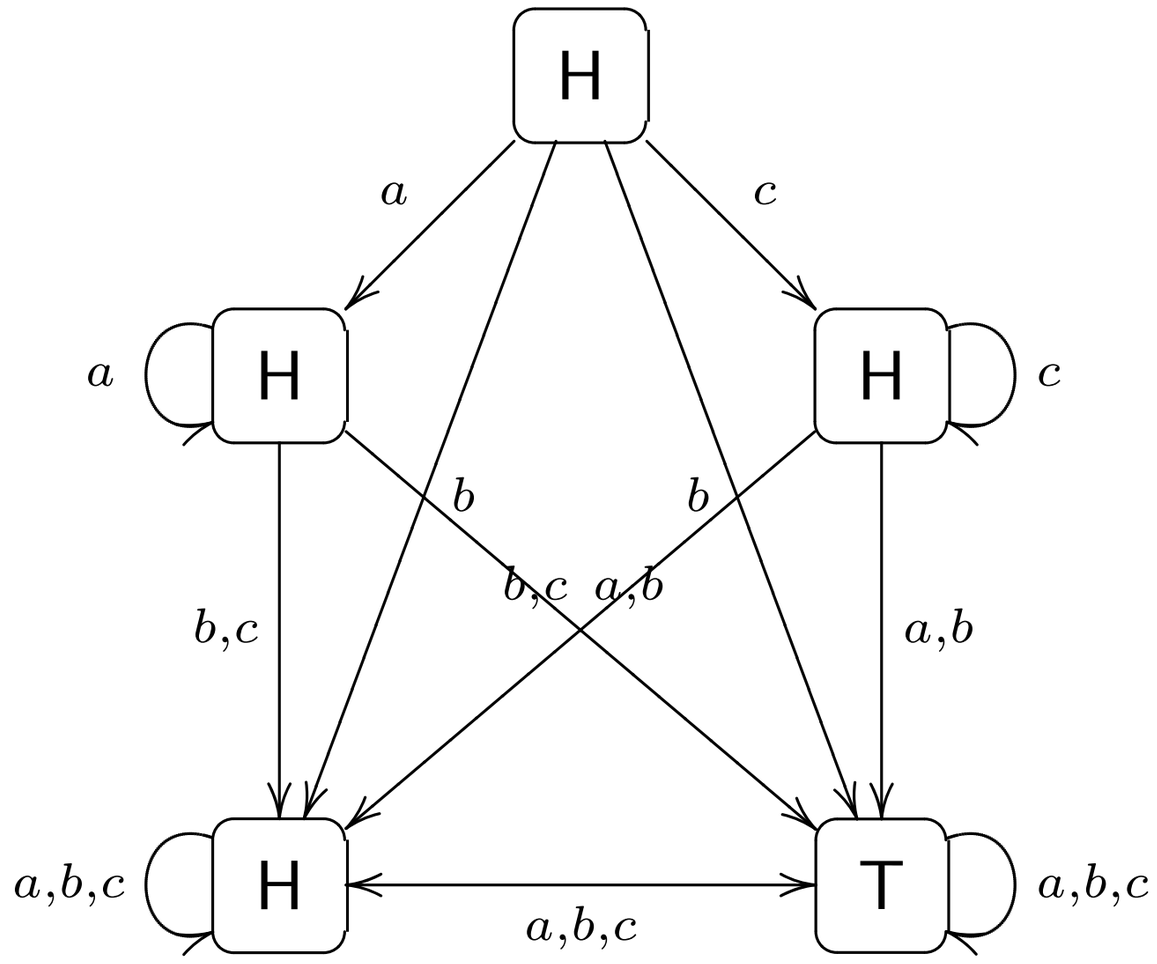
By taking the update product of the original state model



with the event model



we obtain the following updated state model



which correctly represents the informational situation after the simultaneous secret-peeking actions.

Scenario 6

This starts right after Scenario 2, when it was common knowledge that c knew the face. c attempts to send a secret message to a announcing that (he knows that) H is the case. c is convinced that the communication channel is fully secure and reliable; moreover, he thinks that b doesn't even suspect this secret communication is going on.

But, in fact, unknown and unsuspected by c, the message is *intercepted and read* by b, who also stops it (prevents it) from arriving at its intended destination. As a result, *it never makes it to* a, and in fact a never knows or suspects any of this. As for b, he *knows* all of the above: not only now he knows the message, but he knows that he "fooled" everybody, in the way described above.

EXERCISE: *Represent the event in Scenario 6 and compute the updated state model after that event.*

Change of Facts

We can generalize our setting our setting, by allowing for events that **may change some of the facts of the world**, not just the agents' information or beliefs. This means that such actions **may now change the valuation of atomic sentences**.

Formally, this is done by adding to the structure of the event models a **post-condition map** $post$, that associates to each event $e \in E$ some set $POST_e$ of "post-conditions" of the form

$$p := post_e(p),$$

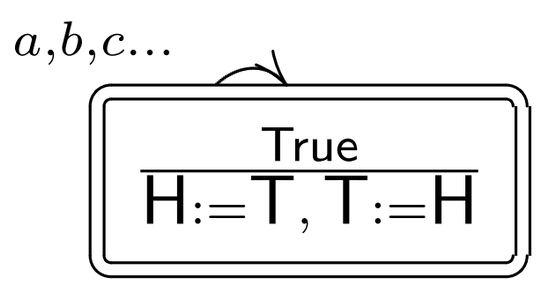
(one for each atomic sentence $p \in \Phi$), where $post_e(p)$ are any sentences in our language.

The intended meaning is that p **WILL become true after event** e **iff** $post_e(p)$ **was true before the event**.

Example: Public Fact Change

In front of everybody, a coin is turned upside down, without anybody looking at the face of the coin: so, after this it is common knowledge that the coin will be Heads up iff it was Tails up before (and vice-versa).

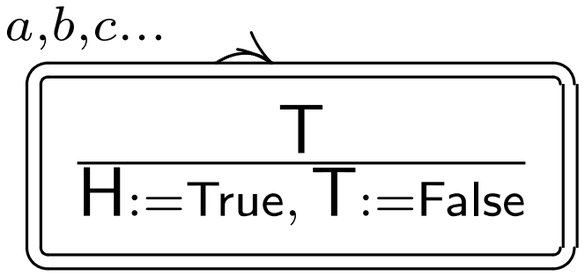
The event model for this is:



The sentence *above the line is the event's precondition* of this event, while **the event's postconditions are below the line.**

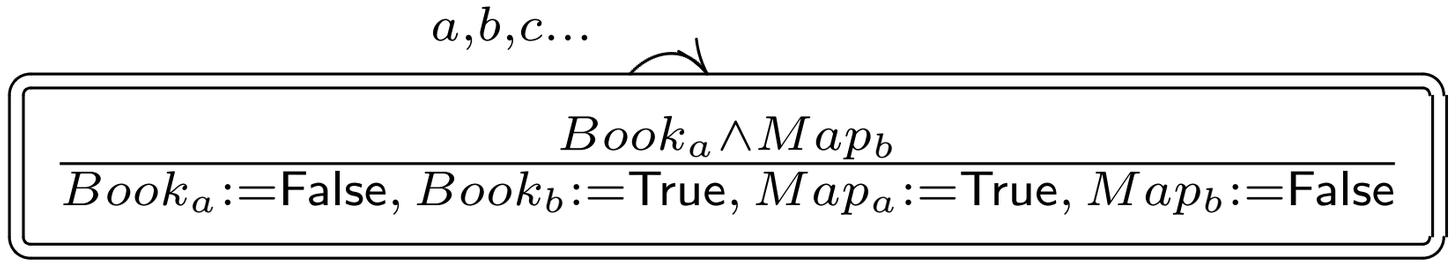
Another Public Fact-Change

In front of everybody, the coin is turned upside down and it is seen by all to be now lying Heads up.



A Public Exchange

In front of everybody, Alice and Bob exchange gifts: she gives him a rare old edition of Darwin's book *The Origin of Species*, while he gives her a valuable 17th century-old map.



Here, $Book_i$ means that *agent i has Darwin's book*, and Map_i means that *agent i has the 17th century map.*

EXERCISE: Model a **secret exchange** between Alice and Bob of the same items, exchange that is unknown to, and unsuspected by, Charles.

The Product Update with Fact-Changing Event Models

It is defined in the same way as before, except that the *valuation* ν' at world (w, e) in the new model $\mathbf{M} \otimes \mathbf{\Sigma}$ is given by *event e's postconditions*:

$$\nu'(w, e) = \{p \in Prop : w \models_{\mathbf{M}} post_e(p)\}.$$

Lecture 3.1. The Logic of Public Announcements

PAL (Public Announcement Logic) was first formalized (including complete axiomatization) by Plaza (1989), and independently by Gerbrandy and Groeneveld (1997).

The logic PAC (Public Announcement Logic with Common knowledge) is obtained by adding to PAC the **common knowledge** operator.

The problem of **completely axiomatizing** PAC was first solved by Baltag, Moss and Solecki (1998).

PAL is just one example of "dynamic epistemic logic".

Dynamic Epistemic Logics study informational changes: the **dynamics** of knowledge/belief.

Dynamic Modalities

To express **informational changes**, dynamic epistemic logics use a new kind of propositional operators, called **dynamic modalities**:

$$[\alpha]\varphi$$

Here, α is the *name of some action* involving some kind of **communication**. Such actions are called epistemic actions, since they knowledge (only) the knowledge/beliefs of the agents.

The *intended meaning* of $[\alpha]\varphi$:

"**if action α is performed then φ will become true after this**".

The "if" part is interpreted as a *conditional*: $[\alpha]\varphi$ is be definition **true** in worlds in which action α **cannot** be performed.

Example: the Public Announcement Modality

An example of such action α is the **(truthful) public announcement** $!\varphi$ of some sentence φ.

So we can write

$$[!\varphi]\psi$$

to mean that: **if a truthful public announcement $!\varphi$ is performed, then ψ will become true after this.**

What happens if the announcement is false?

NOTE: **A truthful public announcement $!\varphi$ can only be performed if φ IS TRUE**.

So $[!\varphi]\varphi$ **is by definition true for ANY ψ in worlds w in which φ is false** (since $!\varphi$ cannot be performed in such worlds).

In particular, **if a false sentence is "truthfully announced", then... everything is true after that (including contradictions)**:

$$\neg\varphi \Rightarrow [!\varphi]\bot$$

where

$$\bot := p \wedge \neg p$$

is any sentence that is "always false" (contradictory).

Public Announcement Logic (PAL)

The syntax of **basic** PAL is obtained by adding to basic multi-modal logic dynamic modalities for public announcements:

$$\varphi \quad ::= \quad p \quad | \quad \neg\varphi \quad | \quad \varphi \wedge \varphi \quad | \quad \Box_a \varphi \quad | \quad [!\varphi]\varphi$$

These formulas are interpreted on multi-agent Kripke models.

If we use K instead of $\Box$ in the above syntax, and interpret formulas on epistemic (S5) models, we obtain **epistemic logic with public announcements**.

Similarly, if we use B instead of $\Box$, we obtaining the **doxastic logic with public announcements**.

Finally, we can also put this together with numerical epistemic logic formulas $n = i$, $K_a n$, obtaining **numerical epistemic logic with public announcements**.

Formal Semantics

The important thing is that **the formal semantics of** $[!\varphi]\psi$ **in a model** $\mathbf{M} = (W, R_1, \ldots, R_n, \nu)$ uses the semantics of ψ in the **updated model** $\mathbf{M}^{!\varphi} = (W', R'_1, \ldots, R'_n, \nu')$ (obtained by **deleting** the worlds that do **not** satisfy φ):

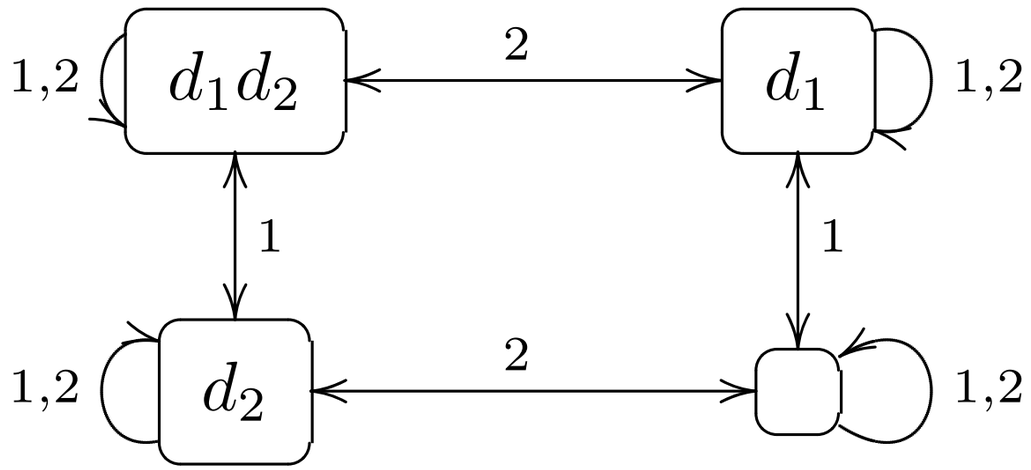
$$w \models_{\mathbf{M}} [!\varphi]\psi \quad \text{iff:} \quad (\text{ if } !\varphi \text{ can be performed, then } w \models_{\mathbf{M}^{!\varphi}} \psi).$$

More precisely:

$$w \models_{\mathbf{M}} [!\varphi]\psi \quad \text{iff} \quad \text{either } w \not\models_{\mathbf{M}} \varphi \text{ or } w \models_{\mathbf{M}^{!\varphi}} \psi.$$

Example

In the Two Muddy Children Example, let **M** be the model



and let $w = (d_1, d_2)$ be the real world (in the upper left corner). Then, *after "truthfully announcing" that child* 1 *is both dirty and clean, child* 2 *will know that she's clean!*

$$w \models_{\mathbf{M}} [!(d_1 \wedge \neg d_1)]K_2\neg d_2.$$

EXERCISE: **Why?**

More Examples

In the same model (in the real world w), we also have that, *after Father's announcement that at least one of them is dirty, they will both know that at least one of them is dirty*:

$$w \models_{\mathbf{M}} [!(d_1 \vee d_2)]\,(K_1(d_1 \vee d_2) \wedge K_2(d_1 \vee d_2))$$

In fact, we have that, *after Father's announcement that at least one of them is dirty, it will be common knowledge that at least one of them is dirty*:

$$w \models_{\mathbf{M}} [!(d_1 \vee d_2)]Ck(d_1 \vee d_2).$$

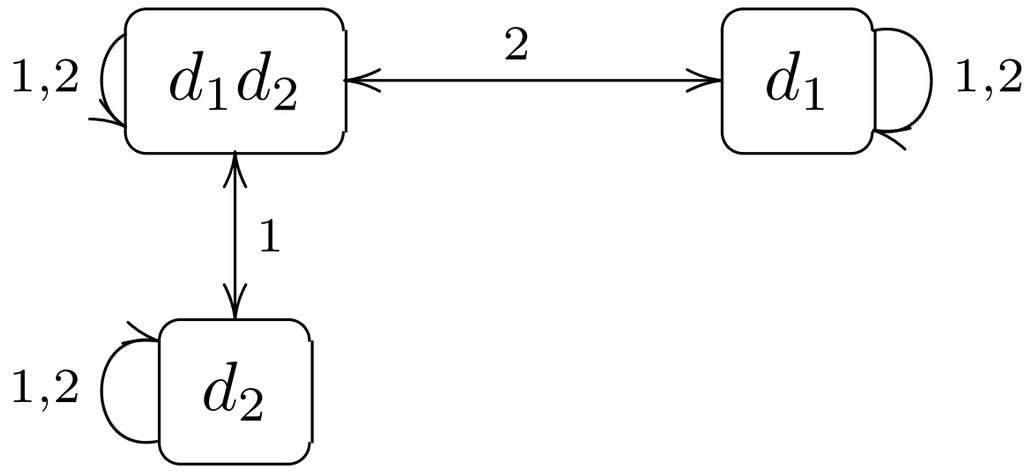
QUESTION: **Why?**

Explanations

By definition of $[!\varphi]\psi$, since $w \models_{\mathbf{M}} d_1 \vee d_2$, we have that

$$w \models_{\mathbf{M}} [!(d_1 \vee d_2)]Ck(d_1 \vee d_2) \quad \text{iff} \quad w \models_{\mathbf{M}^{!(d_1 \vee d_2)}} Ck(d_1 \vee d_2)$$

So we have to look at the truth value of $Ck(d_1 \vee d_2)$ in the updated model $\mathbf{M}^{!(d_1 \vee d_2)}$:



In this updated model $d_1 \vee d_2$ is true in ALL worlds, so we DO have

$$w \models_{\mathbf{M}^{!(d_1 \vee d_2)}} Ck(d_1 \vee d_2)$$

hence the conclusion follows.

Are sentences always known after truthfully announced?

Intuitively, it may seem that the last example can be generalized to "*every sentence becomes common knowledge after is (truthfully) publicly announced*":

$$(?) \qquad [!\varphi]Ck\varphi$$

for every sentence φ.

It may seem even more obvious that "*every sentence becomes KNOWN after is (truthfully) publicly announced*":

$$(??) \qquad [!\varphi]K_a\varphi$$

for every sentence φ and every agent a.

ARE THESE CLAIMS CORRECT??

Well, if they were correct then the following would also be correct:

"*every sentence is (or becomes) TRUE after is (truthfully) publicly announced*"

$$(???) \qquad [!\varphi]\varphi$$

(Question: WHY DOES THIS FOLLOW FROM THE ABOVE CLAIMS?).

"Moore Sentences" become FALSE after being ANNOUNCED!

Unfortunately, all the above claims are WRONG:

there exist SOME sentences φ such that we always have

$$\neg[!\varphi]\varphi$$

and in fact

$$[!\varphi]\neg\varphi.$$

Moreover, such sentences always **become KNOWN TO BE FALSE after being announced**:

$$[!\varphi]K_a\neg\varphi$$

and in fact **they become COMMONLY KNOWN TO BE FALSE after being (publicly) announced**:

$$[!\varphi]Ck\neg\varphi.$$

Such sentences are usually called "*Moore sentences*", after the name of the philosopher who first discussed them.

Examples of Moore Sentences

"*You are dirty but you don't know it*" (where "you" means child 1):

$$d_1 \wedge \neg K_1 d_1$$

"*Both children are dirty but none of them knows he's dirty*":

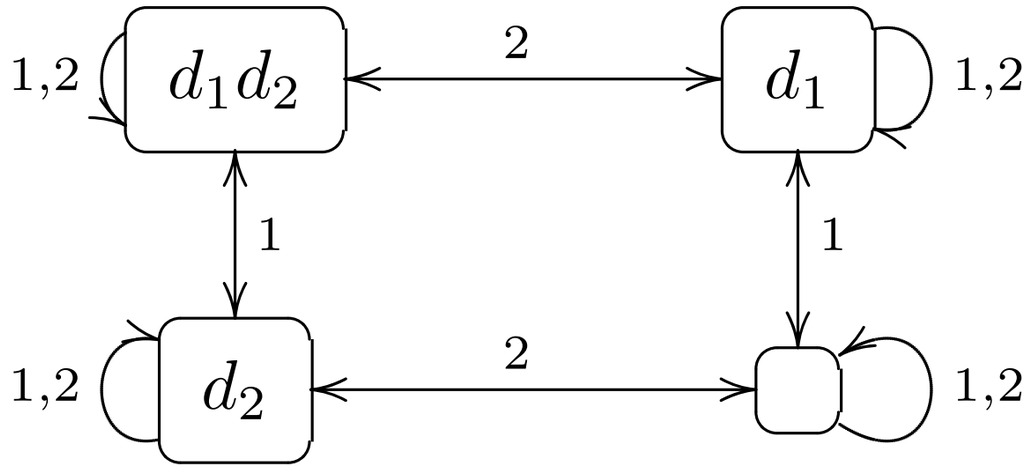
$$d_1 \wedge d_2 \wedge \neg K_1 d_1 \wedge \neg K_2 d_2$$

"*Both of you are dirty but none of you know this* (=that you are both dirty)":

$$d_1 \wedge d_2 \wedge \neg K_1(d_1 \wedge d_2) \wedge \neg K_2(d_1 \wedge d_2)$$

All these are Moore Sentences!

All these sentences are **true** in the real world $w = (d_1, d_2)$ of the initial model $\mathbf{M}$ of the Two Muddy Children Story:



E.g.

$$w \models_{\mathbf{M}} d_1 \wedge \neg K_1 d_1$$

But they all **become (COMMONLY) (KNOWN TO BE) FALSE after they are (truthfully) publicly announced!**

E.g.

$$w \models_{\mathbf{M}} [!(d_1 \wedge \neg K_1 d_1)]\neg(d_1 \wedge \neg K_1 d_1)$$

$$w \models_{\mathbf{M}} [!(d_1 \wedge \neg K_1 d_1)] K_1 \neg (d_1 \wedge \neg K_1 d_1)$$

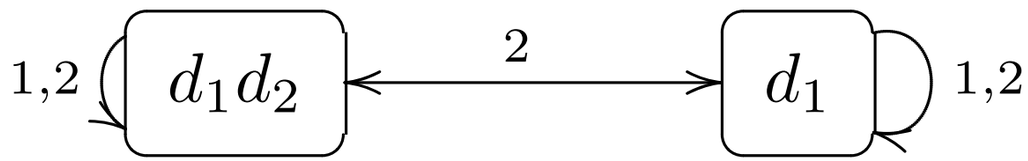
$$w \models_{\mathbf{M}} [!(d_1 \wedge \neg K_1 d_1)] Ck \neg (d_1 \wedge \neg K_1 d_1)$$

Proof

To see this, we have to look at the updated model

$$\mathbf{M}^{!(d_1 \wedge \neg K_1 d_1)}$$

obtained by deleting the worlds in which $d_1 \wedge \neg K_1 d_1$ is FALSE:



NOW (in this model), the announced sentence has become FALSE:

$$w \models_{\mathbf{M}^{!(d_1 \wedge \neg K_1 d_1)}} \neg(d_1 \wedge \neg K_1 d_1)$$

and in fact it became common knowledge that it's false:

$$w \models_{\mathbf{M}^{!(d_1 \wedge \neg K_1 d_1)}} Ck\neg(d_1 \wedge \neg K_1 d_1)$$

WHY?

More Examples

Other examples are to be found in the Alice, Bob and Charles story (the "love triangle"):

"Bob doesn't know about our affair"

$$(affair) \wedge \neg K_b(affair)$$

The Muddy Children Story (with $k > 2$ dirty children) gives us more complex phenomena of the same type:

the sentence

"**nobody knows if he's dirty or not**"

is true in the beginning, it is publicly announced once, after which it is still true,

BUT it becomes false after $k-1$ repeated announcements!

Announcements about Announcements

It is important that in PAL we can iterate all the constructions:
we can announce, not only facts

$$!p$$

or combinations of facts (Boolean formulas)

$$!(p \vee \neg q),$$

but also epistemic formulas

$$!(\neg K_a p)$$

and can even make *announcements about other announcements*

$$!([!q]\neg K_a p).$$

This last fact is essential for having the following nice closure property.

Closure of public announcements under composition

The following is valid on all models:

$$[!\varphi][!\psi]\theta \quad \Leftrightarrow \quad [!(\varphi \wedge [!\varphi]\psi)]\theta$$

Semantically, this captures the **closure of the class of public announcements under sequential compositions**:

performing successively two public announcements

$$!\varphi; !\psi$$

is equivalent to performing only one more complex public announcement

$$!(\varphi \wedge [!\varphi]\psi)$$

EXERCISE: Why not just

$$!(\varphi \wedge \psi)$$

???

Find a **counterexample** to

$$[!\varphi][!\psi]\theta \quad \Leftrightarrow \quad [!(\varphi \wedge \psi)]\theta$$

Validities

The following laws (called "*Reduction Axioms*") are valid on all models:

Atomic Permanence:

$$[!\varphi]p \iff (\varphi \Rightarrow p) \qquad (\text{for } \textit{atomic} \text{ props } p),$$

Announcement-Negation:

$$[!\varphi]\neg\psi \iff (\varphi \Rightarrow \neg[!\varphi]\psi)$$

Announcement-Conjunction:

$$[!\varphi](\psi \wedge \theta) \iff ([!\varphi]\psi \wedge [!\varphi]\theta)$$

Announcement-Knowledge:

$$[!\varphi]\Box_a\psi \iff (\varphi \Rightarrow \Box_a[!\varphi]\psi)$$

Reduction Axioms for Numerical Logic

We also have validities exprssing Reduction Laws for numerical epistemic sentences

Permanence of Value:

$$[!\varphi](n = i) \iff (\varphi \Rightarrow n = i) \qquad \text{(for variables } n \in Var \text{ and numbers } i)$$

Announcement-Numerical Knowledge:

$$[!\varphi]K_a n \iff (\varphi \Rightarrow K_a^{\varphi} n)$$

and more generally

Announcement-Conditional-Numerical Knowledge:

$$[!\varphi]K_a^{\psi} n \iff \left(\varphi \Rightarrow K_a^{\varphi \wedge [!\varphi]\psi} n\right)$$

APPLICATION: Muddy Children

Let d_i be an atomic sentence saying that "*child i is dirty*".

Then the following **abbreviations** can be made in PAC:

$$vision \;:=\; \bigwedge_{a\in\mathcal{A}} \bigwedge_{b\neq a} ((d_b \Leftrightarrow \Box_a d_b) \wedge (\neg d_b \Leftrightarrow \Box_a \neg d_b))$$

$$at\,least\,one \;:=\; \bigvee_{a\in\mathcal{A}} d_a$$

$$exactly\,k \;:=\; \bigvee_{G\subseteq\mathcal{A},|G|=k} \left(\bigwedge_{a\in G} d_a \wedge \bigwedge_{a\notin G} \neg d_a \right)$$

$$nobody\,knows \;:=\; \bigwedge_{a\in\mathcal{A}} (\neg\Box_a d_a \wedge \neg\Box_a \neg d_a)$$

$$dirties\,know \;:=\; \bigwedge_{a\in\mathcal{A}} (d_a \Rightarrow \Box_a d_a)$$

Then the Muddy children scenario can be encoded in the formula:

$$(exactly\, k \wedge C\Box vision) \Rightarrow [!(at\, least\, one)][!(nobody\, knows)]^{k-1} dirties\, know$$

EXERCISE (VERY HARD!):

To prove this sentence, use the Reduction Axioms to push the dynamic modalities, past each other logical operator, under they get to the "bottom" (i.e. in front of atomic sentences), when they can completely eliminated (using the Atomic Permanence law).

After this, we can just check that the resulting sentence is valid on all epistemic models.

Lecture 3.2. Dynamic Epistemic Logics

"**Dynamic Epistemic Logic**" by van Ditmarsch, van der Hoek and Kooi

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

We will follow roughly the last of these (which I did put on Blackboard as pdf).

Dynamic Modalities

For any event $e \in E$, we can consider the corresponding **dynamic modality** $[e]\varphi$. This is a property *of the original model*, expressing the fact that, if action e happens, then φ will come to be true after that.

As before, the semantics of $[e]\varphi$ is given by looking at φ's truth value in the **new (updated) model**:

$$w \models_{\mathbf{M}} [e]\varphi \text{ iff } (w, e) \in \mathbf{M} \otimes \mathbf{\Sigma} \text{ implies } (w, e) \models_{\mathbf{M} \otimes \mathbf{\Sigma}} \varphi$$

The Language of Dynamic Epistemic Logic

The language of basic Dynamic Epistemic Logic (DEL) is simply obtained by adding to basic modal (epistemic) logic dynamic modalities for any given finite event model Σ:

$$\varphi \quad ::= \quad p \quad | \quad \neg\varphi \quad | \quad \varphi \wedge \varphi \quad | \quad \Box_a \varphi \quad | \quad [e]\varphi$$

As before, we can use K instead of $\Box$ in the above syntax when the models are epistemic, and we can use B instead of $\Box$ when the models are doxastic.

Finally, we can also put this together with numerical epistemic logic formulas $n = i$, $K_a n$, obtaining **Numerical** DEL.

(Sequential) Composition of Epistemic Events

Given two event models $\mathbf{\Sigma} = (E, R_1, \ldots, R_n, pre)$ and $\mathbf{\Sigma}' = (E', R_1, \ldots, R_n, pre)$, their **composition** is a new event model

$$\mathbf{\Sigma} \cdot \mathbf{\Sigma}' = (E \times E', R_1, \ldots, R_n, pre), \qquad \text{where}$$

1. $E \times E' := \{(e, e') : e \in E, e' \in E'\}$ is the Cartesian product of the two models.

2. $(e, e')R_a(f, f')$ iff eR_af and $e'R_af'$.

3. $pre_{(e,e')} := pre_e \wedge [e]pre_{e'}$.

NOTATION: We denote by $e \cdot e'$ the pair (e, e') as an event in $\mathbf{\Sigma} \cdot \mathbf{\Sigma}'$. This denote **the succession of** e **followed by** e', but thought of as *one single event*, with precondition $pre_{e \cdot e'} = pre_e \wedge [e]\textsc{Pre}_{e'}$ and accessibility relations: $(e \cdot e')R_a(f \cdot f')$ iff eR_af and $e'R_af'$.

Actions are closed under Sequential Composition

It is easy to see that that *taking the Product Update with a composed event model is equivalent to taking successive Product Updates with the two event models*:

$$\mathbf{S} \otimes (\mathbf{\Sigma} \cdot \mathbf{\Sigma}') = (\mathbf{S} \otimes \mathbf{\Sigma}) \otimes \mathbf{\Sigma}'.$$

This confirms our *intended interpretation*:
intuitively, $e \cdot e'$ is the **sequential composition of the two events**:
the action obtained by **performing first action** e **then action** e'**.**

This is captured by the following validity:

$$[e][f]\varphi \iff [e \cdot f]\varphi.$$

Axioms of DEL (Reduction Laws)

If $e \in E$ is an epistemic event, the following are valid on all models:

- *Atomic Permanence.* For all atomic $p \in \Phi$:

$$[e]p \iff (\, pre_e \Rightarrow p \,)$$

- *Partial Functionality*:

$$[e]\neg\varphi \iff (\, pre_e \Rightarrow \neg[e]\varphi \,)$$

- *Normality*:

$$[e]\,(\varphi \wedge \psi) \iff (\, [e]\varphi \wedge [e]\psi \,)$$

- *"Action-Knowledge Axioms"*: Let $f_1, \ldots, f_n$ be an enumeration of ALL events $f \in E$ such that $eR_a f$. Then:

$$[e]\Box_a\varphi \iff \left(pre_e \Rightarrow \bigwedge_{i=1,n} \Box_a[f_i]\varphi \right)$$

For (conditional) numerical knowledge, the analogue axiom is

$$[e]K_a^{\psi} n \iff \left(pre_e \Rightarrow K_a^{\bigvee_{i=1,n}(pre_{f_i} \wedge [f_i]\psi)} n \right)$$

English Reading

The **Action-Knowledge Axiom** helps us to *compute the state of knowledge/belief* of an agent *after* an event, in terms of the agent's *initial state of knowledge or belief* and of the event's *appearance* to the agent.

In English, the Action-Knowledge axiom says that:

a proposition φ will be known (to an agent a) AFTER an event e iff, assuming that the event e CAN happen, it is ALREADY known (to agent a, BEFORE the event) that φ WILL be true after ANY of the events $f_1, \ldots, f_n$ that agent a THINKS might be happening.

Instances of Action-Knowledge Axiom

We can immediately derive the "Announcement-Knowledge Axiom" for public announcements:

$$[!\theta]\Box_a\varphi \iff (\,\theta \Rightarrow \Box_a[!\theta]\varphi\,)$$

For *fully private announcements* $!_G\theta$ *of* θ *to a subgroup* G (with the insiders $a \in G$ gaining common knowledge that this announcement is made, while outsiders $b \notin G$ don't suspect anything):

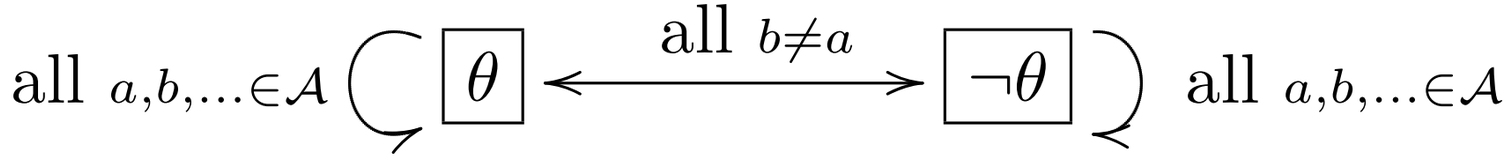
$$\text{all } _{a\in G}\circlearrowright \boxed{\theta} \xrightarrow{\text{all } _{b\notin G}} \boxed{true} \circlearrowright \text{ all } _{a,b,\ldots\in\mathcal{A}}$$

we obtain

$$[!_G\theta]\Box_a\varphi \iff (\,\theta \Rightarrow \Box_a[!_G\theta]\varphi\,), \qquad \text{for insiders } a \in G,$$

$$[!_G\theta]\Box_b\varphi \iff (\,\theta \Rightarrow \Box_b\varphi\,), \qquad \text{for outsiders } b \notin G.$$

Similarly, we can apply it to the "**fair-game announcement**" $Fair_a\theta$, in which *it is common knowledge that an agent a privately learns whether θ or $\neg\theta$ is the case*:



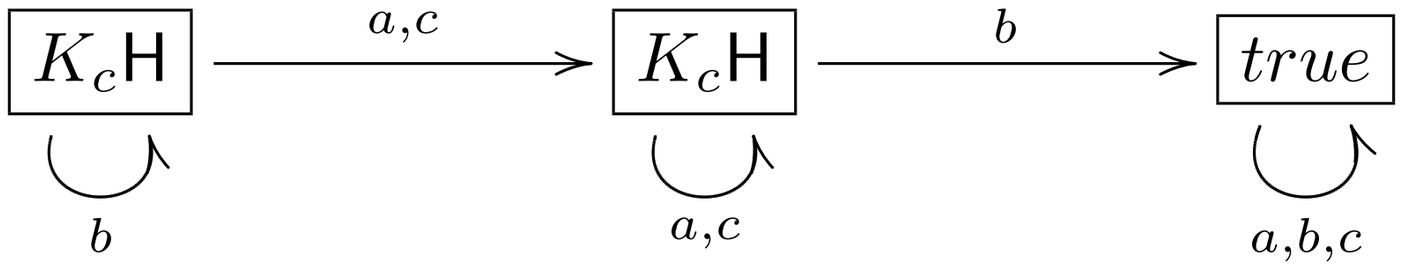
obtaining

$$[Fair_a\theta]\Box_a\varphi \iff (\,\theta \Rightarrow \Box_a[Fair_a\theta]\varphi\,), \qquad \text{and}$$

$$[Fair_a\theta]\Box_b\varphi \iff (\,\theta \Rightarrow \Box_b([[Fair_a\theta]\varphi \wedge [Fair_a\neg\theta]\varphi)\,), \qquad \text{for all } b \neq a.$$

Another Instance: Wiretapping

Recall the "wiretapping" scenario: the supposedly secret message from c to a (saying that $K_c\mathsf{H}$) is secretely intercepted by b (without them suspecting this). So a and c think they have a secret, fully private communication (of the fact that c knew that the coin lies Heads up), but in fact their secret is out: the "deceivers are deceived".



Denote by α the real action (on top), by β the action that a and b think it's happening (secret communication, in the left lower corner); and by τ the action that they think that c thinks it's happening, namely nothing ($skip$): the action in the right lower corner.

Computing Knowledge after Actions

Here, we identify "knowledge" with "correct belief":

$$K_c\varphi \Leftrightarrow (\varphi \wedge \Box_c\varphi).$$

We want to prove that

$$[\alpha]\Box_b\Box_a\Box_c\mathsf{H}$$

i.e.: *after this intercepting action, Bob believes that Alice believes that Charles believes that* H *is true.*

To show this, we use the above axioms to prove that we have the equality (logical equivalence):

$$[\alpha]\Box_b\Box_a\Box_c\mathsf{H} = true$$

Computing Knowledge after Actions

By applying Action-Knowledge axiom, we obtain:

$$[\alpha]\Box_b\Box_a\Box_c\mathsf{H} \quad = \quad K_c\mathsf{H} \rightarrow \Box_b[\alpha]\Box_a\Box_c\mathsf{H}$$

By applying it again, we get:

$$\cdots \quad = \quad K_c\mathsf{H} \rightarrow \Box_b(K_c\mathsf{H} \rightarrow \Box_a[\beta]\Box_c\mathsf{H})$$

Applying it again:

$$\cdots \quad = \quad K_c\mathsf{H} \rightarrow \Box_b(K_c\mathsf{H} \rightarrow \Box_a(K_c\mathsf{H} \rightarrow \Box_c[\beta]\mathsf{H}))$$

Applying the Preservation of Facts:

$$\cdots \quad = \quad K_c\mathsf{H} \rightarrow \Box_b(K_c\mathsf{H} \rightarrow \Box_a(K_c\mathsf{H} \rightarrow \Box_c(K_c\mathsf{H} \rightarrow \mathsf{H})))$$

But

$$K_c\mathsf{H} \rightarrow \mathsf{H} \quad = \quad true$$

(since $K_c\mathsf{H} = (\mathsf{H} \wedge \Box_c\mathsf{H})$). It is easy to check that we also have the following validity

$$\Box_i true \quad = \quad true$$

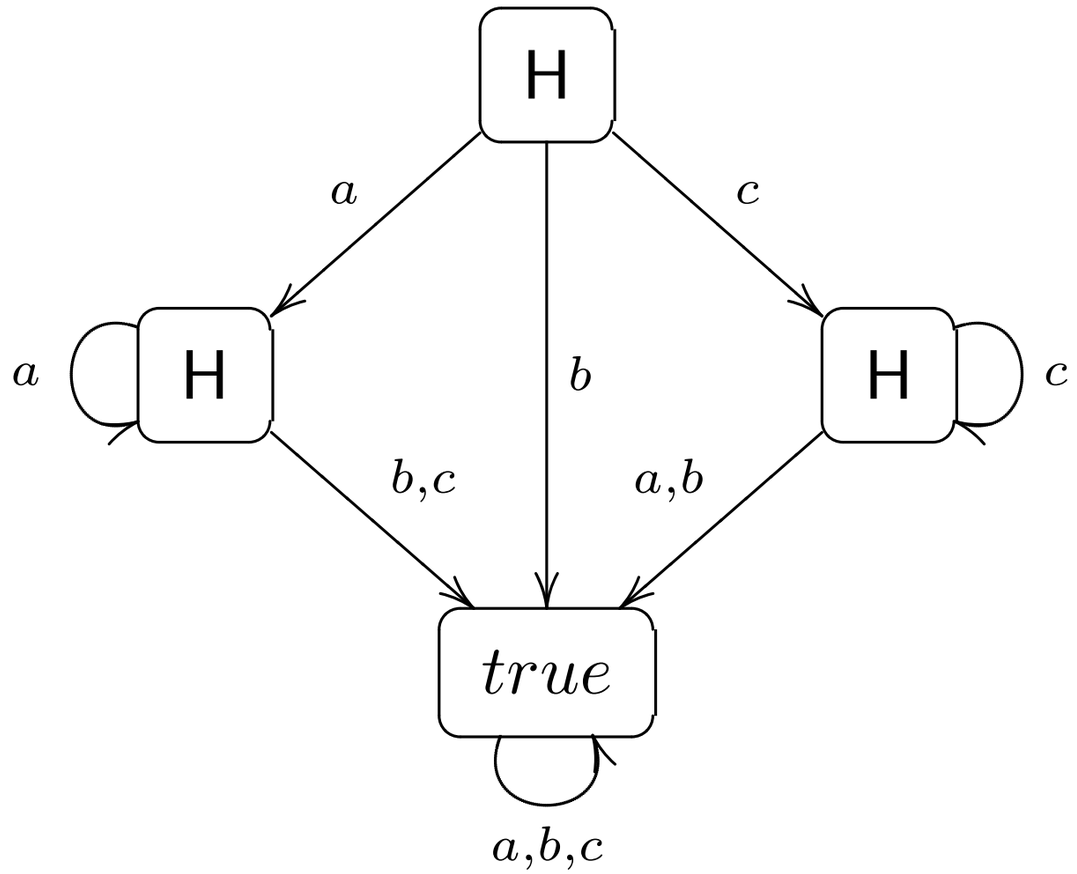
saying that every agent i knows valid sentences. Similarly, by propositional logic, we have

$$K_c\mathsf{H} \Rightarrow \mathsf{True} \quad = \quad true.$$

Repeatedly applying these last two validities, we obtain the desired conclusion.

Simultaneous Peeking

EXERCISE: Apply Action-Knowledge to the "simultaneous peeking"



by which each of agents a and c takes a peek at the coin, secretely and separately from each other (so each of them sees that the coin lies Heads up, but in such a way that nobody else suspects this).

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).

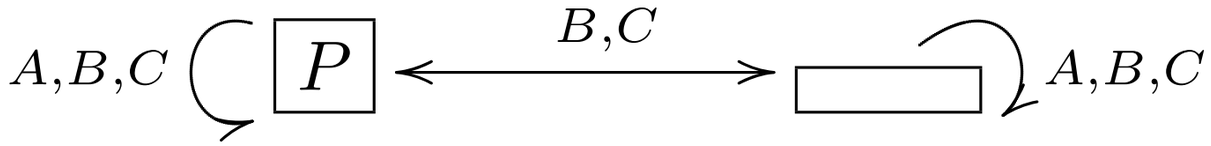
However, suppose that agent C is the only one to notice two features of the specific encryption method: first, that the shape of the already encrypted message can show whether it contains a secret (P or $\neg P$) or it's just junk; second, that without knowing the key or reading the content, he can modify 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

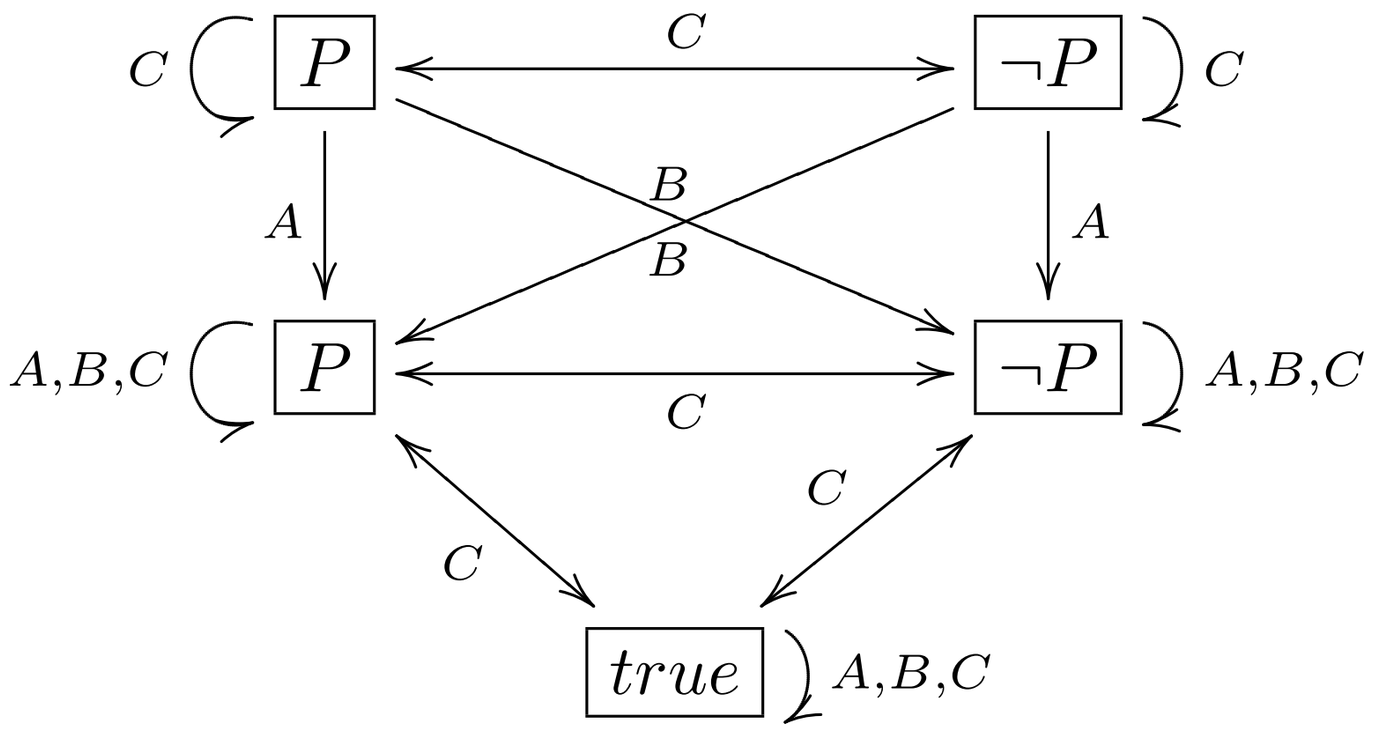
The initial situation in the scenario from the cryptographic attack above is given by:



The *real world* is the one on the left. So P holds in reality, but it would *not* hold at the other possible world (on the right). There are uncertainty arrows for B and C between any two worlds (including identical ones: the loops). This reflects the fact that B and C don't know which of these worlds is the real one: they don't know the secret. In contrast, A knows it, since at each world there is only one A-arrow (the loop).

Representation of the cryptographic attack

The epistemic action α 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) is the top-left action in the following event model:

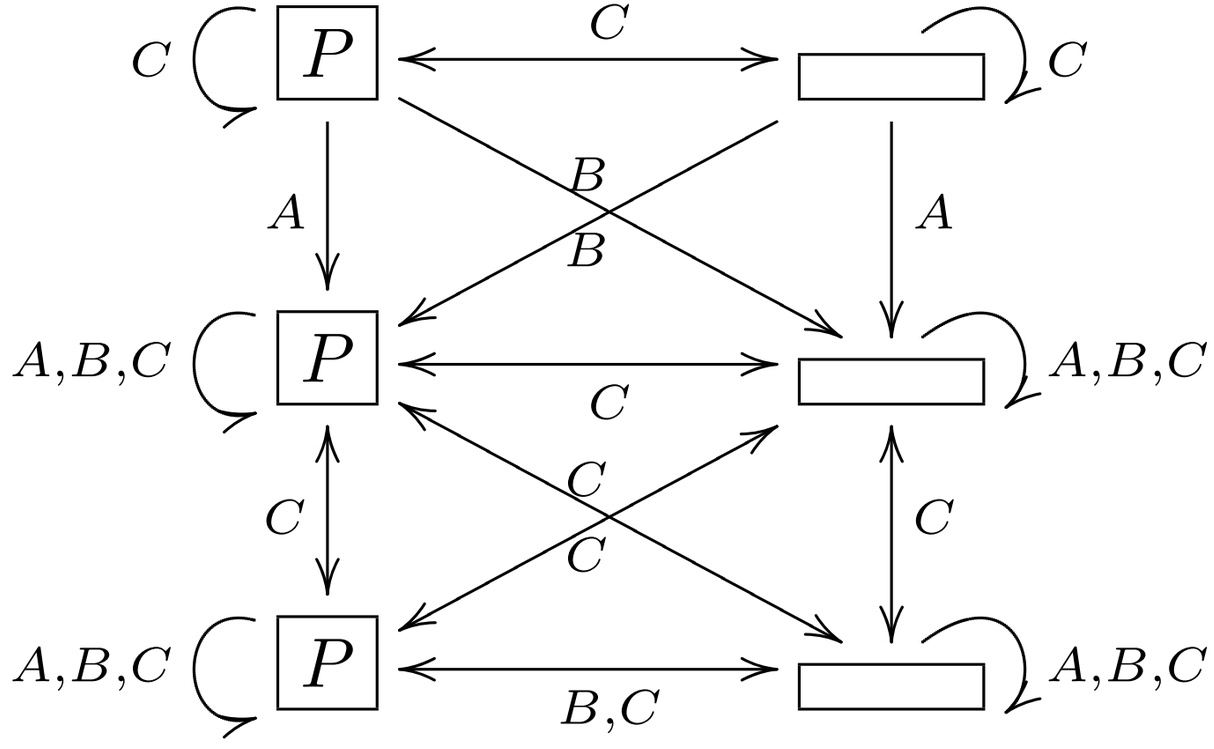


Justification

In the real action α (on top left), the correct "secret" (P) is intercepted (by C), modified and resent to B. The action on the right (call it β) is similar, but corresponding to the other possible case (when the secret was $\neg P$). Only C is aware of the possibility of these interception/modification actions, but he doesn't know which of them is happening: 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.

The update Product

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



Reasoning about the Cryptographic Attack

Let α be the action (on top right of the event model) describing the cryptographic attack. We want to prove that, *after the attack, C will know that A (wrongly) believes that B knows the secret* (P or $\neg P$, whichever is the case, since C still doesn't know which!). This is expressed by the validity of

$$[\alpha]\Box_C(\Box_A\Box_B P \vee \Box_A\Box_B \neg P)$$

To show this, we first apply the Knowledge-Action axiom (using the above notations $\alpha, \beta, \ldots$ for the action nodes in the event model):

$$\begin{aligned}[\alpha]\Box_C(\Box_A\Box_B P \vee \Box_A\Box_B \neg P) \quad = \quad & (P \to \Box_C[\alpha](\Box_A\Box_B P \vee \Box_A\Box_B \neg P)) \wedge \\ & \wedge(\neg P \to \Box_C[\beta](\Box_A\Box_B P \vee \Box_A\Box_B \neg P))\end{aligned}$$

To prove our validity, we need to show that the *formula on the right-hand side is equivalent to $True$.*

EXTRA WARMING-UP EXERCISE (WEEK 3)

To show the above equivalence, prove *first the following two validities*:

$$[\alpha]\Box_A\Box_B P,$$

$$[\beta]\Box_A\Box_B\neg P$$

(show that each of these is equivalent to $True$: use Reduction Axioms to push dynamic modalities inside while also changing the actions, and then eliminate them, using Atomic Permanence and the following equivalencies: $True \wedge True = True$, $[\alpha]True = True$ and $\Box_A True = \Box_B True = \Box_C True = True$).

Then use the validity $[\alpha]\varphi \to [\alpha](\varphi \vee \psi)$, to obtain the validities

$$[\alpha](\Box_A\Box_B P \vee \Box_A\Box_B\neg P, \quad \text{and} \quad [\beta](\Box_A\Box_B P \vee \Box_A\Box_B\neg P).$$

Finally, use $\Box_C True = True$, $(\varphi \Rightarrow True) = True$ and $True \wedge True = True$, to derive the desired equivalence.

Reduction Axioms for Fact-Changing Actions

For **fact-changing actions** (involving *post-conditions*), our "Atomic Permanence Axiom" is NOT valid anymore.

Instead, it has to be *changed* to the following:

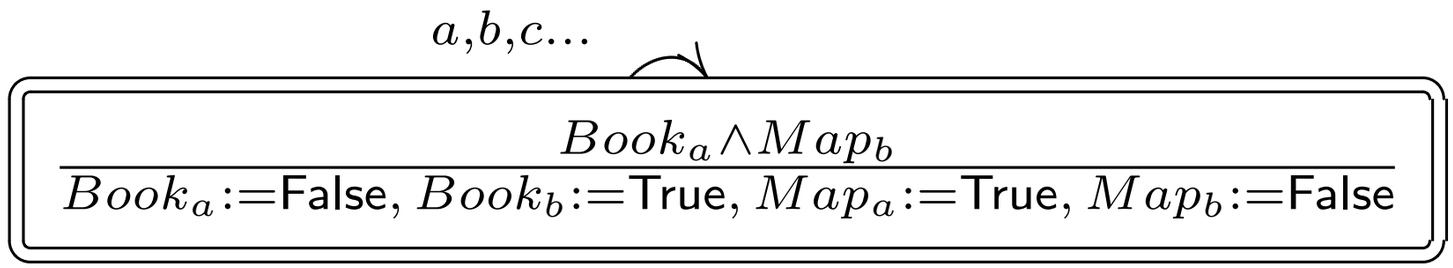
- *Fact-Change Axiom.* For all atomic $p \in \Phi$:

$$[e]p \iff (\, pre_e \Rightarrow post_e(p) \,)$$

EXERCISE: Prove that the above sentence is valid on all models for fact-changing events e.

Recall: Examples of Fact-Changing Actions

In front of everybody, Alice and Bob exchange gifts: she gives him a rare old edition of Darwin's book *The Origin of Species*, while he gives her a valuable 17th century-old map.



Here, $Book_i$ means that *agent i has Darwin's book*, and Map_i means that *agent i has the 17th century map*.

RECALL EXERCISE: Model a **secret exchange** between Alice and Bob of the same items, exchange that is unknown to, and unsuspected by, Charles.

More Useful Exercises

EXERCISE: Denote by α the *public gift-exchange action* in the previous slide. Use reduction axioms (including the Fact-Change Axiom) to show that: AFTER the exchange, *Alice will (correctly) believe that Charles (correctly) believes that Bob has Darwin's book*:

$$[\alpha]\Box_a\Box_c Book_b.$$

EXERCISE (harder!): In contrast, denote by β the *secret* gift-exchange (mentioned as an exercise on the bottom of previous slide). Using reduction axioms (and your modelling of β), show that: if initially it was common knowledge that Alice had the book, then AFTER the secret gift-exchange, *Alice (correctly) believes that Charles* ***wrongly*** *believes that Alice still has the book*:

$$C\Box Book_a \Rightarrow [\beta]\Box_a(\neg Book_a \wedge \Box_c Book_a).$$