# Study Notes

**Filip Rehburg**[1]

filip.rehburg@student.uva.nl

[1] *University of Amsterdam*, Amsterdam, The Netherlands

February 25, 2026

---

**Instructor**: Alexandru Baltag (TheAlexandruBaltag@gmail.com)
**TA**: Giuseppe Manes (giuseppe.manes@student.uva.nl)

# Contents

1 Week 1 ... 5
1.1 (Lecture): Introduction: Motivation, Main Themes, Puzzles ... 5
1.2 (Lecture): Main Themes, Puzzles, and Paradoxes Continued ... 7
1.3 (Lecture): Single-Agent Epistemic-Doxastic Logics: Kripke Models ... 12
2 Week 2 ... 18
2.1 (Lecture): Multi-agent Models and Public Announcement Logic (PAL) ... 18
2.2 (Lecture): PAL Continued ... 21
2.3 (Lecture): Does this one even exist?? ... 21
3 Week 3 ... 22
3.1 (Lecture): "Learnability" and "Knowability" ... 22
3.2 (Lecture): Tutorial 1 ... 22
3.3 (Lecture): The problem of belief revision ... 22
4 Week 4 ... 23
4.1 (Lecture): Cheating and the Failure of Standard DEL ... 23
4.2 (Lecture): ... 29
4.3 (Lecture): ... 29
5 Week 5 ... 30
5.1 (Lecture): ... 30
5.2 (Lecture): ... 30
5.3 (Lecture): ... 30
6 Week 6 ... 31
6.1 (Lecture): ... 31
6.2 (Lecture): ... 31
6.3 (Lecture): ... 31
7 Week 7 ... 32
7.1 (Lecture): ... 32
7.2 (Lecture): ... 32
7.3 (Lecture): ... 32
8 Week 8 ... 33
8.1 (Lecture): ... 33
8.2 (Lecture): ... 33
8.3 (Lecture): ... 33
I Glossary: Definitions and Theorems ... 34

| **Lecture** | **Status** |
|---|---|
| Introduction: Motivation, Main Themes, Puzzles : Section 1.1 | ✓ |
| Main Themes, Puzzles, and Paradoxes Continued : Section 1.2 | ✓ |
| Single-Agent Epistemic-Doxastic Logics: Kripke Models : Section 1.3 | ✓ |
| Multi-agent Models and Public Announcement Logic (PAL) : Section 2.1 | WIP |
| PAL Continued : Section 2.2 | ✗ |
| Does this one even exist?? : Section 2.3 | ✗ |
| "Learnability" and "Knowability" : Section 3.1 | ✗ |
| The problem of belief revision : Section 3.3 | ✗ |
| Cheating and the Failure of Standard DEL : Section 3.3 | WIP |
| : Section 4.2 | $\emptyset$ |
| : Section 4.3 | $\emptyset$ |
| : Section 5.2 | $\emptyset$ |
| : Section 5.3 | $\emptyset$ |
| : Section 6.2 | $\emptyset$ |
| : Section 6.3 | $\emptyset$ |
| : Section 7.2 | $\emptyset$ |
| : Section 7.3 | $\emptyset$ |
| : Section 8.2 | $\emptyset$ |
| Tutorial 1 : Section 3.2 | ✗ |
| : Section 4.1 | $\emptyset$ |
| : Section 5.1 | $\emptyset$ |
| : Section 6.1 | $\emptyset$ |
| : Section 7.1 | $\emptyset$ |
| : Section 8.1 | $\emptyset$ |

8 / 20

**Prompt for generating summaries**

Create a summary of the attached slides including the most important intuition, all mathematical formulas, relevant examples, and theorems, but no proofs. Pay special attention to the provided examples, their continuations and modifications.

Be concise and technical using expert vocabulary. Explain in a suitable manner for a master of Logic student familiar with the relevant background but unfamiliar with the discussed material as of yet. Write the summary in typst. The slides are attached. Only focus on content and leave out organizational information about the course. I am pasting all of this into my typst document where each lecture is a level two heading e.g. == Lecture 1, so subchapters have to be at the correct level, at least three e.g. === Core Intuitions and Definitions.

Important: Wrap the generated typst syntax summary in """ to make it copyable

**Notable features of typst syntax:**
1. if there is more than one letter in a name in typst math block then it needs to be wrapped in "".
2. to make text bold, wrap it in singular stars and to make it italic wrap it in underscores
3. If you are more used to different typesetting languages, typst always uses () as parentheses and only uses {} for set notation

**Style guide:**
1. do not include and or [cite: x] in your output
2. I have defined custom functions to represent definitions, theorems ("theorem"), proofs ("proof"), examples ("example"), intuitions [only use this for informal introductions] ("intuition"), warnings to watch out ("attention"), questions ("question"), calls to recall something learned before ("remember"), note something carefully ("note"), and an info ("info").
   - To define a new concept, call `#def("Name of Concept")[Definition body]`
   - For all others call `#callout(title: "Title", style: "style-name")[Box body]
   - Each box generates a tag #label("def-concept-name-hyphenated"). Refer to any concept you reference back to always @def-concept-name-hyphenated
   - use my custom definitions for common operators such as the set of propositional letters or epistemic operators:

**TODO:** For some reason bullet point markers after 1 have no symbols

**TODO:** fix that def title cannot be a `[content block]`

# Week 1

## Session 1-1 (Lecture): Introduction: Motivation, Main Themes, Puzzles

**Motto of Dynamic Epistemic Logic**

*"The wise sees action and knowledge as one. They see truly."* - Bhagavad Gita

### A Core Intuitions and Definitions

**Example 1: Multi-Agent Systems**

1. **Computation**: a network of communicating computers (e.g., 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**: agents ("principals") using a cryptographic protocol to communicate in private
5. **Economics**: transactions in a market
6. **Society**: social activities
7. **Politics**: diplomacy, war
8. **Science**: a community of scientists, engaged in creating theories, making observations and performing experiments to test their theories

**Def 1** (*Properties of Multi-Agent Systems*):
- *dynamic*: Agents perform *actions* which change the system (via interaction)
- *informational*: Agents acquire, store, process, and exchange *information* about each other and the environment

- *Evolving knowledge*: The knowledge an agent has may *change* in time, due to their or other players' actions.
- Certain actions increase information.
- *General rule*: players try to minimize their uncertainty and increase their knowledge.

**Def 2** (*Knowledge*): Truthful information.

**Def 3** (*Justified Belief*): Information that is plausible, well-justified, probable, but possibly false.

**Def 4** (*Belief Revision*): A sustained, dynamic, self-correcting, truth-tracking action. Non-monotonic. True knowledge can only be recovered by effort. Made more difficult by deceit.

**Question**

Is knowledge a form of belief, or is knowledge more fundamental than belief?

**Def 5** (*Uncertainty*): A corollary of imperfect knowledge or "imperfect information".

**Def 6** (*Game of imperfect information*): A game where some moves are hidden, preventing players from knowing everything that is going on; they only have a partial view of the situation.

- An agent may be *uncertain* () about the real situation at a given time: they cannot *distinguish* between possible outcomes.

*Wrong Beliefs*: Agents…
- … may be induced (even with malicious intent e.g., cheating) to acquire false "certainty" in their drive for more knowledge.
- … causing them to "know" things that are not true (e.g., due to bluffing in poker).
- Wrong beliefs are indistinguishable from true beliefs for an agent once they have become "certainty" (they really think they "know").

**Def 7** (*Strategic Ignorance*): It can be advantageous not to know (or pretend not to).

### B Distributed, Nested, and Common Knowledge

**Def 8** (*Distributed Knowledge*): Potential/virtual knowledge that is not reducible to one individual.

Knowledge that is not necessarily held by any individual agent prior to communication, but is known when multiple agents pool their distinct information.

**Example 2: Distributed Knowledge: Business dealings**

- $A$ knows $B$ made a deal with either $C$ or $E$ (exclusively).
- $B$ actually made a deal with $E$, so $C$ knows $B$ did **not** go make a deal with them.
- Neither $A$ nor $C$ individually know $B$ made a deal with $E$ before communicating.
- If $A$ and $C$ communicate (pool their knowledge), they deduce the truth. The fact is *distributed knowledge* among them.

**Def 9** (*Nested Knowledge*): Knowledge about the knowledge of others, leading to potential infinite regress or deep epistemic reasoning (e.g., "how can you know that I do not know?").

**Def 10** (*Introspection*): An agent's capability (or lack thereof) to reason about their own epistemic state.
- **Known knowns**: things we know we know.
- **Known unknowns**: things we know we do not know.
- **Unknown unknowns**: things we do not know that we do not know.

**Def 11** (*Common Knowledge*): A condition where an entire group knows a fact, everybody knows that everybody knows it, and everybody knows that everybody knows that everybody knows it, ad infinitum.

### Example 3: Common Knowledge vs. 'Everybody Knows'

- Suppose everybody knows the road rules (e.g., red means "stop") and respects them.
- **Question**: Is this enough to drive safely? **No**.
- **Reasoning**: Merely knowing the rule is insufficient if you lack the certainty that **others** know the rules and will abide by them.
- **Resolution**: Safe driving requires the rules to be *Common Knowledge* (Def 11).

## Session 1-2 (Lecture): Main Themes, Puzzles, and Paradoxes Continued

### A Epistemic Puzzles and Paradoxes

### Example 4: Puzzle 0: The Coordinated Attack

Two army divisions (A and B) must attack simultaneously to win. They communicate via messengers over a channel where messages might be captured.

- A sends "attack at dawn" and B receives it.
- B must acknowledge receipt, but A does not know if the acknowledgment will arrive.
- A must acknowledge the acknowledgment, ad infinitum.

**Result**: No finite sequence of successful message deliveries can achieve coordination.

### Remember: Fixpoints and Byzantine Generals

**Def 12** (*Fixpoint*): $x$ is a fixpoint iff $f : X \to X; x = f(x)$.

In the case of Puzzle 0:

$$C\Box\varphi \equiv K_A C\Box\varphi \wedge K_B C\Box\varphi \tag{1}$$

Where $K_X$ is the knowledge operator of agent $X$, $C\Box$ is the common knowledge operator, $\varphi$ is the message about the attack time.

### Intuition: Coordinated Attack Intuition

Achieving *Common Knowledge* (Def 11) over an unreliable communication channel is logically impossible in a finite number of steps. Unbounded nested knowledge (Def 9) does not equate to true common knowledge.

**Example 5: Puzzle 1: To Learn is to Falsify**

$A$ sends an email to her lover $C$: "$B$ doesn't know about us."

$B$ secretly intercepts and reads it.

**Result**: The proposition was true right before reading, but the act of learning the message immediately falsifies it (a dynamic variant of Moore's Paradox).

**Note: Instantaneous truth value change**

**Paradox**: usually learning $\varphi$ means believing phi $\Box\varphi$, but here reading $\varphi$ leads to not believing $\varphi$: $\Box\neg\varphi$.

**Less paradoxical with dynamic thinking**: The truth value of the statement changes instantaneously when $B$ reads and accepts it.

**Attention: Non-standard Belief Revision**

Standard belief-revision postulates (e.g., AGM) fail for complex learning actions where the informational payload refers directly to the epistemic state of the receiver.

**Example 6: Puzzle 2 & 3: Self-Fulfilling and Self-Enabling Falsehoods**

- **Self-Fulfilling**: $A$ falsely believes $B$ knows about her affair and sends a warning message. $B$ intercepts it and thereby learns of the affair. Communicating a false belief makes it true.

  "$B$ doesn't know about us."

- **Self-Enabling**: $C$ (wanting to seduce faithful $A$) forges a message to himself from $A$ saying $B$ knows they are having an affair. $B$ reads it and divorces $A$. $A$, on the rebound, starts an affair with $C$. The transmission of a falsehood causally enables its own validation.

### B The Muddy Children and Epistemic Updates

**Example 7: Puzzle 4: Muddy Children**

4 perfect logicians (children), exactly 3 have dirty faces. They see others but not themselves.
- Father publicly announces: "At least one of you is dirty."
- Father iteratively asks: "Do you know if you are dirty or not?"
- Children answer publicly and simultaneously based strictly on their knowledge without guessing.

**Result**: For 2 rounds, they answer in the negative. In the 3rd round, all 3 dirty children confidently state they are dirty. In the 4th round, the clean child deduces they are clean.

### Socratic Questioning

Discovering answers by asking questions of students. (Wikipedia)

### Intuition: Muddy Children

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

The initial announcement transforms distributed implicit knowledge into public *Common Knowledge* (Def 11).

2. *What's the point of the father's repeated questions?*

The iterated Socratic questioning acts as sequential epistemic updates: public statements of ignorance incrementally eliminate possible worlds in the Kripke model until the true state is uniquely isolated.

### Example 8: Modifications of Muddy Children

- **The Amazon Island**: Isomorphic to Muddy Children. A law mandates wifes to execute their cheating husbands at noon once discovered. Queen announces at least one cheater exists and if somebody's husband is cheating, all other wives know it. With 17 cheaters, for 16 days nothing happens, and all 17 are shot on day 17.
- **The Dangers of Mercy**: Wives of the 17 cheaters secretly decide to spare them, while others believe strict obedience to the law is common knowledge. No shots are fired on day 17. On day 18, all faithful husbands are erroneously shot by their wives, who logically deduce (from flawed public premises) that their husbands must be cheating.

### Example 9: Puzzle 5: Sneaky Children

Children are incentivized for speed and punished for errors. After round 1, two dirty children cheat by secretly confirming to each other they are dirty, thus answering "I know" prematurely in round 2.

- **Honest Children Always Suffer**: The 3rd dirty child logically deduces it must be clean, answers incorrectly in round 3, and is punished.
- **Clean Children Always Go Crazy**: The 4th (clean) child faces a strict contradiction. If it blindly applies monotonic updates via classical logic, it undergoes logical explosion (believing everything).

### C Paradoxes of Induction and Probability

#### Example 10: Puzzle 6: Surprise Exam

Teacher announces an exam next week, but the date will be a surprise (students won't even know the night before).

- **Paradoxical Argumentation**: Students apply backward induction. It cannot be Friday (they'd know Thursday night). By elimination, it cannot be any day. They deduce the announcement is false.
- **Result**: They dismiss the announcement. The exam occurs (e.g., Tuesday) and is indeed a complete surprise.

#### Example 11: Puzzle 7: The Lottery Paradox

A fair lottery with 1,000,000 tickets.

- Probability of ticket $x$ winning is 0.000001.
- It is rational to hold the belief that ticket $x$ will lose.
- This reasoning applies symmetrically to all tickets.
- Yet, the agent knows one ticket will win.

**Result**: The conjunction of highly probable rational beliefs yields a strict logical **inconsistency**.

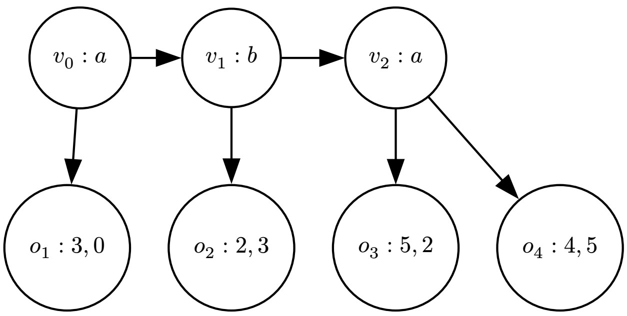
#### Example 12: Puzzle 7 Modification: The Infinite Lottery

An infinite lottery over arbitrary natural numbers. The probability of any given ticket winning is exactly 0. The agent is mathematically correct to believe a specific ticket will not win, yet one must win. Any finite subset of beliefs is consistent, but the infinite global set is inconsistent.

### D Backward Induction and Social Epistemology

#### Example 13: Puzzle 8: The Centipede Game

A sequential game with alternating moves by $a$ and $b$, deciding between stopping the game or continuing:



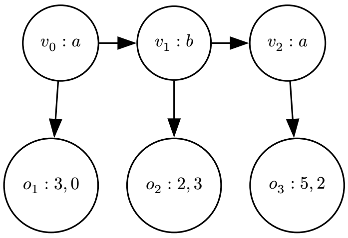

In the leaves ("outcomes" $o_j$) the first number is $a$'s payoff, the second number is $b$'s payoff.

- $v_0 : a$ stops for $o_1(3, 0)$ or continues to $v_1$
- $v_1 : b$ stops for $o_2(2, 3)$ or continues to $v_2$
- $v_2 : a$ stops for $o_3(5, 2)$ or continues to $o_4(4, 5)$

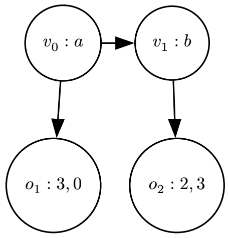
**The Backwards Induction (BI) Method**

- Iteratively eliminate the *obviously* “bad” moves
- Proceeding backwards from the leaves

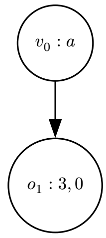
**Elimination Step 1**




**Elimination Step 2**




**Elimination Step 3**




- **BI outcome**: $o_1 : 3, 0$
- *Why not another outcome?*: Strikes many as irrational

**Intuition: The BI Paradox and Rational Pessimism**

- **Aumann’s Argument**: Assuming *Common Knowledge* (Def 11) of Rationality (CKR), backward induction dictates $A$ chooses $o_3$ at $v_2$, so $B$ chooses $o_2$ at $v_1$, so $A$ chooses $o_1$ at $v_0$. The game terminates immediately at a suboptimal Pareto outcome.
- **Counterargument**: If $B$ reaches $v_1$, he observes $A$ violating CKR (she didn’t stop at $v_0$). If $B$ adopts **Rational Pessimism**—assuming $A$ is irrational and will thus choose $o_4$ at $v_2$—he should continue. If $A$ anticipates this belief revision, her initial deviation becomes strictly rational. The epistemic foundation of backward induction contradicts its own counterfactuals.

**E Social Epistemology**

Group dynamics often deviate from ideal individualized epistemic logic due to the recursive nature of social evidence.

**Def 13** (*Pluralistic Ignorance*): A situation where the group collectively knows or acts upon less information than the individuals possess privately. Often observed in totalitarian regimes where public behavior contradicts private beliefs.

**Example 14: Puzzle 9: Wisdom vs. Madness of the Crowds**

- **Wisdom of the Crowds**: Distributed group knowledge often empirically exceeds the most expert individual (e.g., aggregating independent estimates).
- **Madness of the Crowds**: Systems can fail systematically due to cascading social epistemology.

**Intuition: Information Cascades**

An information cascade occurs when agents base their decisions on the observable behavior of prior agents rather than their own private evidence, leading to a breakdown of *epistemic democracy* (the wisdom of crowds).

### Example 15: The Black and White Urn Problem

**Setup:** One urn is in a room. It is either Urn B (2/3 black marbles) or Urn W (2/3 white marbles). Agents enter one by one, draw a marble, replace it, and publicly record their guess of the urn on a blackboard. **The Cascade:** 1. Voter 1 draws Black and guesses Urn B.

2. Voter 2 draws Black and guesses Urn B.
3. Voter 3 draws White. However, the public evidence (two B votes) combined with their private evidence (one W draw) yields an aggregate evidence of (B, B, W). The rational epistemic choice is still to guess Urn B.

**Result:** From Voter 3 onwards, everyone will vote Urn B regardless of their private draw. If the first two voters happened to draw the minority color (probability $\frac{1}{9}$), the entire crowd of $n$ voters will lock into the wrong conclusion.

### Example 16: Biological and Geopolitical Cascades

- **Army Ant Circular Mill:** If an army ant loses the pheromone trail, it is biologically programmed to follow the ant directly in front of it. This simple rule works locally but can result in a massive recursive loop (a death spiral up to 400m in diameter) where the ants walk in a circle until they die.
- **The Men Who Stare at Goats (Cold War):** A French newspaper published a fabricated story about US military research into psychic weapons. Soviet intelligence read this, assumed it was a cover-up, and initiated their own psychic research program. US intelligence discovered the Soviet program and, assuming the Soviets were onto a real threat, started their own actual research program, sparking a 30-year arms race built on an initial cascade of false information.

## Session 1-3 (Lecture): Single-Agent Epistemic-Doxastic Logics: Kripke Models

### A Syntax and Core Definitions

Single-agent epistemic-doxastic logic expands standard propositional logic to formally capture an agent's knowledge and beliefs.

$$\varphi ::= p \mid \neg\varphi \mid \varphi \wedge \varphi \mid \mathrm{K}\varphi \mid \mathrm{B}\varphi \tag{2}$$

where $p \in \mathrm{Prop}$.

**Def 14** (*Single-Agent, pointed Epistemic-Doxastic Model*): Is a tuple $\mathbf{S} = (S, S_0, \|\cdot\|, s_\star)$, where

- $S$: A set of *ontic* states defining the agent's *epistemic state* (epistemically possible).
- $S_0$: A non-empty subset $S_0 \subseteq S$, called the *sphere of beliefs* or the agent's *doxastic state.*
- $\|\cdot\| : \mathrm{Prop} \to \mathcal{P}(S)$: A *valuation* map assigning atomic propositions to sets of states.
- $s_\star \in S$: The designated "actual world" representing the real state of the world.

*Sphere-based*: represents beliefs as nested layers of possible worlds, ranking worlds by their plausibility

## B Semantics

**Intuition: Interpretation**

- **Epistemic state**: state of the agent's knowledge: they belief $s_\star$ is among $S$, but cannot distinguish between $s_i, s_j \in S; i \neq j$.
- **Doxastic state**: the agent beliefs $s_\star \in S_0$

**Notation: Truth**

We write the following if $\varphi$ is *true* in world $w$. When the model $\mathbf{S}$ is fixed, we skip the subscript.

$$w \underset{\mathbf{S}}{\vDash} \varphi \tag{3}$$

**Note: Atomic logical connectives**

We interpret negation $\neg$ and conjunction $\wedge$ as atomic logical connectives, but disjunction $\vee$, the conditional $\rightarrow$, and the biconditional $\leftrightarrow$ as compound connectives.

**Def 15** (*Truth in an Interpretation*): A sentence $\varphi$ is true in a model $\mathbf{S}$ under the valuation map $\|\cdot\|_{\mathbf{S}}$ if

i. $\varphi = p; p \in \mathrm{Prop}$: $w \vDash p$ iff $w \in \|p\|$,
ii. $\varphi = \neg\psi$: $w \vDash \neg\psi$ iff $w \nvDash \psi$,
iii. $\varphi = \psi \wedge \chi$: $w \vDash \psi \wedge \chi$ iff $w \vDash \psi$ and $w \vDash \chi$,
iv. $\varphi = \mathrm{K}\varphi$: $w \vDash \mathrm{K}\varphi$ iff $\forall s \in S, s \vDash \varphi$,
v. $\varphi = \mathrm{B}\varphi$: $w \vDash \mathrm{K}\varphi$ iff $\forall s \in S_0, s \vDash \varphi$.

**Def 16** (*Validity*): A sentence $\varphi$ is **valid** in a model $\mathbf{S}$ if it is true at every state $w \in \mathbf{S}$.

**Def 17** (*Satisfiability*): A sentence $\varphi$ is **satisfiable** in a model $\mathbf{S}$ if it is true at some state $w \in \mathbf{S}$.

**Note: Semantics of Knowledge and Belief**

The universal quantifier over the domain of possibilities is interpreted as knowledge or belief.

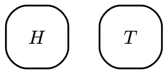
- **Knowledge** ($\mathrm{K}\varphi$): Truth in all epistemically possible worlds.
- **Belief** ($\mathrm{B}\varphi$): Truth in all doxastically possible worlds within the sphere of beliefs.

## C Learning and Mistaken Updates

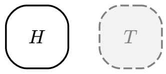
Learning corresponds to world elimination. An update with a sentence $\varphi$ is the operation of deleting all non-$\varphi$ possibilities from the model.

### Example 17: The Concealed Coin and Mistaken Updates

**Base Scenario:** A coin is on the table; the agent does not know if it is Heads ($H$) or Tails ($T$).



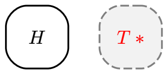

**Standard Update:** The agent looks and sees $H$. The $T$ world is eliminated, and only the $H$ epistemic possibility survives.



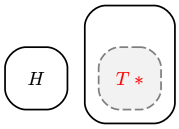

#### Note: Update as World Elimination

In general, updating corresponds to world elimination: an update with a sentence $\varphi$ is simply the operation of deleting all the non-$\varphi$ possibilities.

**Mistaken Update:** The agent mistakenly believes they saw $H$. If we eliminate $T$, the actual world $s_\star$ is no longer in the agent's model, making it impossible to evaluate objective truth.

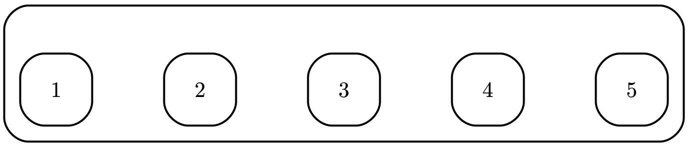



**Resolution (Third-Person Models):** We maintain an objective perspective where the real possibility always remains in the global model $S$, even if the agent believes it to be impossible. The sphere of beliefs $S_0$ ($\boxdot$) is restricted to $T$, meaning the agent believes $H$, but their belief is false because $s_\star \in S/S_0$.

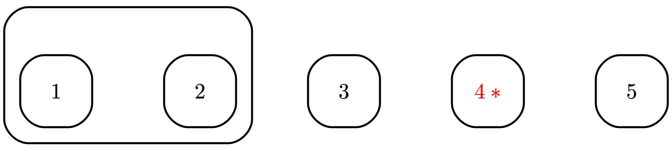



### Example 18: Continuation of Puzzle 6: Surprise Exam

Situation before the teacher's announcement (we don't know $s_\star$: no star in the figure):




A student beliefs (for some reason) the exam will be on Monday or Tuesday, but it is on Thursday ($s_\star = 4$):



### D Kripke Semantics for Epistemic-Doxastic Logic

Sphere models can be generalized using Kripke semantics to allow for varying strengths of knowledge and belief.

**Remember: Kripke Model**

**Def 18** (*Kripke Model*): A Kripke model is a tuple $\mathbf{S} = \left(S, \{R_i\}_{i \in I}, \|.\|, s_\star\right)$ with set of states $S$, accessibility relations $R_i$, valuation $\|\cdot\|$, and actual state $s_\star$.

**Def 19** (*Epistemic-Doxastic Kripke Model*): To model knowledge K and belief B, this becomes $(S, \sim, \to, \|.\|, s_\star)$, where $\sim$ is the epistemic relation (for K) and $\to$ is the doxastic relation (for B).

For atomic sentences and for Boolean connectives, we use the same semantics (and notations) as on epistemic-doxastic models.

**Def 20** (*Kripke modalities*): For every sentence $\varphi$, we can define a new sentence using the *universal Kripke modality* $[R_i]$ by universally quantifying over $R_i$ accessible worlds. The dual *existential Kripke modality* $\langle R_i \rangle$ is given by

$$\langle R_i \rangle \varphi := \neg [R_i] \neg \varphi. \tag{4}$$

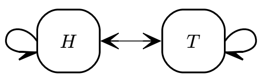
**Notation: Kripke modalities: subscript**

If $R$ is unique, we abbreviate $[R_i]\varphi$ as $\Box\varphi$, and $\langle R_i \rangle\varphi$ as $\Diamond\varphi$.

**Def 21** (*Truth in an interpretation continued: Kripke modalities*): We continue Def 15 by adding vi. $\varphi = [R_i]\varphi$: $w \vDash [R_i]\varphi$ iff $v \vDash \varphi \forall v : wR_iv$.

**Example 19: Example 17: Concealed coin continued**

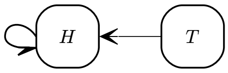
The agent's knowledge in the concealed coin scenario can be represented as:




- The arrows represent the **epistemic relation** $\sim$, capturing the agent's uncertainty about the state of the world.
- An arrow from state $s$ to state $t$ means that if $s_\star = s$, the agent could not distinguish between $s$ and $t$.

### Example 20: Example 17: Concealed coin continued

The agent's belief after the mistaken update can be represented as:




- The arrows represent the **epistemic relation** $\sim$, capturing the agent's uncertainty about the state of the world.
- An arrow from state $s$ to state $t$ means that if $s_\star = s$, the agent could not distinguish between $s$ and $t$.

### Remember: Named axioms in Modal Logic

Certain axioms have set names in Modal Logic:

- (**K**) Basic Modal Logic
- (**T**) Reflexivity $\Box\varphi \to \varphi$
- (**4**) Transitivity $\Box\varphi \to \Box\Box\varphi$
- (**5**) Euclideaness $\Diamond\varphi \to \Box\Diamond\varphi$
- (**D**) Seriality $\Box\varphi \to \Diamond\varphi$
- *Weak Epistemic Model* (**S4**) = (**K**) + (**T**) + (**4**): No negatice introspection (only relfex. & trans.)
- *Epistemic Model*: (**S5**) = (**K**) + (**T**) + (**5**)
  - Note: An (**S5**)-model is one where the accessibility relations are equivalence relations: reflexive, transitive, symmetric (with the other two properties, Symmetry $\equiv$ Euclidean)
- *Doxastic Model*: (**KD45**) = (**K**) + (**D**) + (**4**) + (**5**)
  - Note: Doxastic Models are not symmetric, but serial ($\forall s : \exists t : s \to t$), transitive, Euclidean

### Theorem 1: Axioms and Relational Properties

A Kripke model satisfying all the below conditions on the relations $\sim$ and $\to$ is called an **epistemic-doxastic Kripke model**.

Validities for Knowledge (Equivalence relation $\sim$, giving an S5 model):

i. **Veracity** ($\mathrm{K}\varphi \Rightarrow \varphi$): $\sim$ is reflexive.
ii. **Positive Introspection** ($\mathrm{K}\varphi \Rightarrow \mathrm{KK}\varphi$): $\sim$ is transitive (**4**).
iii. **Negative Introspection** ($\neg\mathrm{K}\varphi \Rightarrow \mathrm{K}\neg\mathrm{K}\varphi$): $\sim$ is Euclidean (and symmetric) (**5**).

Validities for Belief (KD45 model properties for $\to$):

iv. **Consistency** ($\neg B(\varphi \wedge \neg\varphi)$): $\to$ is serial.
v. **Positive Introspection** ($\mathrm{B}\varphi \Rightarrow \mathrm{BB}\varphi$): $\to$ is transitive.
vi. **Negative Introspection** ($\neg\mathrm{B}\varphi \Rightarrow \mathrm{B}\ \neg\mathrm{B}\varphi$): $\to$ is Euclidean.

KB Interaction Properties:

vii. **Knowledge implies Belief** ($\mathrm{K}\varphi \Rightarrow \mathrm{B}\varphi$): If $s \to t$ then $s \sim t$.
viii. **Strong Positive Introspection** ($\mathrm{B}\varphi \Rightarrow \mathrm{KB}\varphi$): If $s \sim t$ and $t \to w$ then $s \to w$.
ix. **Strong Negative Introspection** ($\neg\mathrm{B}\varphi \Rightarrow \mathrm{K}\neg\mathrm{B}\varphi$): If $s \sim t$ and $t \to w$ then $s \to w$.

**Observations**:

1. Epistemic-doxastic Kripke models are equivalent to Simple Epistemic-Doxastic Models(Def 14)

2. The epistemic relation is completely determined by the doxastic relation.

**Sound and Complete Proof System for single agent epistemic-doxastic logic**:

- Axioms:
  - From above: i. - iv., vii. - ix.
  - All propositional tautologies
  - Modus Ponens: from $\varphi$ and $(\varphi \rightarrow \psi)$ infer $\psi$
  - Necessitation: from $\varphi$ infer $\mathrm{K}\varphi$ and $\mathrm{B}\varphi$
  - Kripke's axioms for K and B:
    - $(\mathrm{K}\varphi \rightarrow \psi) \rightarrow (\mathrm{K}\varphi \rightarrow \mathrm{K}\psi)$
    - $(\mathrm{B}\varphi \rightarrow \psi) \rightarrow (\mathrm{B}\varphi \rightarrow \mathrm{B}\psi)$

**Generalization**

- It is convenient to have a more general semantics where the above do not hold
  - Introspection is not universally accepted
  - People may believe they know things they don't actually know
  - There might be "crazy" agents with inconsistent beliefs

**Theorem 2: Equivalence of Models**

Every epistemic-doxastic sphere model $S = (S, S_0, \|.\|, s_\star)$ is completely equivalent to an epistemic-doxastic Kripke model $S' = (S, \sim, \rightarrow, \|.\|, s_\star)$ that satisfies the same sentences at $s_\star$.

**Attention: Logical Omniscience**

Any Kripke modality validates axiom K ($K(\varphi \Rightarrow \psi) \Rightarrow (\mathrm{K}\varphi \Rightarrow K\psi)$) and the Necessitation rule (if $\varphi$ is valid, $\mathrm{K}\varphi$ is valid). Consequently, Kripke semantics models "ideal reasoners" with unlimited inference powers who know/believe all logical entailments, failing to capture bounded rationality.

# Week 2

## Session 2-1 (Lecture): Multi-agent Models and Public Announcement Logic (PAL)

### A Multi-Agent Kripke Models & Modalities

**Def 22** (*Multi-Agent Kripke Model*): A multi-agent Kripke model is a tuple

$$\mathbf{S} = \left(S, \left\{\underset{a}{\rightarrow}\right\}_{a \in \mathcal{A}}, \|\cdot\|\right) \tag{5}$$

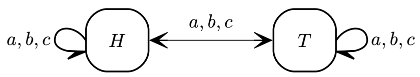
where $\mathcal{A}$ is a set of labels representing the names of epistemic agents.

**Def 23** (*Epistemic/ Doxastic Modalities*): For every sentence $\varphi$, we define $\Box_a \varphi$ by universally quantifying over $\underset{a}{\rightarrow}$-accessible worlds: $s \underset{S}{\models} \Box_a \varphi \Leftrightarrow t \underset{S}{\models} \varphi$ for all $t$ such that $s \underset{a}{\rightarrow} t$. This is interpreted as knowledge, denoted $\mathrm{K_a}\varphi$, or belief, denoted $\mathrm{B_a}\varphi$.

Its existential dual $\Diamond_a \varphi := \neg \Box_a \neg \varphi$ denotes epistemic/ doxastic possibility.

**Example 21: The Concealed Coin**

Two players $a$, $b$, along with a referee $c$ play a game. The referee throws a fair coin so nobody knows the outcome.




Using concatenated arrows, we can express iterated knowledge. For instance, $b$ knows that $a$ does not know the outcome but knows it is Heads ($H$) or Tails ($T$):

$$w \models \Box_b(\neg \Box_a H \wedge \neg \Box_a T) \wedge \Box_b \Box_a (H \vee T) \tag{6}$$

### B Common Knowledge

**Def 24** (*Common Knowledge (Group)*): Common knowledge within a group $G \subseteq \mathcal{A}$, denoted $C\Box_{\mathrm{G}}\varphi$, is evaluated by quantifying over all worlds accessible by any finite concatenation of arrows within $G$:

$$s \underset{S}{\models} C\Box_{\mathrm{G}} \varphi \Leftrightarrow t \underset{S}{\models} \varphi \tag{7}$$

for every $t$ and every finite chain $s = s_0 \underset{a_1}{\rightarrow} s_1 \underset{a_2}{\rightarrow} \ldots \underset{a_n}{\rightarrow} s_n = t$ with $a_1, \ldots, a_n \in G$.

**Notation: Common Knowledge**

**Full common knowledge**: In the case that $G = \mathcal{A}$, we omit the subscript and write $C\Box\varphi$. **Knowledge/ Belief**: In epistemic-doxastic models we have both

- *common knowledge* $Ck$ and
- *common true belief* $Cb$.

**Common Knowledge equivalence to Kripke Modality**

$C\Box_G\varphi$ is equivalent to the Kripke modality for the reflexive-transitive closure of the union of all epistemic relations: $\left[\left(\bigcup_{a\in G} \xrightarrow[a]{}\right)^*\right]$.

**Remember**

**Def 25** (*Reflexive-transitive closure*): Given a relation $R$, its *reflexive-transitive closure* $R^*$ is defined by:

$$wR^*v \text{ iff } \exists \text{ finite chain (length } n \geq 0) : w = w_0Rw_1R...Rw_n = v \tag{8}$$

**Intuition: Common Knowledge as Infinite Conjunction**

Let $E_G\varphi := \bigwedge_{a\in G} \Box_a\varphi$ ("everybody in group $G$ knows $\varphi$").

Then, common knowledge $C\Box_{\mathrm{G}}\varphi$ (Def 24) is semantically equivalent to the infinite conjunction:

$$\varphi \wedge E_G\varphi \wedge E_GE_G\varphi \wedge ... \tag{9}$$

**Attention: Infinitary definitions**

The most used modal-epistemic languages are *finitary* s.t. $C\Box_{\mathrm{G}}$ cannot be defined as the infinite conjunction, which is impossible to form.

Instead, $C\Box_{\mathrm{G}}$ is interpreted as a **primitive** operator induced by the semantic clause in Info 2

**Theorem 3: Validities for Common Modalities**

- **Fixed-Point Axiom** (Mix): $C\Box_{\mathrm{G}}\varphi \Rightarrow (\varphi \wedge E_GC\Box_{\mathrm{G}}\varphi)$
- **Induction Axiom**: $C\Box_{\mathrm{G}}(\varphi \Rightarrow E_G\varphi) \Rightarrow (\varphi \Rightarrow C\Box_{\mathrm{G}}\varphi)$

### B.1 Syntax

**Epistemic logic with common knowledge**

$$\varphi ::= p \mid \neg\varphi \mid \varphi \wedge \varphi \mid \mathrm{K}_a\varphi \mid Ck_{\mathrm{G}}\varphi \tag{10}$$

**Doxastic logic with common true belief**

$$\varphi ::= p \mid \neg\varphi \mid \varphi \wedge \varphi \mid \mathrm{B}_a\varphi \mid Cb_{\mathrm{G}}\varphi \tag{11}$$

**Epistemic-doxastic logic with common knowledge and common (true) belief**

$$\varphi ::= p \mid \neg\varphi \mid \varphi \wedge \varphi \mid \mathrm{K}_a\varphi \mid \mathrm{B}_a\varphi \mid Ck_{\mathrm{G}}\varphi \mid Cb_{\mathrm{G}}\varphi \tag{12}$$

**Complete axiomatization:**

- Multi-agent versions of the axioms in Theorem 1 (modalities labeled with agents)
- Fixed-Point and Induction Axioms (Theorem 3) for both $Ck_{\mathrm{G}}$ and $Cb_{\mathrm{G}}$
- Kripke axioms for both $Ck_{\mathrm{G}}$ and $Cb_{\mathrm{G}}$: $C\Box_{\mathrm{G}}(\varphi \to \psi) \to (C\Box_{\mathrm{G}}\varphi \to C\Box_{\mathrm{G}}\psi)$
- Necessitation for both $Ck_{\mathrm{G}}$ and $Cb_{\mathrm{G}}$: $\varphi \to C\Box_{\mathrm{G}}\varphi$

## C Distributed Knowledge

**Def 26** (*Distributed Knowledge (Group)*): Distributed knowledge within a group $G$, denoted $D\Box_{\mathrm{G}}\varphi$ or $Dk_{\mathrm{G}}\varphi$, is obtained by quantifying over all worlds simultaneously accessible by all arrows for agents in $G$: $s \underset{S}{\vDash} D\Box_G\varphi \Leftrightarrow t \underset{S}{\vDash} \varphi$ for every $t$ such that $s \underset{a}{\rightarrow} t$ holds for all $a \in G$.

### Intuition: Epistemic Potential

Def 26 captures the implicit (or virtual) knowledge of the group: what the agents in $G$ could come to know if they pooled all their private knowledge.

### Distributed Knowledge equivalence to Kripke Modality

$D\Box_G$ is equivalent to the Kripke modality corresponding to the intersection of epistemic relations

$$\bigcap_{a\in G} \underset{a}{\rightarrow} \tag{13}$$

### Note: Interpretations of distributed modalities

- when the relations $\underset{a}{\rightarrow}$ are reflexive: $Dk_G$ as some sort of distributed knowledge
- when the relations $\underset{a}{\rightarrow}$ represent beliefs: $Db_{\mathrm{G}}$ may be inconsistent

### Example 22: Two Muddy Children

Two children (1 & 2) have dirty foreheads ($d_1$, $d_2$). Each sees the other but not themselves. In the real world $w = (d_1, d_2)$, neither knows both are dirty, but it is distributed knowledge: $w \vDash \neg \mathrm{K}_1 d_1 \wedge d_2 \wedge \neg \mathrm{K}_2 d_1 \wedge d_2 \wedge Dk(d_1 \wedge d_2)$.

### Theorem 4: Validities for Distributed Knowledge

- $\mathrm{K_a}\varphi \Rightarrow Dk\varphi$
- $(\mathrm{K_a}\varphi \wedge \mathrm{K_b}\psi) \Rightarrow Dk(\varphi \wedge \psi)$

## D Dynamics & Public Announcements

**Def 27** (*Public Announcement Logic*): A public announcement $!\varphi$ is a joint update that deletes all non-$\varphi$ worlds from a model. The model transformer maps $S$ to $S^{!\varphi} = \left(S_\varphi, \underset{\varphi}{\rightarrow}, \|\cdot\|_\varphi\right)$, where $S_\varphi = \|\varphi\|_S$, while the relations as well as valuations are restricted to $S_\varphi$. The dynamic modality is evaluated as: $s \underset{S}{\vDash} [!\varphi]\psi \Leftrightarrow t \underset{S^{!\varphi}}{\vDash} \psi$ for all $t \in S^{!\varphi}$ such that $s \overset{!\varphi}{\underset{S}{\rightarrow}} t$.

### Theorem 5: PAL Reduction Axioms

PAL (Def 27) allows translating dynamic formulas to basic modal logic via reduction axioms:
- **Atomic Permanence**: $[!\varphi]p \Leftrightarrow (\varphi \Rightarrow p)$
- **Announcement-Negation**: $[!\varphi]\neg\psi \Leftrightarrow (\varphi \Rightarrow \neg[!\varphi]\psi)$
- **Announcement-Conjunction**: $[!\varphi](\psi_1 \wedge \psi_2) \Leftrightarrow ([!\varphi]\psi_1 \wedge [!\varphi]\psi_2)$
- **Announcement-Knowledge**: $[!\varphi]\Box_a\psi \Leftrightarrow (\varphi \Rightarrow \Box_a[!\varphi]\psi)$

### Expressivity & Succinctness

PAL has the exact same expressivity as basic modal logic because dynamic modalities can be eliminated using the reduction axioms. However, PAL is exponentially more succinct.

### E Moore Sentences & Paradoxes

**Def 28** (*Moore Sentences*): Sentences that become false after being truthfully announced. For a Moore sentence $\varphi$, we have $[!\varphi]\neg\varphi$. Therefore, they become known to be false after being announced: $[!\varphi]\mathrm{K_a}\neg\varphi$.

### Example 23: Moore Sentences in Muddy Children

"You are dirty but you do not know it" is a Moore sentence for child 1: $d_1 \wedge \neg \mathrm{K}_1 d_1$. This is initially true in the muddy children model. However, after it is publicly announced, child 1 learns they are dirty. The second conjunct becomes false, making the entire sentence false.

*Continuation*: In the grand finale of the muddy children puzzle, children repeatedly announce their ignorance: $!(\bigwedge_i (\neg\mathrm{K_i}d_i \wedge \neg\mathrm{K_i}\neg d_i))$. Each round deletes worlds where a child would have known their state, converting distributed knowledge into common knowledge until the dirty children deduce they are dirty.

### Theorem 6: Closure under Composition

Performing two successive public announcements is equivalent to a single, more complex announcement: $[!\varphi][!\psi]\theta \Leftrightarrow [!(\varphi \wedge [!\varphi]\psi)]\theta$.

**TODO**: Continue here [click]

## Session 2-2 (Lecture): PAL Continued

## Session 2-3 (Lecture): Does this one even exist??

# Week 3

## Session 3-1 (Lecture): "Learnability" and "Knowability"

## Session 3-2 (Lecture): Tutorial 1

## Session 3-3 (Lecture): The problem of belief revision

# Week 4

## Session 4-1 (Lecture): Cheating and the Failure of Standard DEL

Based on `DEL 2019-20 Lectures 4.2.pdf`

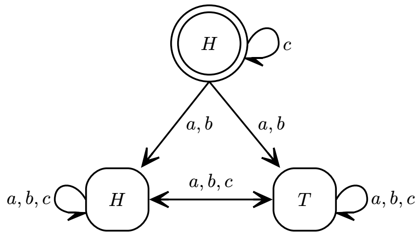
### A The Failure of Standard DEL

**⚠ Attention: DEL Failure: The Problem with Standard Updates**

Standard Dynamic Epistemic Logic (DEL) update mechanisms fail when an agent is confronted with new information that contradicts their previously held *false* beliefs. Under the standard update product, all doxastic relations originating from the real world are eliminated. This empty sphere of beliefs results in the agent believing everything (inconsistent beliefs), violating the consistency axiom (D).

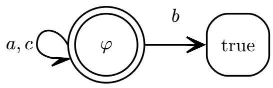
**Example 24: Counterexample: Scenario 4**

*Scenario 4*: Recall the state model immediately after taking a peak:

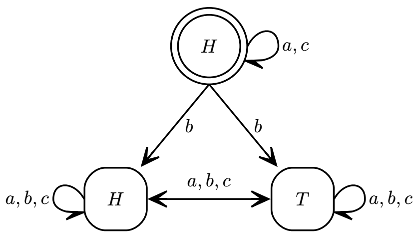



$c$ privately knows that the coin lies heads up: $\varphi = \mathrm{K}_c H$ (also: $\neg \mathrm{K}_{a,b}\varphi$).

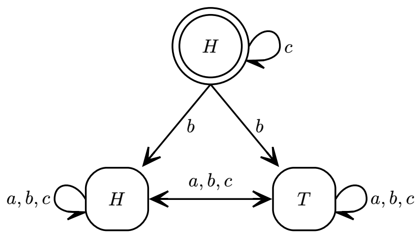
*Scenario 5*: $c$ sends a secret announcement to $a$: $!_{c,a}\varphi$, the event model is:




Intuitive updated state model:

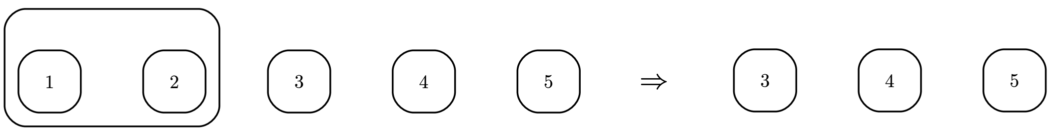



Actual updated state model:




**Example 25: Surprise Exam continued**

If an agent strongly believes an exam is on Monday or Tuesday, and updates with $!(\neg 1)$ (not Monday) and then $!(\neg 2)$ (not Tuesday), the resulting model has no states left in their belief sphere $S_0 = \emptyset$. The agent has inconsistent beliefs, violating axiom (**D**) Consistency of Beliefs.



### Example 26: Newton

*It gets worse*: Agent $a$ used to believe that Newton was the first to discover the laws of gravitation ($p$) after being inspired by being hit on the head by a falling apple ($q$). Belief set: $T = \{p, q, (p \wedge q)\}$

$a$ learned this was a myth: $\neg(p \wedge q)$. Belief set $T = \{p, q, \neg(p \wedge q)\}$. Inconsistent!

$a$ needs to remove $p$ or $q$ from the belief set, logic cannot tell $a$ which.

**B Belief Revision and AGM Theory**

We can fix our update with **Belief Revision Theory**.

### The Problem of Belief Revision

What happens if an agent $a$ learns a new fact $\varphi$ that contradicts previous beliefs?

$a$ has to give up some previous beliefs. But which of them? All of them?

No, $a$ should try to maintain as many previous beliefs as possible, while still accepting the new fact $\varphi$ and without arriving at a contradiction.

### Intuition: AGM Theory Intuition

Standard Belief Revision Theory (AGM[1]) attempts to solve this via an axiomatic approach on theories (belief sets) $T$.

Given input $\varphi$, AGM defines:
- *Expansion operator* $T + \varphi$: $T \cup \{\varphi\}$ closed under logical inference (which can be inconsistent if the new information contradicts $T$).
- *Revision operator* $T * \varphi$: maintains consistency: only adds consistent inference results

### Note: Standard AGM fails to capture higher-order beliefs.

**Def 29** (*AGM Postulates for Belief Revision*): Let $T$ be a theory and $\varphi, \psi$ be formulas. The AGM revision operator $*$ satisfies:
1. **Closure:** $T * \varphi$ is a belief set.
2. **Success:** $\varphi \in T * \varphi$.
3. **Inclusion:** $T * \varphi \subseteq T + \varphi$.
4. **Preservation:** If $\neg\varphi \notin T$ then $T + \varphi \subseteq T * \varphi$.
5. **Vacuity:** $T * \varphi$ is inconsistent iff $\vdash \neg\varphi$.
6. **Extensionality:** If $\vdash \varphi \leftrightarrow \psi$, then $T * \varphi = T * \psi$.
7. **Subexpansion:** $T * (\varphi \wedge \psi) \subseteq (T * \varphi) + \psi$ (Note: symmetry of conjunction).
8. **Superexpansion:** If $\neg\psi \notin T * \varphi$, then $T * (\varphi \wedge \psi) \supseteq (T * \varphi) + \psi$.

### Question: Are the postulates 'correct'?

This is impossible to say without formal semantics; no defintion for $*$. AGM only defines syntax.

[1]Name after its authors: Carlos Alchourrón, Peter Gärdenfors, and David Makinson (1985).

**⚠ Attention: Higher-Order Beliefs and AGM**

AGM postulates become inconsistent when applied to higher-order beliefs. For a Moore sentence $\varphi := p \wedge \neg \mathrm{B}p$, the Success postulate requires believing $\varphi$ after learning it, which forces an introspective agent to acquire inconsistent beliefs. Furthermore, Vacuity is too liberal, allowing revision with any consistent sentence even if the agent already **knows** its negation.

**Limiting Vacuity** $AGM^K$: Updated axioms (Def 29)

- Vacuity states: successful revision with *any* logically consistent sentence $\varphi$ (not a contradiction)
- Accounting for the agent's knowledge $T$: if $\neg\varphi \in T$, should never revise with $\varphi$

$\Rightarrow$ restrict Vacuity to formulas that are logically consistent and consistent with $T$:

$$\text{if } \neg \mathrm{K} \neg \varphi \text{ then } T \star \varphi \text{ is consistent} \tag{14}$$

### C Defining a more expressive language for Belief Revision

*Strategy* for defining a more expressive language + proof system:

1. **Syntax**: (Non-monotonic) Conditional Logic: Conditional belief operators as contingency plans for belief revision.
2. **Semantics**: Conditional Logic
   - Grove sphere models (Lewis-Stalnaker semantics for counterfactual conditionals)
   - Spohn ordinal ranking models
   - Preferential models (J. Halpern)
   - Belief revision models (O. Board)
   - Plausibility models (Baltag, Smets)
   - Probabilistic models and spaces (Popper, Brandenburger)

### C.1 Sphere Models

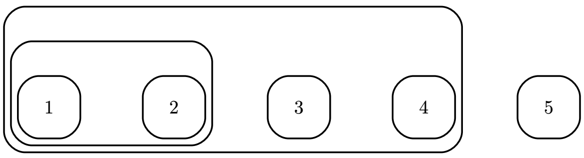
**ⓘ Intuition: Fallback Beliefs**

To accommodate belief revision semantically, agents need a contingency plan—weaker secondary beliefs they can fall back on if their primary beliefs are contradicted. This is formalized using nested spheres $S_0, S_1, \ldots$ or plausibility orders over states.

**Example 27: Suprise Exam**

Student $a$ has tiered levels of belief:

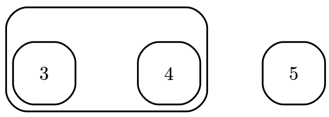
- *strongest*: $S_0 = \{1, 2\}$
- *a bit weaker*: $S_1 = \{1, 2, 3, 4\}$
- *weakest* (implicit): $S_2 = S = \{1, 2, 3, 4, 5\}$




If $a$'s first belief $S_0$ (Monday or Tuesday) is wrong, then $a$'s contingency is to belief in $S_1$ (Wednesday or Thursday).

Now, after updates $!\neg 1, !\neg 2$, $a$ still holds consistent beliefs:

We can repeat this with the implicit last sphere of belief $S_2$, maintaining consistency and allowing for automatic belief revision.



**Remember: Well-foundedness and converse well-foundedness**

**Def 30** (*Well-foundedness*): Means there are **no infinite descending chains** ($S_0 > S_1 > \ldots > S_n > \ldots$). This guarantees that every non-empty subset has a minimal element (e.g., a "closest" world or "smallest" sphere).

**Def 31** (*Converse well-foundedness*): Means there are **no infinite ascending chains** ($S_0 < S_1 < \ldots < S_n < \ldots$) of better and better worlds. This guarantees that every non-empty subset has a maximal element (e.g., a "most plausible" world).

**Def 32** (*Single-Agent Sphere Model for Belief Revision (Grove Model)*): A Grove model is a tuple $\mathbf{S} = (S, \mathcal{F}, \|\cdot\|, s_\star)$, where $\mathcal{F}$ is a nested, well-founded, and exhaustive family of subsets of $S$ (spheres) such that:

1. Nested: $\forall S', S'' \in \mathcal{F}$, either $S' \subseteq S''$ or $S'' \subseteq S'$.
2. Smallest intersecting sphere:

$$\forall P \subseteq S \text{ with } P \neq \emptyset, \quad \exists S' \in \mathcal{F} : \forall S'' \in \mathcal{F} : P \cap S'' \neq \emptyset \Leftrightarrow S' \subseteq S'' \tag{15}$$

3. Exhaustive: $\cap \mathcal{F} \neq \emptyset$ and $S = \cup \mathcal{F}$.

The smallest sphere $S_0 = \cap \mathcal{F}$ represents the agent's strongest beliefs.

**Note**

The above is an extension of simple models (not yet Kripke).

**Spohn Ordinals** Because the family of spheres $\mathcal{F}$ is well-founded, we can sequentially identify the smallest spheres and index them:

- **Smallest sphere:** $S_0 := \cap \mathcal{F}$, which has the property that $S_0 \subseteq S'$ for all spheres $S' \in \mathcal{F}$.
- **Next smallest:** $S_1 \in \mathcal{F} \setminus \{S_0\}$, such that $S_1 \subseteq S'$ for all remaining spheres $S' \in \mathcal{F} \setminus \{S_0\}$.
- **Indexing:** This allows the family $\mathcal{F}$ to be indexed by ordinals (or natural numbers in finite cases) up to some ordinal $\beta$: $S_0 \subset S_1 \subset \ldots \subset S_\alpha \subset S_{\alpha+1} \subset \ldots \subset S_\beta = S$

**Def 33** (*Spohn Ordinal / Degree of Implausibility*): For every world $w \in S$, the Spohn ordinal $\mathrm{ord}(w)$ of world $w$ is defined as the *least ordinal* $\alpha$ such that $w \in S_\alpha$, it represents the *"degree of implausibility"* of $w$.

**Def 34** (*Belief and Knowledge in Grove models*): As in epistemic-doxastic models (Def 19): by quantifying respectively over

- $S$ for knowledge, and over
- $S_0$ for belief. (Student question: quantify over $S_1, S_2, \ldots$ for weaker beliefs?)

**Def 35** (*Updates in Grove models*): An *update* $!\varphi$ with a sentence $\varphi$ is defined on full sphere models $\mathbf{S} = (S, \mathcal{F}, \|\cdot\|, s_\star)$ similarly as on sphere-based epistemic-doxastic models (Def 19), except the family of spheres is restricted to worlds in $\|\varphi\|_{\mathbf{S}}$, the new family of spheres is

$$\mathcal{F}' = \{S' \cap \|\varphi\|_{\mathbf{S}} \mid S' \in \mathcal{F} : S' \cap \|\varphi\|_{\mathbf{S}} \neq \emptyset\} \tag{16}$$

### C.2 Plausibility Models

**Remember**

**Def 36** (*Preorder*): Reflexive and transitive binary relation $R$ on set $S$:

$$\forall s \in S : s \leq s \text{ and } \forall s, t, w \in S : (s \leq t \land t \leq w) \to s \leq w \tag{17}$$

**Def 37** (*Totality of a binary relation*):

$$\forall s, t \in S : s \leq t \lor t \leq s \tag{18}$$

**Def 38** (*Single-Agent Plausibility Model*): A plausibility model is a tuple $\mathbf{S} = \left(S, \underset{\text{pl.}}{\leq}, \|\cdot\|, s_\star\right)$ where:

- $S$ is a non-empty set of states.
- $\underset{\text{pl.}}{\leq}$ is a converse-well-founded total preorder (*plausibility order*).
- Converse-well-foundedness ensures that every non-empty set $P \subseteq S$ has a set of most plausible states: best $P = \left\{s \in P : s \underset{\text{pl.}}{\leq} t \text{ for all } t \in P\right\} \neq \emptyset$.

**Intuition: Plausibility order**

$s \underset{\text{pl.}}{\leq} t$ means:

1. $s$ is at least as plausible as $t$.
2. $t$ is in at least as many spheres as $s$.
3. Independent of what agent $a$ learns, as long as $s$ is consistent with $a$'s beliefs and $t$ is epistemically possible, $t$ is also consistent with $a$'s beliefs.

**Note: Equivalence of Spheres and Plausibility**

Grove models (Def 32) and plausibility models (Def 38) are mathematically equivalent. The plausibility relation can be extracted via

$$s \underset{\text{pl.}}{\leq} t \Leftrightarrow \forall S' \in \mathcal{F}(s \in S' \to t \in S') \tag{19}$$

An alternative statement using Spohn Ordinals (Def 33):

$$s \underset{\text{pl.}}{\leq} t \Leftrightarrow \operatorname{ord}(s) \geq \operatorname{ord}(t). \tag{20}$$

Conversely, spheres can be generated by

$$\mathcal{F} := \{w^{\leq} : w \in S\}; w^{\leq} = \left\{s \in S : w \underset{\text{pl.}}{\leq} s\right\}. \tag{21}$$

**Notation: Strict plausibility**

A bit of syntactic sugar: abbreviate $\left(s \leq t \text{ and } t \not\leq_{\text{pl.}}\right)$ as $s < t$.

**Note: Most plausible states**

Totality + converse well-foundedness together are equivalent to requiring that in every set of states $S$ there are some "most plausible" ones: for every $P \subseteq S$, if $P$ is non-empty then the set

$$\mathrm{bestP} = \max_{\underset{\text{pl.}}{\leq}} P := \left\{ s \in P : t \underset{\text{pl.}}{\leq} s \forall t \in P \right\} \tag{22}$$

is also nonempty.

### Interpretation Map on Plausibility models

cont

Figure 27: continue here

### D Conditional Beliefs and The Logic of Knowledge

**Def 39** (*Knowledge and Belief in Plausibility Models*):
- **Knowledge:** $s \vDash \mathrm{K}\varphi$ iff $\|\varphi\|_S = S$.
- **Belief:** $s \vDash \mathrm{B}\varphi$ iff best $S \subseteq \|\varphi\|_S$.

**Def 40** (*Conditional Belief*): A proposition $\varphi$ is believed conditional on $\psi$, denoted $\mathrm{B}^\psi\varphi$, if $\varphi$ is true in the most plausible $\psi$-worlds:

$$\left\|\mathrm{B}^\psi\varphi\right\|_S = \{s \in S : \text{best } \|\psi\|_S \subseteq \|\varphi\|_S\} \tag{23}$$

essentially: in the smallest sphere

plausibility model drawing: assuming reflexivity and transitivity because preorder (not drawn) arrow means BELOW

**Example 28: Surprise Exam**

**Note: Conditional Belief as Belief Revision**

We can semantically capture the AGM(Def 29) revision operator using conditional belief (Def 40): $T * \varphi := \{\theta : s_\star \vDash \mathrm{B}^\varphi\theta\}$. This interpretation guarantees that all modified AGM axioms are sound.

**Theorem 7: Axiomatization of Knowledge and Conditional Belief**

The complete logic bridging knowledge and conditional belief includes:

- **Necessitation:** From $\vdash \varphi$ infer $\vdash \mathrm{B}^{\psi}\varphi$ and $\vdash \mathrm{K}\varphi$
- **Normality:** $\mathrm{B}^{\psi}\varphi \Rightarrow \theta \Rightarrow \left(\mathrm{B}^{\psi}\varphi \Rightarrow \mathrm{B}^{\psi}\theta\right)$
- **Truthfulness of Knowledge:** $\vdash \mathrm{K}\varphi \Rightarrow \varphi$
- **Persistence of Knowledge:** $\vdash \mathrm{K}\varphi \Rightarrow \mathrm{B}^{\psi}\varphi$
- **Full Introspection:** $\vdash \mathrm{B}^{\psi}\varphi \Rightarrow \mathrm{K}\mathrm{B}^{\psi}\varphi$ and $\vdash \neg\mathrm{B}^{\psi}\varphi \Rightarrow \mathrm{K}\neg\mathrm{B}^{\psi}\varphi$
- **Success of Belief Revision:** $\vdash \mathrm{B}^{\varphi}\varphi$
- **Consistency of Belief Revision:** $\vdash \neg\mathrm{K}\neg\varphi \Rightarrow \neg\mathrm{B}^{\varphi}\mathrm{False}$
- **Inclusion:** $\vdash \mathrm{B}^{\varphi\wedge\psi}\theta \Rightarrow \mathrm{B}^{\varphi}\psi \Rightarrow \theta$
- **Rational Monotonicity:** $\vdash \neg\mathrm{B}^{\varphi}\neg\psi \wedge \mathrm{B}^{\varphi}\theta \Rightarrow \mathrm{B}^{\varphi\wedge\psi}\theta$

**Example 29: Professor Wine**

Professor Wine knows he is either a genius ($g$) or drunk ($d$). He doesn't feel drunk, so he believes he is a sober genius ($g \wedge \neg d$). However, if he realized he was drunk, he'd conditionally believe he was a drunk non-genius. Formalized: $\mathrm{B}g$, $\mathrm{K}g \vee d$, $\mathrm{B}\neg d$, $\mathrm{B}^{d}\neg g$, and actually $d \wedge g$. This demonstrates that his true belief ($g$) is not knowledge, because it is easily lost upon learning $d$. Conditional belief safely

**Continue** click here: slide 47/70

do not confuse counterfactual and conditional beliefs

$\vdash B^{\varphi}\varphi$ is valdi in this logic

we continue from slide 66 in l4-2

## Session 4-2 (Lecture):

## Session 4-3 (Lecture):

# Week 5

## Session 5-1 (Lecture):

## Session 5-2 (Lecture):

## Session 5-3 (Lecture):

# Week 6

## Session 6-1 (Lecture):

## Session 6-2 (Lecture):

## Session 6-3 (Lecture):

# Week 7

## Session 7-1 (Lecture):

## Session 7-2 (Lecture):

## Session 7-3 (Lecture):

# Week 8

## Session 8-1 (Lecture):

## Session 8-2 (Lecture):

## Session 8-3 (Lecture):

# Glossary: Definitions and Theorems

## List of Definitions


Def 1 Properties of Multi-Agent Systems . . . 5
Def 2 Knowledge . . . 5
Def 3 Justified Belief . . . 5
Def 4 Belief Revision . . . 5
Def 5 Uncertainty . . . 5
Def 6 Game of imperfect information . . . 6
Def 7 Strategic Ignorance . . . 6
Def 8 Distributed Knowledge . . . 6
Def 9 Nested Knowledge . . . 6
Def 10 Introspection . . . 6
Def 11 Common Knowledge . . . 6
Def 12 Fixpoint . . . 7
Def 13 Pluralistic Ignorance . . . 11
Def 14 Single-Agent, pointed Epistemic-Doxastic Model . . . 12
Def 15 Truth in an Interpretation . . . 13
Def 16 Validity . . . 13
Def 17 Satisfiability . . . 13
Def 18 Kripke Model . . . 15
Def 19 Epistemic-Doxastic Kripke Model . . . 15
Def 20 Kripke modalities . . . 15
Def 21 Truth in an interpretation continued: Kripke modalities . . . 15
Def 22 Multi-Agent Kripke Model . . . 18
Def 23 Epistemic/ Doxastic Modalities . . . 18
Def 24 Common Knowledge (Group) . . . 18
Def 25 Reflexive-transitive closure . . . 19
Def 26 Distributed Knowledge (Group) . . . 20
Def 27 Public Announcement Logic . . . 20
Def 28 Moore Sentences . . . 21
Def 29 AGM Postulates for Belief Revision . . . 24
Def 30 Well-foundedness . . . 26
Def 31 Converse well-foundedness . . . 26
Def 32 Single-Agent Sphere Model for Belief Revision (Grove Model) . . . 26
Def 33 Spohn Ordinal / Degree of Implausibility . . . 26
Def 34 Belief and Knowledge in Grove models . . . 26
Def 35 Updates in Grove models . . . 26
Def 36 Preorder . . . 27
Def 37 Totality of a binary relation . . . 27
Def 38 Single-Agent Plausibility Model . . . 27
Def 39 Knowledge and Belief in Plausibility Models . . . 28
Def 40 Conditional Belief . . . 28


## List of Theorems


Theorem 1 Axioms and Relational Properties . . . 16
Theorem 2 Equivalence of Models . . . 17

Theorem 3 Validities for Common Modalities ........ 19
Theorem 4 Validities for Distributed Knowledge ........ 20
Theorem 5 PAL Reduction Axioms ........ 21
Theorem 6 Closure under Composition ........ 21
Theorem 7 Axiomatization of Knowledge and Conditional Belief ........ 29