

DYNAMIC EPISTEMIC LOGIC

Study Notes

Filip Rehburg¹

filip.rehburg@student.uva.nl

¹ *University of Amsterdam*, Amsterdam, The Netherlands

February 28, 2026

Instructor: Alexandru Baltag (TheAlexandruBaltag@gmail.com)

TA: Giuseppe Manes (giuseppe.manes@student.uva.nl)

Do not distribute, please send this link: https://github.com/frehburg/mol_DEL_notes

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, Common & Distributed Knowledge, and Updates	18
2.2	(Lecture): Public Announcement Logic (PAL)	22
2.3	(Lecture): PAL continued	28
3	Week 3	29
3.1	(Lecture): Learnability and Knowability	29
3.2	(Lecture): Tutorial 1	29
3.3	(Lecture): The problem of belief revision	29
4	Week 4	30
4.1	(Lecture): Cheating and the Failure of Standard DEL	30
4.2	(Lecture): Dynamic Belief Revision in the general case	39
4.3	(Lecture):	41
4.4	Homework 1	41
5	Week 5	48
5.1	(Lecture):	48
5.2	(Lecture):	48
5.3	(Lecture):	48
6	Week 6	49
6.1	(Lecture):	49
6.2	(Lecture):	49
6.3	(Lecture):	49
7	Week 7	50
7.1	(Lecture):	50
7.2	(Lecture):	50
7.3	(Lecture):	50
8	Week 8	51
8.1	(Lecture):	51
8.2	(Lecture):	51
8.3	(Lecture):	51
I	Glossary: Definitions and Theorems	52

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, Common & Distributed Knowledge, and Updates : Section 2.1	✓
Public Announcement Logic (PAL) : Section 2.2	WIP
PAL continued : Section 2.3	X
Learnability and Knowability : Section 3.1	X
Tutorial 1 : Section 3.2	X
The problem of belief revision : Section 3.3	X
Cheating and the Failure of Standard DEL : Section 4.1	✓
Dynamic Belief Revision in the general case : Section 4.2	WIP
: Section 4.3	∅
: Section 5.1	∅
: Section 5.2	∅
: Section 5.3	∅
: Section 6.1	∅
: Section 6.2	∅
: Section 6.3	∅
: Section 7.1	∅
: Section 7.2	∅
: Section 7.3	∅
: Section 8.1	∅
: Section 8.2	∅
: Section 8.3	∅

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:

Regex for replacing cite: `\[cite: (\d+,)+\d+\]`

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 $f : X \rightarrow X; x = f(x)$.

In the case of Puzzle 0:

$$C\Box\varphi \equiv K_A C\Box\varphi \wedge K_B C\Box\varphi \quad (1)$$

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

① 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 φ means believing φ $\Box\varphi$, but here reading φ leads to not believing φ : $\Box\neg\varphi$.

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.

❶ Socratic Questioning

Discovering answers by asking questions of students. (Wikipedia)

❷ 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 wives 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**.

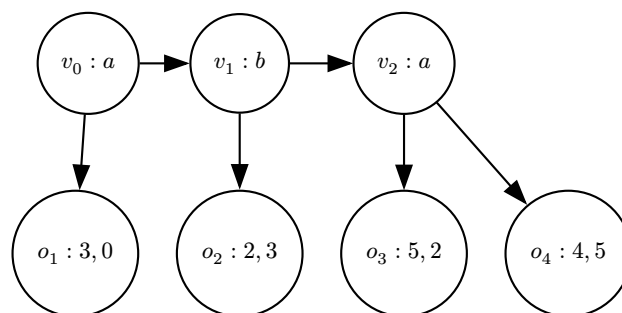
Example 12: Puzzle 7 Modification: The Infinite Lottery

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:



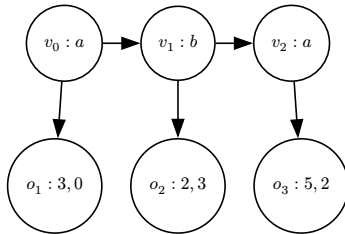
In the leaves ("outcomes" o_j) the first number is a 's payoff, the second number is b 's payoff.

- $v_0 : a$ 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 $o_4(4, 5)$

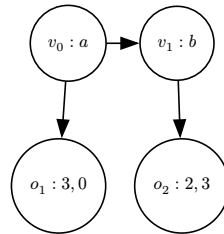
The Backwards Induction (BI) Method

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

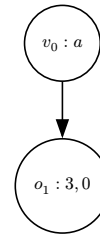
Elimination Step 1



Elimination Step 2



Elimination Step 3



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

i 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 (they 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.

i 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.

$$\varphi ::= p \mid \neg\varphi \mid \varphi \wedge \varphi \mid K\varphi \mid B\varphi \quad (2)$$

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

Def 14 (*Single-Agent, pointed Epistemic-Doxastic Model*): Is a tuple $\mathbf{S} = (S, S_0, \|\cdot\|, s_*)$, 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*.
- $\|\cdot\| : \text{Prop} \rightarrow \mathcal{P}(S)$: A *valuation* map assigning atomic propositions to sets of states.
- $s_* \in S$: 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_* is among S , but cannot distinguish between $s_i, s_j \in S; i \neq j$.
- **Doxastic state:** the agent beliefs $s_* \in S_0$

Notation: Truth

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

$$w \models_{\mathbf{S}} \varphi \quad (3)$$

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 φ is true in a model \mathbf{S} under the valuation map $\|\cdot\|_{\mathbf{S}}$ if

- $\varphi = p; p \in \text{Prop}: w \models p$ iff $w \in \|p\|$,
- $\varphi = \neg\psi: w \models \neg\psi$ iff $w \not\models \psi$,
- $\varphi = \psi \wedge \chi: w \models \psi \wedge \chi$ iff $w \models \psi$ and $w \models \chi$,
- $\varphi = K\varphi: w \models K\varphi$ iff $\forall s \in S, s \models \varphi$,
- $\varphi = B\varphi: w \models B\varphi$ iff $\forall s \in S_0, s \models \varphi$.

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

Def 17 (*Satisfiability*): A sentence φ 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** ($K\varphi$): Truth in all epistemically possible worlds.
- **Belief** ($B\varphi$): 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 φ is the operation of deleting all non- φ 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.



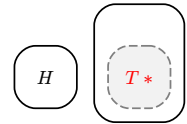
Note: Update as World Elimination

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

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

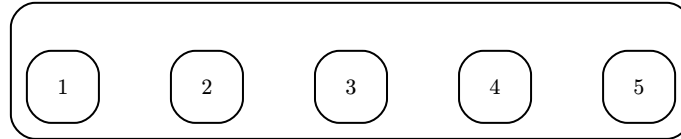


Resolution (Third-Person Models): We maintain an objective perspective where the real possibility always remains in the global model S , even if the agent believes it to be impossible. The sphere of beliefs S_0 (\ominus) is restricted to T , meaning the agent believes H , but their belief is false because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

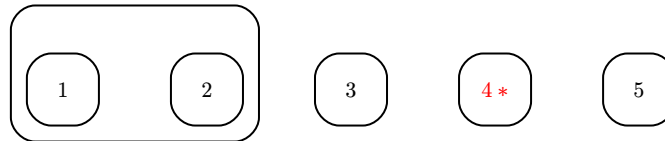


Example 18: Continuation of Puzzle 6: Surprise Exam

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



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



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 $\mathbf{S} = (S, \{R_i\}_{i \in I}, \|\cdot\|, s_*)$ with set of states S , accessibility relations R_i , valuation $\|\cdot\|$, and actual state s_* .

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

$$\langle R_i \rangle \varphi := \neg [R_i] \neg \varphi. \quad (4)$$

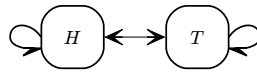
Notation: Kripke modalities: subscript

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. $\varphi = [R_i]\varphi$: $w \models [R_i]\varphi$ iff $v \models \varphi \forall v : wR_iv$.

Example 19: Example 17: Concealed coin continued

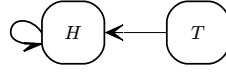
The agent's knowledge in the concealed coin scenario can be represented as:



- 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 $s_* = s$, 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 $\Box\varphi \rightarrow \varphi$
- **(4)** Transitivity $\Box\varphi \rightarrow \Box\Box\varphi$
- **(5)** Euclideaness $\Diamond\varphi \rightarrow \Box\Diamond\varphi$
- **(D)** Seriality $\Box\varphi \rightarrow \Diamond\varphi$
- *Weak Epistemic Model* **(S4)** = **(K)** + **(T)** + **(4)**: No negative introspection (only reflx. & 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 ($\forall s : \exists t : s \rightarrow t$), transitive, Euclidean

Theorem 1: Axioms and Relational Properties

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

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

- Veracity** ($K\varphi \Rightarrow \varphi$): \sim is reflexive.
- Positive Introspection** ($K\varphi \Rightarrow KK\varphi$): \sim is transitive (**4**).
- Negative Introspection** ($\neg K\varphi \Rightarrow K\neg K\varphi$): \sim is Euclidean (and symmetric) (**5**).

Validities for Belief (**KD45** model properties for \rightarrow):

- Consistency** ($\neg B(\varphi \wedge \neg\varphi)$): \rightarrow is serial.
- Positive Introspection** ($B\varphi \Rightarrow BB\varphi$): \rightarrow is transitive.
- Negative Introspection** ($\neg B\varphi \Rightarrow B\neg B\varphi$): \rightarrow is Euclidean.

KB Interaction Properties:

- Knowledge implies Belief** ($K\varphi \Rightarrow B\varphi$): If $s \rightarrow t$ then $s \sim t$.
- Strong Positive Introspection** ($B\varphi \Rightarrow KB\varphi$): If $s \sim t$ and $t \rightarrow w$ then $s \rightarrow w$.
- Strong Negative Introspection** ($\neg B\varphi \Rightarrow K\neg B\varphi$): If $s \sim t$ and $t \rightarrow w$ then $s \rightarrow w$.

 **Note: 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 φ and $(\varphi \rightarrow \psi)$ infer ψ
 - Necessitation: from φ infer $K\varphi$ and $B\varphi$
 - Kripke's axioms for K and B:
 - $(K\varphi \rightarrow \psi) \rightarrow (K\varphi \rightarrow K\psi)$
 - $(B\varphi \rightarrow \psi) \rightarrow (B\varphi \rightarrow B\psi)$

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

 **Attention: Logical Omniscience**

Any Kripke modality validates axiom K ($K(\varphi \Rightarrow \psi) \Rightarrow (K\varphi \Rightarrow K\psi)$) and the Necessitation rule (if φ is valid, $K\varphi$ 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, Common & Distributed Knowledge, and Updates

A Multi-Agent Kripke Models & Modalities

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

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

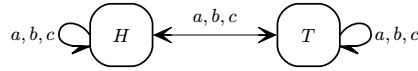
where \mathcal{A} is a set of labels representing the names of epistemic agents.

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

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):

$$w \models \Box_b (\neg \Box_a H \wedge \neg \Box_a T) \wedge \Box_b \Box_a (H \vee T) \quad (6)$$

B Common Knowledge

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

$$s \models_S C\Box_G \varphi \Leftrightarrow t \models_S \varphi \quad (7)$$

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

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: $\left[\left(\bigcup_{a \in G} \rightarrow_a\right)^*\right]$.

🔍 Remember

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

$$wR^*v \text{ iff } \exists \text{ finite chain (length } n \geq 0) : w = w_0 R w_1 R \dots R w_n = v \quad (8)$$

❗ Intuition: Common Knowledge as Infinite Conjunction

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

Then, common knowledge $C\Box_G\varphi$ (Def 24) is semantically equivalent to the infinite conjunction:

$$\varphi \wedge E_G\varphi \wedge E_GE_G\varphi \wedge \dots \quad (9)$$

⚠ Attention: Infinitary definitions

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

Instead, $C\Box_G$ is interpreted as a **primitive** operator induced by the semantic clause in Info 2

📖 Theorem 3: Validities for Common Modalities

- **Fixed-Point Axiom** (Mix): $C\Box_G\varphi \Rightarrow (\varphi \wedge E_GC\Box_G\varphi)$
- **Induction Axiom**: $C\Box_G(\varphi \Rightarrow E_G\varphi) \Rightarrow (\varphi \Rightarrow C\Box_G\varphi)$

B.1 Syntax

Epistemic logic with common knowledge | Doxastic logic with common true belief

$$\varphi ::= p \mid \neg\varphi \mid \varphi \wedge \varphi \mid K_a\varphi \mid Ck_G\varphi \quad (10) \qquad \varphi ::= p \mid \neg\varphi \mid \varphi \wedge \varphi \mid B_a\varphi \mid Cb_G\varphi \quad (11)$$

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

$$\varphi ::= p \mid \neg\varphi \mid \varphi \wedge \varphi \mid K_a\varphi \mid B_a\varphi \mid Ck_G\varphi \mid Cb_G\varphi \quad (12)$$

Complete Axiomatization for Common Knowledge:

- Multi-agent versions of the axioms in Theorem 1 (modalities labeled with agents)
- Fixed-Point and Induction Axioms (Theorem 3) for both Ck_G and Cb_G
- Kripke axioms for both Ck_G and Cb_G : $C\Box_G(\varphi \rightarrow \psi) \rightarrow (C\Box_G\varphi \rightarrow C\Box_G\psi)$
- Necessitation for both Ck_G and Cb_G : $\varphi \rightarrow C\Box_G\varphi$

C Distributed Knowledge

Def 26 (*Distributed Knowledge (Group)*): Distributed knowledge within a group G , denoted $D\Box_G\varphi$ or $Dk_G\varphi$, is obtained by quantifying over all worlds simultaneously accessible by all arrows for agents in G : $s \models_S D\Box_G\varphi \Leftrightarrow t \models_S \varphi$ for every t such that $s \xrightarrow{a} t$ 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 communicated all their private knowledge.

Public announcements: By communicating, private knowledge can become (common) knowledge.

Distributed Knowledge equivalence to Kripke Modality

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

$$\bigcap_{a \in G} \xrightarrow{a} \quad (13)$$

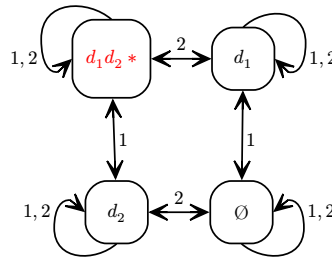
Note: Interpretations of distributed modalities

- when the relations \xrightarrow{a} are reflexive: Dk_G as some sort of distributed knowledge
- when the relations \xrightarrow{a} represent beliefs: Db_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 $s_* = (d_1, d_2)$, neither knows both are dirty, but it is distributed knowledge:

$$s_* \models \neg K_1(d_1 \wedge d_2) \wedge \neg K_2(d_1 \wedge d_2) \wedge Dk(d_1 \wedge d_2) \quad (14)$$



Theorem 4: Validities for Distributed Knowledge

The following are valid on all epistemic models:

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

Complete Axiomatization of Distributed Knowledge:

- Axioms for multi-agent epistemic logic (subset of multi-agent **S5**)
- Fixed-Point and Induction Axioms (Theorem 3) for $D\Box_G$
- Kripke axiom $D\Box_G: D\Box_G(\varphi \rightarrow \psi) \rightarrow (D\Box_G\varphi \rightarrow D\Box_G\psi)$
- Necessitation for $D\Box_G: \forall a \in G : \Box_a\varphi \rightarrow D\Box_G\varphi$

D Dynamics & Public Announcements

Def 27 (*Updates on Sphere Models*): When new information φ is learned with absolute certainty, the resulting situation is represented by performing an update $!\varphi$ on the original model \mathbf{S} . This corresponds to simply deleting all non- φ -worlds from \mathbf{S} .

The new epistemic and doxastic relations are the restrictions of the old ones to the new set of states.

- **New set of worlds:**

$$S' = S \cap \|\varphi\|_{\mathbf{S}} = \|\varphi\|_{\mathbf{S}} = \{w \in S \mid w \models_{\mathbf{S}} \varphi\} \quad (15)$$

- **New sphere of beliefs:**

$$S'_0 = S_0 \cap \|\varphi\|_{\mathbf{S}} = \{w \in S_0 \mid w \models_{\mathbf{S}} \varphi\} \quad (16)$$

- **New valuation:**

$$\|p\|_{\mathbf{S}'} = \|p\|_{\mathbf{S}} \cap \|\varphi\|_{\mathbf{S}} \quad (17)$$

Def 28 (*Updates on Kripke Models*): Similar as above (Def 27) and also need to delete all doxastic and epistemic arrows connecting to or from a deleted world. As above, the new doxastic and epistemic relations are restrictions to the new set of states $\|\varphi\|_{\mathbf{S}}$.

⚠ Attention: Assumption: truthful updates

An update $!\varphi$ means absolute certainty of φ . Thus, update $!\varphi$ can only be performed on a model if φ is true at the real world: $s_{\star} \models_{\mathbf{S}} \varphi$ because absolute certainty implies truth.

≡ Example 23: Concealed Coin: Announcement with absolute certainty

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 c): they all see it's Heads up (**H**), and they all see that the others see it etc.

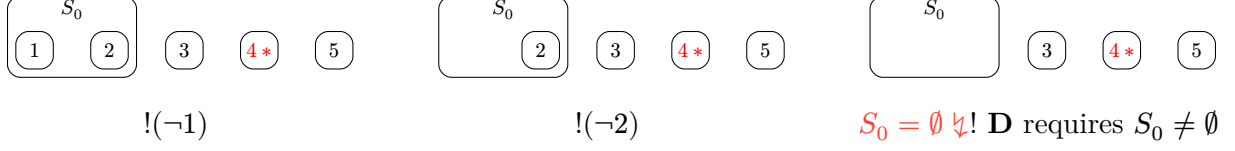
This announcement comes with **absolute certainty** and thus truth.

So this is a *public announcement* that the coin lies Heads up. We denote this event by $!\mathbf{H}$. Intuitively, after the announcement, we have common knowledge of **H**: $CK_{a,b,c}\mathbf{H}$

Example 24: Surprise Exam: Empty sphere of beliefs

The student believes the exam will be either on Monday (1) or Tuesday (2), but actually the exam takes place on Thursday ($s_* = 4$).

Sphere model representation of the situation with updates:



The Problem of Belief Revision

In doxastic models, if an announcement contradicts prior beliefs, the update might erase all worlds within an agent's sphere of beliefs.

This results in an empty sphere of beliefs $S_0 = \emptyset$, violating Theorem 1 iv. Consistency of Beliefs (**D**) meaning the agent has inconsistent beliefs and believes everything.

Question: How can we revise our beliefs without arriving at inconsistent beliefs?

We will revisit this later.

Session 2-2 (Lecture): Public Announcement Logic (PAL)

A Introduction to DEL and PAL

Background

1. Public Announcement Logic (PAL) [indep.: Plaza (1989), Gerbrandy and Groeneveld (1997)].
2. PAC = PAL + $C\Box$ operator [complete axiomatization: Baltag, Moss, and Solecki (1998)].

Remember: Propositional Dynamic Logic (PDL)

Def 29 (*PDL Dynamic Modalities*): Given a model $\mathbf{S} = (S, \{R\}, \|\cdot\|)$, program α , $[\alpha]$ is a universal Kripke modality, semantically defined via the relation $R_\alpha \subseteq S \times S$ for a world $s \in S$:

$$s \models_{\mathbf{S}} [\alpha] \text{ iff } \forall t \in S : sR_\alpha t : t \models_{\mathbf{S}} \varphi \quad (18)$$

The dual (existential) dynamic modality is defined as:

$$\langle \alpha \rangle \varphi := \neg [\alpha] \neg \varphi \quad (19)$$

Notation: Transition relations

The relations R_α are interpreted as *transition relations* in DEL; they relate an input-state s to possible output-states t that might result by executing program α on input s .

Thus, $sR_\alpha t$ is often written as $s \rightarrow^\alpha t$

Def 30 (*Dynamic Epistemic Logic*): *Dynamic Epistemic Logic* (DEL) is a formal framework in logic and computer science used to model how the knowledge and beliefs of multiple agents change over time in response to specific events or communications.

Expressing informational changes:

- DEL borrows dynamic modalities $[\alpha]$ from PDL.
- Interpret PDL programs as epistemic actions α
- *Intended meaning*: if action α is performed (in the current state), then φ will become true afterwards.

DEL vs PDL Semantics

Unlike PDL, DEL considers transition relations between states living in *different* models. Initial models are only supposed to represent the epistemic situation at a given moment and might not contain states corresponding to all possible action outputs. While one could mathematically convert an “open” system into a “closed” one containing all future states, in a truly open system the number of possible actions is a proper class, making the model too large.

Thus, DEL performs updates by interpreting epistemic actions such as public announcements as *model transformers* (Def 32).

B Public Announcement Logic

B.1 Syntax

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

$$\varphi ::= p \mid \neg\varphi \mid \varphi \wedge \varphi \mid K_a\varphi \mid [!\varphi]\varphi \quad (20)$$

B.2 Semantics

- PAL formulas are interpreted on multi-agent Kripke models:
 - *epistemic S5 models*: epistemic logic with public announcements,
 - *doxastic K45 models*: doxastic logic with public announcements
 - *Note*: Axiom (D) (consistency of beliefs) is inconsistent with public announcements.

Def 31 (*DEL semantics of the dynamic modalities*): are given by:

For any state $s \in S$

$$s \models_S [\alpha]\varphi \Leftrightarrow \forall t \in S^\alpha : s \rightarrow_S^\alpha t : t \models_{S^\alpha} \varphi \quad (21)$$

Intuition: General framework for epistemic actions as model transformers

An epistemic action α acts as a *model transformer* consisting of:

1. A map $S \mapsto S^\alpha$ that takes any initial Kripke model S into an “updated” model S^α .
2. A binary transition relation $\rightarrow_S^\alpha \subseteq S \times S^\alpha$ between the input-states of the initial model S and the output-states (living in the updated model S^α).

Def 32 (*Public Announcement as a Model Transformer*): A public announcement $!\varphi$ (Def 28) acting as a *model transformer*

1. maps any model $\mathbf{S} = (S, \xrightarrow{a}, \|\cdot\|)$ to a new model $\mathbf{S}^{!\varphi} = (S_\varphi, \xrightarrow[\varphi]{a}, \|\cdot\|_\varphi)$, given by:
 - **Domain:** $S_\varphi := \|\varphi\|_{\mathbf{S}}$
 - **Relations** $\xrightarrow[\varphi]{a}$: $s \xrightarrow[\varphi]{a} t$ iff $s \xrightarrow{a} t$, for $s, t \in S_\varphi$.
 - **Valuation:** $\|p\|_\varphi := \|p\|_S \cap S_\varphi$.
2. Contains the transition relation $\rightarrow_{\mathbf{S}}^{!\varphi}$, relating any state $s \in \mathbf{S}$ satisfying φ to the same state s in the restricted model $\mathbf{S}^{!\varphi}$.

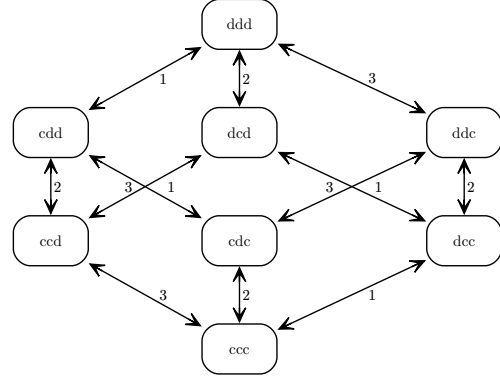
C Solving the Muddy Children Puzzle

Example 25: Muddy Children

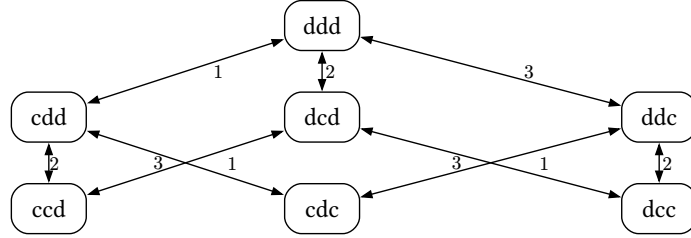
Note: Epistemic **S5** model \Rightarrow reflexivity, but skipped drawing reflexive arrows

- 3 children: $\{1, 2, 3\}$,
- d_i denotes “child i is dirty”.
- There are 8 possible worlds.
- Let $s_* = ddc$: where children 1 and 2 are dirty, and child 3 is clean: $d_1 \wedge d_2 \wedge \neg d_3$.

In the initial model, it is common knowledge that each child knows if the others are dirty or not, but does not know if they themselves are dirty.



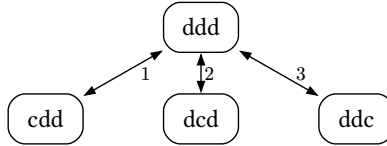
1. **First Announcement:** The Father announces “At least one of you is dirty”. If he is an infallible source, this public announcement is an update $!(d_1 \vee d_2 \vee d_3)$, which deletes the ccc world.



2. **First Round of Questioning:** The children answer “I don’t know if I am dirty or not”. Assuming they tell the truth, this is a public update:

$$!\left(\bigwedge_i (\neg K_i d_i \wedge \neg K_i \neg d_i)\right) \quad (22)$$

This eliminates worlds where a child would know their state (like dcc , cdc , ccd). After this update, in the real world ddc , children 1 and 2 now *know* they are dirty, while child 3 still does not know.



3. **Second Round of Questioning:** Children 1 and 2 answer that they know, while child 3 answers that they don’t. This constitutes a new public update:

$$!(K_1 d_1 \wedge K_2 d_2 \wedge \neg K_3 d_3 \wedge \neg K_3 \neg d_3) \quad (23)$$

This eliminates all remaining worlds except for $s_* = ddc$. Now, child 3 knows it’s clean!

***ddc**

D Axiomatizations and Expressivity

Theorem 5: Complete Axiomatizations & Reduction Axioms

A complete axiomatization of basic PAL is given by combining the standard multi-agent modal logic axioms, Kripke's axiom, the Necessitation rule for dynamic modalities $[\![\varphi]\!]$, and the following

Reduction Axioms (Recursion Axioms):

- **Atomic Permanence:** $[\![\varphi]\!]p \Leftrightarrow (\varphi \Rightarrow p)$
- **Announcement-Negation:** $[\![\varphi]\!]\neg\psi \Leftrightarrow (\varphi \Rightarrow \neg[\![\varphi]\!]\psi)$
- **Announcement-Conjunction:** $[\![\varphi]\!](\psi \wedge \theta) \Leftrightarrow ([\![\varphi]\!]\psi \wedge [\![\varphi]\!]\theta)$
- **Announcement-Knowledge:** $[\![\varphi]\!]K_a\psi \Leftrightarrow (\varphi \Rightarrow K_a[\![\varphi]\!]\psi)$

Note: These are axiom schemata, rather than single axioms, and the logic is *not* closed under substitution. Epistemic/doxastic logic with public announcements is completely axiomatized by taking **S5/K45** axioms alongside the Reduction Axioms stated for K or B .

Expressivity and Succinctness of PAL

By recursively applying the Reduction Axioms, any PAL formula can be rewritten into an equivalent formula in basic modal (epistemic/doxastic) logic. Thus, PAL has the exact same expressivity as basic modal logic, meaning all dynamic modalities are eliminable.

However, there is a difference in succinctness: PAL is exponentially more succinct. For example, capturing the full n -agent Muddy Children scenario (including announcements) in basic epistemic logic yields a vastly more complex formula than doing so in PAL.

E Moore Sentences

Def 33 (*Moore Sentences*): It is a misconception that every sentence becomes known, becomes true, or becomes common knowledge after being truthfully publicly announced. A **Moore sentence** is a sentence φ that becomes FALSE immediately after it is truthfully publicly announced.

Therefore, the following claims are universally WRONG for all formulas:

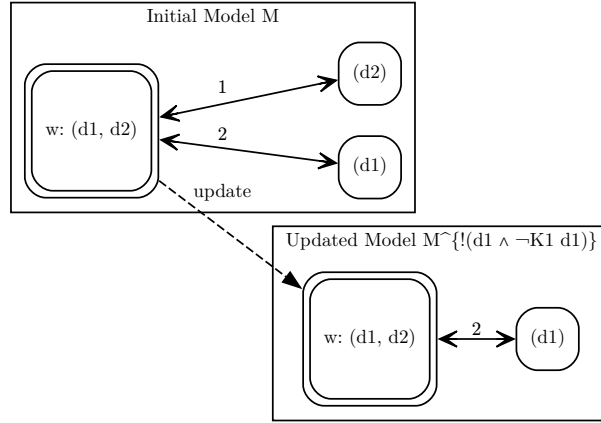
- $[\![\varphi]\!]\varphi$
- $[\![\varphi]\!]K_a\varphi$
- $[\![\varphi]\!]Ck\varphi$

Moreover, Moore sentences always become COMMONLY KNOWN TO BE FALSE after being publicly announced:

$$[\![\varphi]\!]Ck\neg\varphi \quad (24)$$

Example 26: Moore Sentences in Muddy Children

Consider the statement “You are dirty but you don’t know it” addressed to child 1: $d_1 \wedge \neg K_1 d_1$. In the initial real world $w = (d_1, d_2)$ of the Two Muddy Children story, this sentence evaluates to true. We can represent the initial epistemic state and its update:



- In the updated model $M^{!(d_1 \wedge \neg K_1 d_1)}$, the world (d_2) has been deleted.
- In the remaining worlds, child 1 now *knows* they are dirty, meaning the previously announced sentence $d_1 \wedge \neg K_1 d_1$ has become false.
- In fact, it becomes common knowledge that the statement is false: $w \models_{M^{! \varphi}} Ck \neg (d_1 \wedge \neg K_1 d_1)$.

Example 27: Further Examples of Moore Sentences

- **Muddy Children (2 children):** “Both children are dirty but none of them knows he’s dirty”:

$$d_1 \wedge d_2 \wedge \neg K_1 d_1 \wedge \neg K_2 d_2 \quad (25)$$

- **Muddy Children (2 children):** “Both of you are dirty but none of you know this (=that you are both dirty)”:

$$d_1 \wedge d_2 \wedge \neg K_1 d_1 \wedge d_2 \wedge \neg K_2 d_1 \wedge d_2 \quad (26)$$

- **Alice, Bob, and Charles story (“love triangle”):** “Bob doesn’t know about our affair”:

$$(\text{affair}) \wedge \neg K_b \text{affair} \quad (27)$$

- **Muddy Children ($k > 2$):** The sentence “nobody knows if he’s dirty or not” is true initially, and remains true after being publicly announced once. BUT it becomes false (a Moore sentence) after $k - 1$ repeated announcements!

F Iteration and Closure

Announcements about Announcements

It is important that in PAL we can iterate all constructions. We can announce not only facts $!p$ or combinations of facts $!(p \vee \neg q)$, but also epistemic formulas $!(\neg K_a p)$. We can even make announcements about other announcements: $!([q] \neg K_a p)$. This is essential for the closure property.

Theorem 6: Closure Under Composition

Performing two public announcements successively $!\varphi;!\psi$ is equivalent to performing a single, more complex public announcement $!(\varphi \wedge [!\varphi]\psi)$. This semantic closure of public announcements under sequential composition is captured by the following valid schema:

$$[!\varphi][!\psi]\theta \Leftrightarrow [!(\varphi \wedge [!\varphi]\psi)]\theta \quad (28)$$

Continue at AFTER THE EXAMPLE: [\[click here\]](#) (slide 11/29 in DEL actual lecture 2.pdf)

Session 2-3 (Lecture): PAL continued

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

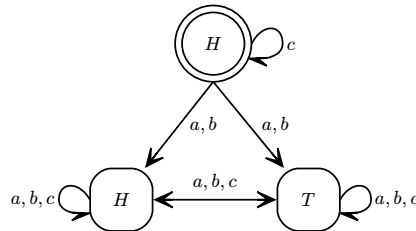
A The Failure of Standard DEL

⚠ 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).

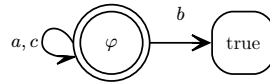
≡ Example 28: Counterexample: Scenario 4

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

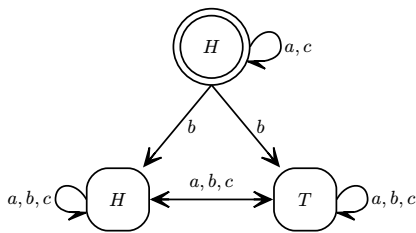


c privately knows that the coin lies heads up: $\varphi = K_c H$ (also: $\neg K_{a,b} \varphi$).

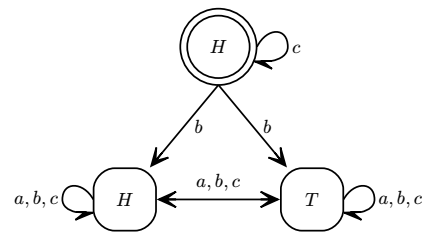
Scenario 5: c sends a secret announcement to a : $!_{c,a} \varphi$, the event model is:



Intuitive updated state model:

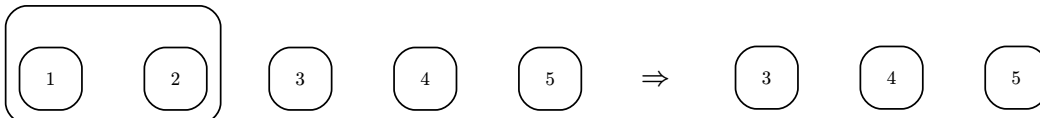


Actual updated state model:



≡ Example 29: 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 30: 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**.

i The Problem of Belief Revision

What happens if an agent a learns a new fact φ 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 φ and without arriving at a contradiction.

i Intuition: AGM Theory Intuition

Standard Belief Revision Theory (AGM¹) attempts to solve this via an axiomatic approach on theories (belief sets) T .

Given input φ , 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 34 (*AGM Postulates for Belief Revision*): Let T be a theory and φ, ψ 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 definition for $*$. AGM only defines syntax.

¹Name 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 $\varphi := p \wedge \neg Bp$, the Success postulate requires believing φ 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 34)

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

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

$$\text{if } \neg K\neg\varphi \text{ then } T \star \varphi \text{ is consistent} \quad (29)$$

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

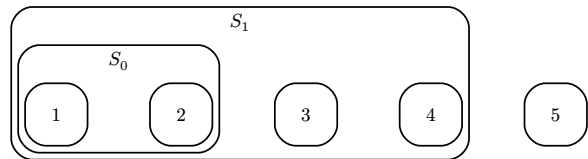
i Intuition: Fallback Beliefs

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, \dots or plausibility orders over states.

≡ Example 31: Surprise Exam

Student a has tiered levels of belief:

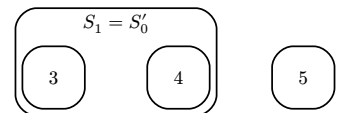
- *strongest*: $S_0 = \{1, 2\}$
- *a bit weaker*: $S_1 = \{1, 2, 3, 4\}$
- *weakest* (implicit): $S_2 = S = \{1, 2, 3, 4, 5\}$



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 $!-1, !-2$, 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 35 (*Well-foundedness*): Means there are **no infinite descending chains** ($S_0 > S_1 > \dots > S_n > \dots$). This guarantees that every non-empty subset has a minimal element (e.g., a “closest” world or “smallest” sphere).

Def 36 (*Converse well-foundedness*): Means there are **no infinite ascending chains** ($S_0 < S_1 < \dots < S_n < \dots$) of better and better worlds. This guarantees that every non-empty subset has a maximal element (e.g., a “most plausible” world).

Def 37 (*Single-Agent Sphere Model for Belief Revision (Grove Model)*): A Grove model is a tuple $\mathbf{S} = (S, \mathcal{F}, \|\cdot\|, s_*)$, 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:

$$\forall P \subseteq S \text{ with } P \neq \emptyset, \exists S' \in \mathcal{F} : \forall S'' \in \mathcal{F} : P \cap S'' \neq \emptyset \Leftrightarrow S' \subseteq S'' \quad (30)$$

3. Exhaustive: $\bigcap \mathcal{F} \neq \emptyset$ and $S = \bigcup \mathcal{F}$.

The smallest sphere $S_0 = \bigcap \mathcal{F}$ 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:** $S_0 := \bigcap \mathcal{F}$, 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 β : $S_0 \subset S_1 \subset \dots \subset S_\alpha \subset S_{\alpha+1} \subset \dots \subset S_\beta = S$

Def 38 (*Spohn Ordinal / Degree of Implausibility*): For every world $w \in S$, the Spohn ordinal $\text{ord}(w)$ of world w is defined as the *least ordinal* α such that $w \in S_\alpha$, it represents the “degree of implausibility” of w .

Def 39 (*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, \dots for weaker beliefs?)

Def 40 (*Updates in Grove models*): An *update* $!\varphi$ with a sentence φ is defined on full sphere models $\mathbf{S} = (S, \mathcal{F}, \|\cdot\|, s_*)$ similarly as on sphere-based epistemic-doxastic models (Def 19), except the family of spheres is restricted to worlds in $\|\varphi\|_S$, the new family of spheres is

$$\mathcal{F}' = \{S' \cap \|\varphi\|_S \mid S' \in \mathcal{F} : S' \cap \|\varphi\|_S \neq \emptyset\} \quad (31)$$

C.2 Plausibility Models

Remember

Def 41 (*Preorder*): Reflexive and transitive binary relation R on set S :

$$\forall s \in S : s \leq s \text{ and } \forall s, t, w \in S : (s \leq t \wedge t \leq w) \rightarrow s \leq w \quad (32)$$

Def 42 (*Totality of a binary relation*):

$$\forall s, t \in S : s \leq t \vee t \leq s \quad (33)$$

Def 43 (*Single-Agent Plausibility Model*): A plausibility model is a tuple $\mathbf{S} = \left(S, \leq_{\text{pl.}}, \|\cdot\|, s_\star \right)$

where:

- S is a non-empty set of states (*possible worlds*).
- $\leq \subseteq S \times S$ is a converse-well-founded total preorder (*plausibility order*).
- $\|\cdot\|$ assigns a set of worlds $\|p\|_{\mathbf{S}} \subset S$ to each $p \in \text{Prop}$ (*valuation*).

Intuition: Plausibility order

$s \leq_{\text{pl.}} t$ 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 37) and plausibility models (Def 43) are mathematically equivalent. The plausibility relation can be extracted via

$$s \leq_{\text{pl.}} t \Leftrightarrow \forall S' \in \mathcal{F} : (s \in S' \rightarrow t \in S') \quad (34)$$

An alternative statement using Spohn Ordinals (Def 38):

$$s \leq_{\text{pl.}} t \Leftrightarrow \text{ord}(s) \geq \text{ord}(t). \quad (35)$$

Conversely, spheres can be generated by

$$\mathcal{F} := \{w^\leq : w \in S\}; w^\leq = \left\{ s \in S : w \leq_{\text{pl.}} s \right\}. \quad (36)$$

Notation: Strict plausibility

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

$$\text{best}P = \max_{\leq_{\text{pl.}}} P := \left\{ s \in P \mid \forall t \in P : t \leq_{\text{pl.}} s \right\} \quad (37)$$

is also nonempty: $\text{best}P \neq \emptyset$.

Def 44 (*Interpretation Map on Plausibility models*): We extend the valuation $\|p\|_S$ to an interpretation map $\|\cdot\|_S$ for all propositional formulas using standard Boolean connectives.

- **Knowledge:** Truth in all possible worlds. A sentence φ is known iff its interpretation is the whole state space:

$$\|K\varphi\|_S = \{s \in S : \|\varphi\|_S = S\} \quad (38)$$

Meaning $\|K\varphi\|_S = S$ iff $\|\varphi\|_S = S$, and \emptyset otherwise.

- **Belief:** Truth in all the most plausible worlds.

$$\|B\varphi\|_S = \{s \in S : \text{best } S \subseteq \|\varphi\|_S\} \quad (39)$$

Meaning $\|B\varphi\|_S = S$ iff $\text{best } S \subseteq \|\varphi\|_S$, and \emptyset otherwise.

D Conditional Beliefs and The Logic of Knowledge

Def 45 (*Conditional Belief*): Let $P, Q \subseteq S$ be two propositions over a model \mathbf{S} and let φ, ψ be sentences. We say that at any world $s \in S$,

- $B^Q P$: P is believed conditional on Q , if P is true in the most plausible Q -worlds:

$$\text{best}Q \subseteq P \quad (40)$$

- $B^Q \psi$: ψ is believed conditional on Q , if ψ is true in the most plausible Q -worlds:

$$\text{best}Q \subseteq \|\psi\|_{\mathbf{S}} \quad (41)$$

- $B^\varphi \psi$: ψ is believed conditional on φ , if $\|\psi\|_{\mathbf{S}}$ is believed given $\|\varphi\|_{\mathbf{S}}$:

$$\|B^\varphi \psi\|_{\mathbf{S}} = \{s \in S : \text{best}\|\varphi\|_S \subseteq \|\psi\|_S\} \quad (42)$$

 **Intuition: Conditional Beliefs as Contingency Plans**

Think of $B^\varphi \psi$ as contingency plans for belief change:

In case find out φ , change belief to ψ .

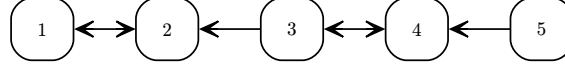
 **Note: Conditional Belief as Belief Revision**

We can semantically capture the AGM(Def 34) revision operator using conditional belief (Def 45): $T * \varphi := \{\theta : s_* \models B^\varphi \theta\}$. This interpretation guarantees that all modified AGM axioms are sound.

Example 32: Surprise Exam

When drawing a plausibility model:

- Preorder: Assume reflexivity and transitivity \Rightarrow (not drawn)
- $s \rightarrow t: t \leq_{\text{pl.}} s$ (e.g., $4 \leq_{\text{pl.}} 5$ and $5 \rightarrow 4$).

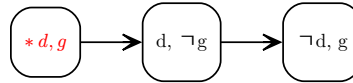


- $B(1 \vee 2), B^{\neg(1 \vee 2)}(3 \vee 4)$

Example 33: Professor Wine - Formalization

Professor Wine knows there are only two explanations for feeling like a genius: he is a genius (g) or he's drunk (d). He doesn't feel drunk, so he believes he is a sober genius. If he realized he was drunk, he would conditionally believe his genius feeling was just the drink (drunk non-genius). In reality, he is both drunk and a genius.

- **Assumptions:** $Bg, Kg \vee d, B\neg d, B^d\neg g$, and $d \wedge g$.
- **The Model:** Worlds are $(d, g), (d, \neg g), (\neg d, g)$. The actual world is (d, g) .
- **Plausibility Order:** $(\neg d, g) < (d, \neg g) < (d, g)$. No $(\neg d, \neg g)$ world exists because he knows $Kg \vee d$.
- **Conclusion:** He believes he is a genius ($(d, g) \models Bg$), but he does not **know** it, since $(d, \neg g)$ is a possible world. True belief is not knowledge because it can be lost upon learning new facts (like learning d).



Theorem 7: Full Introspection of Conditional Beliefs

Strong introspection holds for knowledge and standard beliefs, and it importantly extends to conditional beliefs:

$$B^\psi \varphi \rightarrow KB^\psi \varphi \quad (43)$$

$$\neg B^\psi \varphi \rightarrow K\neg B^\psi \varphi \quad (44)$$

Note: Knowledge vs. True Belief

Knowledge implies true belief ($K\varphi \rightarrow \varphi \wedge B\varphi$), but the converse is false in this setting.

Not every true belief qualifies as “knowledge”, as it lacks stability against belief revision. (Recall Example 33)

E Kripke Semantics and Belief Modalities

⚠ Attention: Difference from Standard Kripke Semantics

While plausibility models are single-agent Kripke models, the semantics of belief is **not** given by the standard Kripke semantics for the plausibility relation \leq_{pl} .

Belief is **not** the Kripke modality for the plausibility relation.

❓ Question: Knowledge as a Kripke Modality

For a set S , we must have:

$$s \models K\varphi \Leftrightarrow \forall t : (sR_{\text{knowledge}}t \rightarrow (t \models \varphi)) \quad (45)$$

📖 Theorem 8: Belief and Conditional Belief as Kripke Modalities

We can define appropriate accessibility relations to make belief a Kripke modality:

- **Doxastic Accessibility Relation** R_{belief} : $sR_{\text{belief}}t \Leftrightarrow t \in \text{best}S$
Then $s \models B\varphi$ iff $\forall t (sR_{\text{belief}}t \rightarrow t \models \varphi)$.
Endowed with R_{belief} , plausibility models become **KD45** doxastic models.
- **Conditional Doxastic Accessibility Relation** R_{belief}^ψ :

$$sR_{\text{belief}}^\psi t \Leftrightarrow t \in \text{best } \|\psi\|_S \quad (46)$$

Then conditional belief is a Kripke modality: $s \models B^\psi\varphi$ iff $\forall t (sR_{\text{belief}}^\psi t \rightarrow t \models \varphi)$.

✍ Note: Seriality and Consistency of Conditional Belief

The conditional doxastic arrows R_{belief}^ψ are **not necessarily serial**. They are serial only if the condition ψ is consistent with the agent's knowledge (i.e., $\neg K\neg\psi$ or $\|\psi\|_S \neq \emptyset$).

- **Interpretation:** Revision is restricted by knowledge. If φ is known to be false, the agent should not be able to revise with φ .
- **Warning - Counterfactual vs. Conditional:**
 - *Counterfactually*, an agent may consistently imagine possibilities that go against their knowledge.
 - *Conditionally*, consistent beliefs must be based on possibilities consistent with their knowledge.

Interpreting AGM^K using conditional beliefs: Given a plausibility model $\mathbf{S} = (S, \leq_{\text{pl}}, \|\cdot\|, s_\star)$

$$T = \{\theta \in L \mid s_\star \models_{\mathbf{S}} B\theta\} \quad (47)$$

where L is the language of epistemic-doxastic logic.

Note: the agent's current theory T consists of all the sentences believed in model \mathbf{S} at the real world s_\star .

For a sentence $\varphi \in L$:

$$T \star \varphi := \{\theta \in L \mid s_\star \models_{\mathbf{S}} B^\varphi\theta\} \quad (48)$$

If we interpret revision with φ in terms of doxastic conditioning with φ , all the AGM^K axioms are sound. In fact AGM^K are a subset of the following.

Theorem 9: Axiomatization of Knowledge and Conditional Belief

The complete logic bridging knowledge and conditional belief includes:

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

F Dropping Well-foundedness

Intuition: Infinite Models and Generalization

In finite models, converse well-foundedness of $\leq_{\text{pl.}}$ is automatically satisfied. If we drop this condition for infinite cases (keeping only the totality of the preorder $\leq_{\text{pl.}}$), we can no longer guarantee the existence of “most plausible” states (best S might be empty).

Def 46 (*Belief in Non-Wellfounded Models*): We redefine conditional belief as “truth in all worlds that are plausible enough”:

$$\|B^\psi \varphi\|_S = \{s \in S : \exists w \in \|\psi\|_S (\|\psi\|_S \cap w^{\leq} \subseteq \|\varphi\|_S)\} \quad (49)$$

where $w^{\leq} = \left\{ t \in S : w \leq_{\text{pl.}} t \right\}$ is the sphere determined by w .

Theorem 10: Properties of Generalized Plausibility Models

- On converse-wellfounded models, this new definition is equivalent to the standard one.
- The logic of conditional beliefs on totally preordered (generalized) plausibility models is **exactly the same** as on converse-wellfounded plausibility models (the proof system remains sound and complete).
- **Key Differences:** Belief is **not** a Kripke modality in generalized models. Most importantly, the set of sentences that are believed may be **inconsistent** in such models, even though any finite subset of them is consistent.

Session 4-2 (Lecture): Dynamic Belief Revision in the general case

A Generalizing Dynamic Belief Revision

i Intuition: Intuition

Standard updates represent a very specific kind of learning where the new information is “hard” and comes with a warranty of truthfulness[cite: 15]. The goal of dynamic belief revision is to generalize this to “softer” forms of learning, answering how we can compute an agent’s new higher-order beliefs after learning new information that might not be absolutely certain.

Def 47 (*Belief Upgrade*): A belief upgrade is a model transformer T , defined as a partial map from a plausibility model $S = (S, \leq, \|\cdot\|, s_*)$ to a new plausibility model $T(S) = (S', \leq', \|\cdot\| \cap S', s_*)$. The upgrade is defined ONLY if the real world s_* survives it, meaning $s_* \in S'$. The domain where T can be performed is denoted as $\text{Dom}_{S(T)} := S'$.

Def 48 (*Soft vs Hard Upgrades*):

- **Soft Upgrade**: A total map where $\text{Dom}_{S(T)} = S$ for all models. It does not add “hard” knowledge but only changes the agent’s beliefs or belief-revision plans based on “soft information”.
- **Hard Upgrade**: Shrinks the state set to a proper subset $S' \subset S$, effectively adding new absolute knowledge[cite: 40].

B Formal Semantics of Upgrades

i Dynamic Operators

We extend the language with dynamic operators $[T]\psi$ to express that ψ will surely be true after upgrade T .

- Semantics: $s \models [T]\psi$ iff $s \in \text{Dom}_{S(T)} \Rightarrow s \models_{T(S)} \psi$.
- Dual modality: $\|\langle T \rangle \psi\|_s = \|\psi\|_{T(S)}$.

C Types of Upgrades and Trust

Different upgrades correspond to different attitudes toward the reliability of the information source.

Def 49 (*Update*): Denoted by $!\varphi$. Used when the source is infallible (guaranteed truthful). It deletes all non- φ states and keeps the same plausibility order among the remaining states[cite: 51].

Def 50 (*Radical Upgrade*): Denoted by $\uparrow \varphi$ (Lexicographic Revision). Used when the source is fallible but highly reliable (strongly believed). All φ -worlds become strictly more plausible than all non- φ -worlds, while the old ordering remains within the two respective zones.

Def 51 (*Conservative Upgrade*): Denoted by $\dagger \varphi$ (Minimal Revision). Used when the source is trusted “barely” (simply believed, but easily given up). Only the “best” φ -worlds become better than all other worlds; the old order remains everywhere else[cite: 54].

D Strong Belief

Def 52 (*Strong Belief*): A formula φ is strongly believed ($\text{Sb}(\varphi)$) if:

1. It is consistent with knowledge: $\|\varphi\|_S \neq \emptyset$.
2. All φ -worlds are strictly more plausible than all non- φ -worlds: $s > t$ for every $s \in \|\varphi\|_S$ and $t \notin \|\varphi\|_S$.

Strong belief is believed until proven wrong. It implies belief but is **not** closed under logical inference ($\text{Sb}(\varphi \wedge \psi) \neq \text{Sb}(\varphi) \wedge \text{Sb}(\psi)$).

E The Wine Example

Let d denote being drunk and g denote being a genius. The arrows \rightarrow point to strictly more plausible states.

Example 34: Initial Setup and Update

- **Initial Model:** $d, g \rightarrow d, \neg g \rightarrow \neg d, g$. Albert holds a strong false belief that he is sober ($\neg d$) [cite: 140].
- **Update** ($!d$): Albert sees a flawless blood test. The state $\neg d, g$ is deleted. The model becomes $d, g \rightarrow d, \neg g$. He now has knowledge that he is drunk: Kd [cite: 73].

Example 35: Radical vs Conservative Upgrades

Suppose instead Albert hears he is drunk from Mary Curry, a trusted but fallible friend[cite: 83].

- **Radical Upgrade** ($\uparrow d$): Albert strongly believes Mary[cite: 151]. We promote all d -worlds. Result: $\neg d, g \rightarrow d, g \rightarrow d, \neg g$. Albert’s strong belief reverts; he now strongly believes he is drunk[cite: 164].
- **Conservative Upgrade** ($\dagger d$): Albert has fragile trust in Mary[cite: 167]. We promote only the **most plausible** d -world, which was $d, \neg g$. Result: $d, g \rightarrow \neg d, g \rightarrow d, \neg g$. He acquires a weak belief in d ($Bd \wedge B^g \neg d$); if told he is a genius, he will revert to believing he was sober.

F Theorems and Properties

Theorem 11: Knowledge and Belief Induction

- Updates induce knowledge: $[!\varphi]K\text{BEFORE } \varphi$.
- Conservative upgrades induce belief: $\neg K\neg\varphi \Rightarrow [\dagger \varphi]B\text{BEFORE } \varphi$.
- Radical upgrades induce strong belief: $\neg K\neg\varphi \Rightarrow [\uparrow \varphi]\text{Sb}(\text{BEFORE } \varphi)$.
- Conservative upgrades are special radical upgrades: $\dagger \varphi = \uparrow (\text{best } \varphi)$.

Theorem 12: Compositionality

- Updates are closed under sequential composition: The sequence of $!\varphi$ then $!\psi$ is equivalent to $!(\varphi \wedge [!\varphi]\psi)$.
- Radical upgrades are **not** closed under composition. **Counter-example:** Doing $\uparrow(d \wedge g)$ followed by $\uparrow d$ on the initial Wine model results in $\neg d, g \rightarrow d, \neg g \rightarrow d, g$, which cannot be reached by any single radical or conservative upgrade.

Theorem 13: Reduction Laws

To fully axiomatize the logic, we must pre-encode dynamic behavior into static conditional beliefs.

- Belief under update: $[!\psi]B\varphi \Leftrightarrow (\psi \rightarrow B^\psi[!\psi]\varphi)$.
- Conditional belief under update: $[!\psi]B^\theta\varphi \Leftrightarrow (\psi \rightarrow B^{\psi \wedge [!\psi]\theta}[!\psi]\varphi)$ [cite: 247].

Session 4-3 (Lecture):

Session 4-4 Homework 1

This homework is worth 100 points in total. This homework is **individual** (so, **no collaboration!**). Please **type in capital letters your full name**. Your answers should be **typed** as PDF, except for the drawings, which can scanned hand-drawings integrated with your pdf.

1. (35 points) A non-standard version of the so-called *Singapore Problem* goes as follows:

Albert and Bernard just met Cheryl. “When is your birthday?” Albert asks Cheryl. Cheryl thinks a second and says, “Boys, I’m not going to tell you this, but I’ll give you some clues, see how smart you are. Then I’ll go out on a date with whoever of you finds the answer first”. Then she writes down a list of 10 dates:

May 15, May 16, May 19, June 14, July 17, July 18, August 14, August 15, August 17, September 14, September 16.

“My birthday is one of these,” she tells them. “Now I am going to tell to one of you the month (and only the month) of my birthday, and tell to the other the day (and only the day) of my birthday.” Then Cheryl whispers in Albert’s ear the month (and only the month) of her birthday. To Bernard, she whispers the day (and only the day).

“Do you think that Albert can figure out my exact birthday now?” she asks Bernard.

Bernard: “Huh! I know for sure that Albert doesn’t know when it is!”

Albert (after hearing Bernard’s statement): “Yes, but I also I know that right now poor Bernard doesn’t know it either!”

Bernard (after listening to Albert): “Well, well, guess what: now I know your birthday, but... I also know that Albert knows it too! So hmm, you won’t go out with either of us then?! Do you have another test for us? Or shall we throw a fair coin?”

The problem is: *When is Cheryl’s birthday?*

- (a) (7 points) First, start by representing (drawing) the epistemic situation immediately after Cheryl gives the boys their pieces of information (but before she starts questioning them). Represent all the facts, as well as each agent’s knowledge in this situation, using a Kripke model M with three agents $\mathcal{A} = \{\text{Albert, Bernard, Cheryl}\}$ and ten atomic propositions

$$\text{Prop} = \{\text{May, June, July, August, September, 14, 15, 16, 17, 18, 19}\}, \quad (50)$$

where the first four indicate “Cheryl’s birthday is in month x ” and the others indicate “Cheryl’s birthday is in day y ”. (You may disregard all the information about going out on a date, that is just for fun.) **Draw the model (as a graph)**, don’t just describe it in words or formulas!

Answer to Exercise 1. (a)

Let

$$\mathcal{A} = \{a, b, c\}$$

$$\text{Prop}_M = \{M, \text{Jn}, \text{Jl}, A, S\}$$

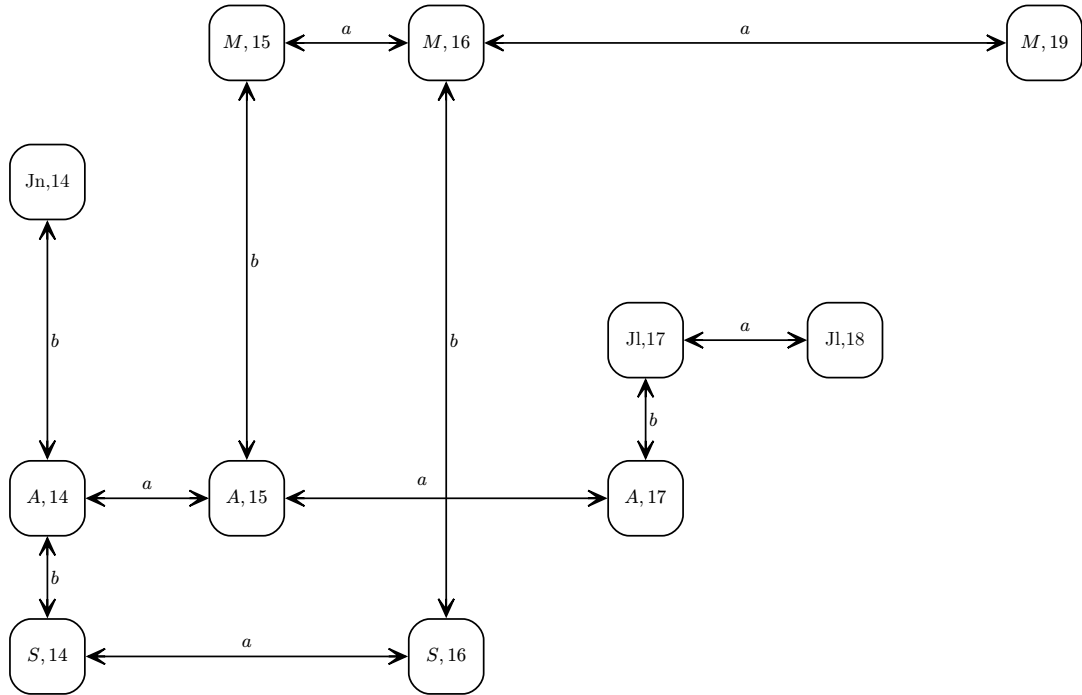
$$\text{Prop}_D = \{14, 15, 16, 17, 18, 19\}$$

$$\text{Prop} = \text{Prop}_M \cup \text{Prop}_D$$

$$S = \{(M, 15), (M, 16), (M, 19), (\text{Jn}, 14), (\text{Jl}, 17), (\text{Jl}, 18), (A, 14), (A, 15), (A, 17), (S, 14), (S, 16)\}$$

It is common knowledge that a knows the birthday month and b knows the date.
 $Ck\left(\bigvee_{m \in \text{Prop}_M} K_a m\right) \wedge Ck_{a,b,c}\left(\bigvee_{d \in \text{Prop}_D} K_b d\right)$.

The following is a graph describing epistemic model $\mathbf{S} = (S, \{\sim_\alpha\}_{\alpha \in \mathcal{A}}, \|\cdot\|)$. Since this is an epistemic **S5** model, all relations are reflexive, transitive, and thus symmetric. Reflexive and transitive arrows are omitted for visual clarity. Since Cheryl c knows her own birthday, all worlds form their own equivalence class under \sim_c ($\forall s = (m, d), t = (m', d') \in S : s \sim_c t \Leftrightarrow (m = m' \wedge d = d')$).



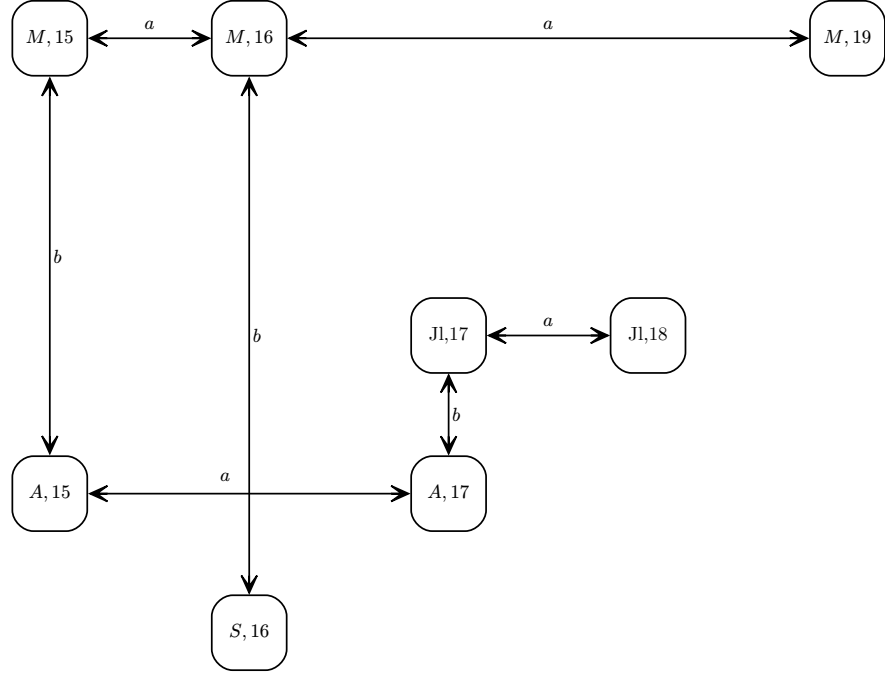
(b) (7 points) As for the Muddy Children, write down an epistemic sentence encoding Bernard’s first announcement (that he knows Albert doesn’t know when her birthday is). As for the Muddy Children, interpret Bernard’s first announcement as a truthful public announcement, and represent (draw) the updated model M' after this update.

Answer to Exercise 1. (b)

Bernard's first announcement encoded as an epistemic sentence is:

$$\varphi_1 = K_b \left(\bigwedge_{(m,d) \in S} (\neg K_a(m \wedge d)) \right) \quad (51)$$

The updated model S' :



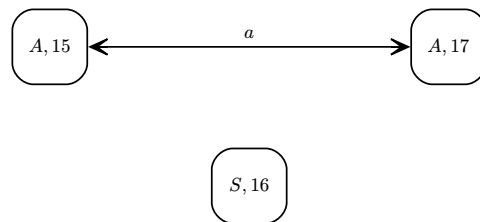
(c) (7 points) Do the same as in previous part for Albert's first announcement (that now he also knows that Bernard doesn't know Cheryl's birthday): encode his statement as an epistemic sentence, and represent (draw) the updated model M'' after this new update.

Answer to Exercise 1. (c)

Albert's first announcement encoded as an epistemic sentence is:

$$\psi_1 = \left(K_a \bigwedge_{(m,d) \in S} (\neg(m \wedge d)) \right) \wedge \left(K_a \bigwedge_{(m,d) \in S} (\neg K_b(m \wedge d)) \right) \quad (52)$$

The updated model S'' :



(d) (7 points) Do the same for Bernard's second announcement (that now he knows the birthday, but that he also knows that Albert knows the birthday as well), computing the updated model M'' . Use this to solve the puzzle: when is Cheryl's birthday?

Answer to Exercise 1. (d)

Bernard's second announcement encoded as an epistemic sentence is:

$$\varphi_2 = \left(\bigvee_{(m,d) \in S} K_b(m \wedge d) \wedge K_b \left(\bigvee_{(m,d) \in S} K_a(m \wedge d) \right) \right) \quad (53)$$

From S'' we can deduce that c 's birthday is September the 16th ($s_\star = (S, 16)$).

(e) (7 points) Write a sentence in the language of Public Announcement Logic (PAL) saying that: "after Bernard's first announcement, followed by Albert's first announcement, Bernard knows Cheryl's birthday and he also knows that Albert knows it".

Answer to Exercise 1. (e)

$$![\varphi_1][\psi_1]\varphi_2$$

2. (35 points) Alice and Bob have each some positive natural number $n_a, n_b \in \{1, 2, \dots, 12\}$ written on their forehead. It is common knowledge that
1. each of them can see the other's number (Alice can see n_b and Bob can see n_a), but neither of them can see his/her number;
 2. the two children are perfect logicians.
 3. both numbers are no larger than 12 (i.e. $1 \leq n_a, n_b \leq 12$);
 4. the two numbers are related by the function $g : \mathbb{N} \rightarrow \mathbb{N}$ given by: $g(n) = 1$, if $n = 2^k$ for some $k > 0$; $g(n) = n + 2$, if n is odd; $g(n) = n - 2$ if n is even but not a power of 2.

$$g(n) = \begin{cases} 1 & \text{if } n = 2^k \text{ for some } k > 0 \\ n + 2 & \text{if } n \text{ is odd} \\ n - 2 & \text{if } n \text{ is even, but not a power of two} \end{cases} \quad (54)$$

4. So it is common knowledge that either $n_a = g(n_b)$ or else $n_b = g(n_a)$.

The Father asks them, repeatedly: "Do you know your own number?". The two are supposed to answer truthfully, publicly, simultaneously (without any other communication). They both answer "I don't know" to the first 3 questions, after which Alice answers "Yes, now I know my number" to the 4th question (while Bob still answers "I don't know").

(a) (10 points) What is Alice's number?

Answer to Exercise 2 (a)

$$n_a = 3.$$

- (b) (5 points) Will Bob ever know his number (without looking in the mirror or being told the number by Alice)? If so, when will he answer "I know my number"?
- (c) (20 points) Prove your conclusions in (a) and (b) semantically, by drawing the initial epistemic model, then applying repeated updates with the (semantic information conveyed by the) children's answers. Draw each intermediary model, and explain your conclusions.

Answer to Exercise 2 (c)

All values of $g(n)$ computed:

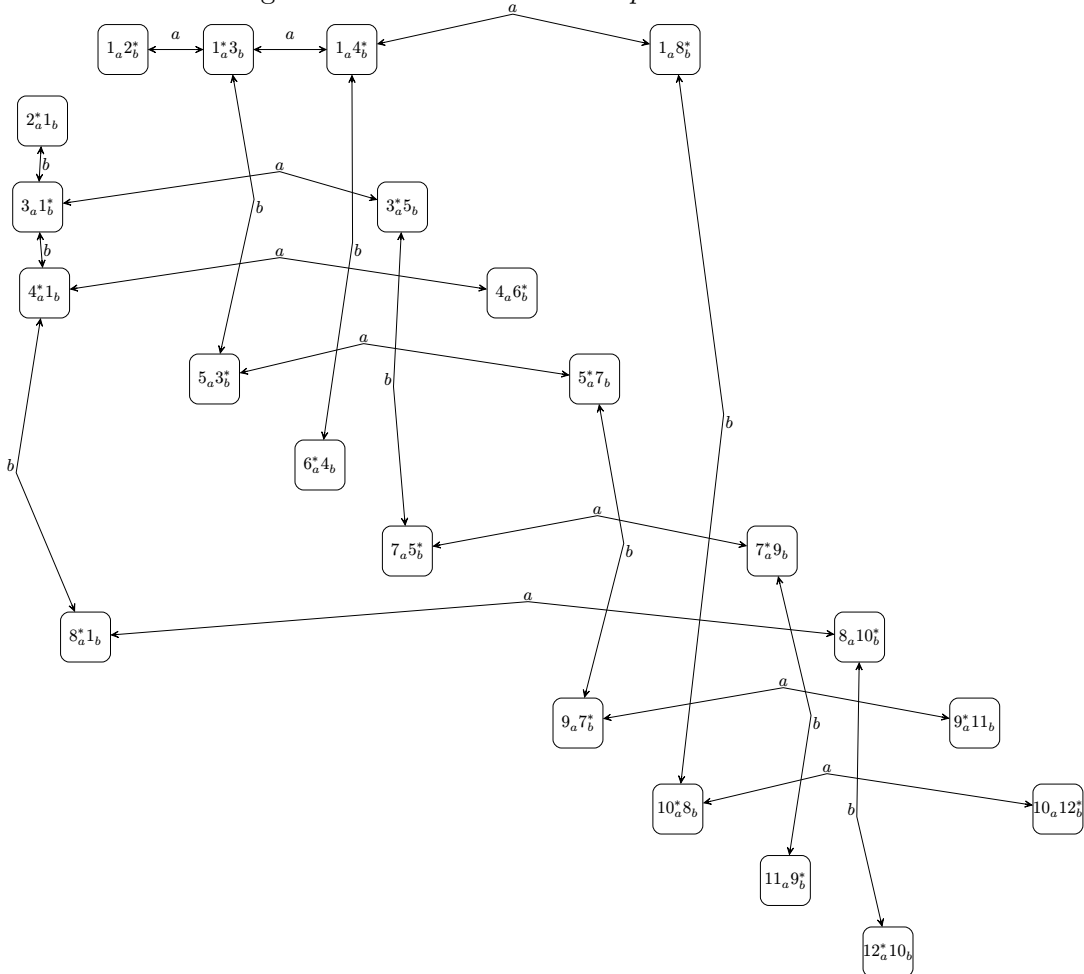
- $n = 1, g(n) = 3$
 $n = 2, g(n) = 1$
 $n = 3, g(n) = 5$
 $n = 4, g(n) = 1$
 $n = 5, g(n) = 7$
 $n = 6, g(n) = 4$
 $n = 7, g(n) = 9$
 $n = 8, g(n) = 1$
 $n = 9, g(n) = 11$
 $n = 10, g(n) = 8$
 $n = 11, g(n) = 13 > 12 \nmid$
 $n = 12, g(n) = 10$

(55)

Either $x = a$ and $y = b$ or vice versa

	possible values for n_x		
x sees	$i = 1$	$i = 2$	$i = 3$
$n_y = 1$	{2, 3, 4, 8}	{}	{}
$n_y = 2$	{1}	{}	{}
$n_y = 3$	{1, 5}	{}	{}
$n_y = 4$	{1, 6}	{}	{}
$n_y = 5$	{3, 7}	{}	{}
$n_y = 6$	{4}	{}	{}
$n_y = 7$	{5, 9}	{}	{}
$n_y = 8$	{1, 10}	{}	{}
$n_y = 9$	{7, 11}	{}	{}
$n_y = 10$	{8, 12}	{}	{}
$n_y = 11$	{9}	{}	{}
$n_y = 12$	{10}	{}	{}

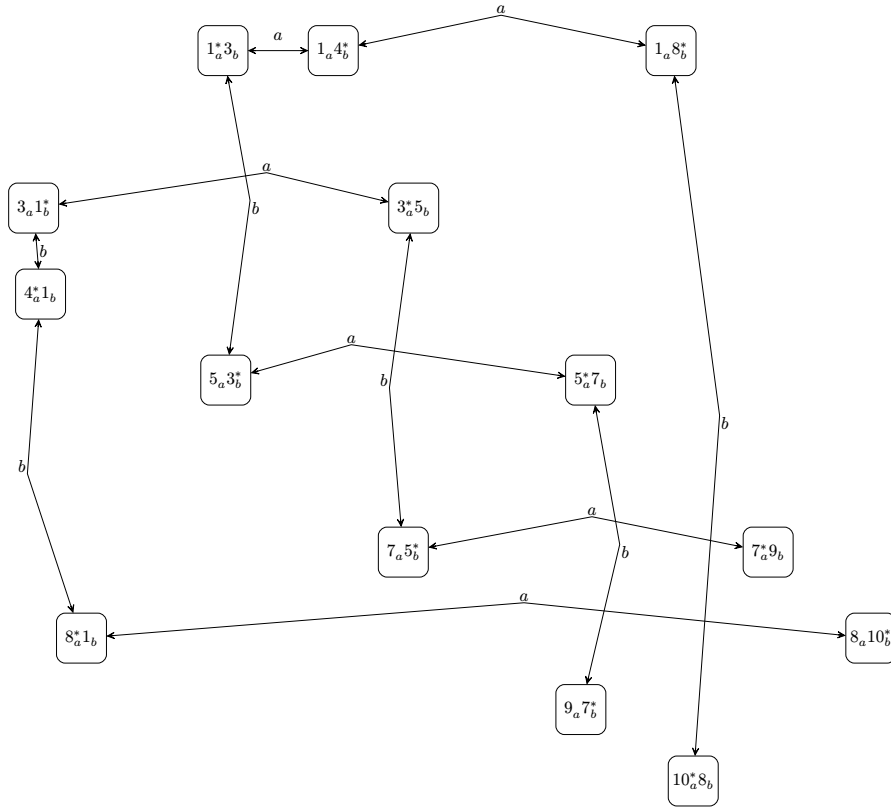
I fought with dot for a long time but the edges just don't want to be straight lines. So it looks ugly but it contains the right information. *Initial set up:*



Answer to Exercise 2 (c) continued

Iteration 1: After the first announcements of ignorance.

Eliminations:

**Answer to Exercise 2 (c)**

test

1. (30 points)

1. (15 points) Is the following formula valid or not on epistemic models with two agents (a and b)? $K_a(Dp \vee D \text{ not } p) \Rightarrow D(K_a p \vee K_a \text{ not } p)$ If your answer is yes, then give a semantic argument for this (by reasoning on possible worlds in any arbitrary epistemic model, using the semantic

clauses for K , D_k and propositional connectives). If your answer is no, then provide a counterexample (by drawing an epistemic model, and presenting

September 16

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	Updates on Sphere Models	21
Def 28	Updates on Kripke Models	21
Def 29	PDL Dynamic Modalities	22
Def 30	Dynamic Epistemic Logic	23
Def 31	DEL semantics of the dynamic modalities	23
Def 32	Public Announcement as a Model Transformer	24
Def 33	Moore Sentences	26
Def 34	AGM Postulates for Belief Revision	31
Def 35	Well-foundedness	33
Def 36	Converse well-foundedness	33
Def 37	Single-Agent Sphere Model for Belief Revision (Grove Model)	33
Def 38	Spohn Ordinal / Degree of Implausibility	33
Def 39	Belief and Knowledge in Grove models	33
Def 40	Updates in Grove models	33
Def 41	Preorder	34
Def 42	Totality of a binary relation	34
Def 43	Single-Agent Plausibility Model	34
Def 44	Interpretation Map on Plausibility models	35

Def 45 Conditional Belief	35
Def 46 Belief in Non-Wellfounded Models	38
Def 47 Belief Upgrade	39
Def 48 Soft vs Hard Upgrades	39
Def 49 Update	39
Def 50 Radical Upgrade	39
Def 51 Conservative Upgrade	40
Def 52 Strong Belief	40

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 Complete Axiomatizations & Reduction Axioms	26
Theorem 6 Closure Under Composition	28
Theorem 7 Full Introspection of Conditional Beliefs	36
Theorem 8 Belief and Conditional Belief as Kripke Modalities	37
Theorem 9 Axiomatization of Knowledge and Conditional Belief	38
Theorem 10 Properties of Generalized Plausibility Models	38
Theorem 11 Knowledge and Belief Induction	40
Theorem 12 Compositionality	41
Theorem 13 Reduction Laws	41