Dynamic Epistemic Logic

Study Notes

Filip Rehburg[1]

filip.rehburg@student.uva.nl

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

February 20, 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

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	Θ
Multi-agent Models and Public Announcement Logic (PAL) : Section 2.1	✗
PAL Continued : Section 2.2	✗
Does this one even exist?? : Section 2.3	✗
"Learnability" and "Knowability" : Section 3.1	✗
The problem of belief revision : Section 3.3	✗
: Section 3.3	✗
: Section 4.2	✗
: Section 4.3	✗
: Section 5.2	✗
: Section 5.3	✗
: Section 6.2	✗
: Section 6.3	✗
: Section 7.2	✗
: Section 7.3	✗
: Section 8.2	✗
Tutorial 1 : Section 3.2	✗
: Section 4.1	✗
: Section 5.1	✗
: Section 6.1	✗
: Section 7.1	✗
: Section 8.1	✗

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 [cite_start] or [cite: x] in your output
2. I have defined custom functions to represent definitions, theorems (“theorem”), proofs (“proof”), examples (“example”), intuitions (“intuition”), warnings to watch out (“attention”), questions (“question”) and calls to remember (“remember”).
   - To define a new concept, call `#def(“Name of Concept”)[Definition body]`
   - For all others call `#box(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

1. Week 1

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

1.1.1. Core Intuitions and Definitions

Example: 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?

Motto of Dynamic Epistemic Logic

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

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).

1.1.2. 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: 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: 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

1.2.1. Epistemic Puzzles and Paradoxes

Example: Puzzle 0: The Coordinated Attack

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

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

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

Remember: Fixpoints and Byzantine Generals

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

In the case of Puzzle 0:

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

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

Intuition: Coordinated Attack Intuition

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

Example: 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 phi $\Box\varphi$, 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: 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.

1.2.2. The Muddy Children and Epistemic Updates

Example: Puzzle 4: Muddy Children

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

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

Socratic Questioning

Discovering answers by asking questions of students. (Wikipedia)

Intuition: Muddy Children

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

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

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

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

Example: Modifications of Muddy Children

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

Example: 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).

1.2.3. Paradoxes of Induction and Probability

Example: Puzzle 6: Surprised Children (Unexpected Hanging)

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: 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**.

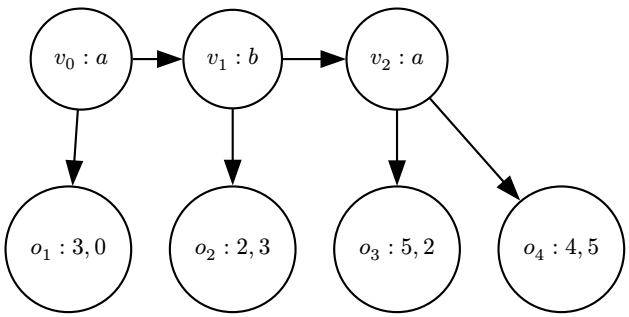
Example: Puzzle 7 Modification: The Infinite Lottery

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.

1.2.4. Backward Induction and Social Epistemology

Example: Puzzle 8: The Centipede Game

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



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

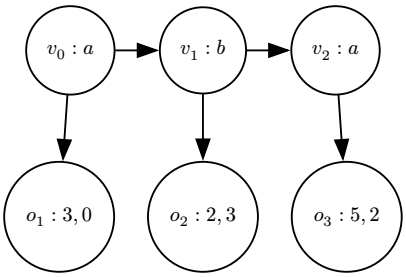
- $v_0 : a$ stops for $o_1(3, 0)$ or continues to v_1

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

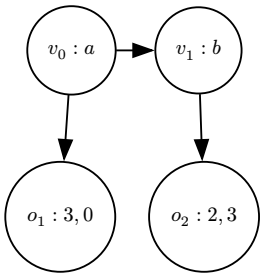
The Backwards Induction (BI) Method

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

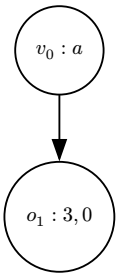
Elimination Step 1



Elimination Step 2



Elimination Step 3



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

Intuition: The BI Paradox and Rational Pessimism

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

Example: 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.
  Pluralistic Ignorance: Group members privately reject a norm but incorrectly assume others accept it (e.g., no one asking questions in a confusing lecture).
  Informational Cascades: Sequential decision-making where rational agents ignore their private signals to follow public actions (e.g., sequentially guessing urn colors based on previous skewed guesses).
  The Circular Mill: Biological equivalent where army ants follow the ant in front, creating an endless, fatal loop.
  The Human Mill: Cold War arms races driven by circular, self-fulfilling falsehoods (e.g., nations mimicking adversary research based entirely on forged intelligence).

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

2. Week 2

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

Session 2-2 (Lecture): PAL Continued

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

3. Week 3

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

Session 3-2 (Lecture): Tutorial 1

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

4. Week 4

Session 4-1 (Lecture):

Session 4-2 (Lecture):

Session 4-3 (Lecture):

5. Week 5

Session 5-1 (Lecture):

Session 5-2 (Lecture):

Session 5-3 (Lecture):

6. Week 6

Session 6-1 (Lecture):

Session 6-2 (Lecture):

Session 6-3 (Lecture):

7. Week 7

Session 7-1 (Lecture):

Session 7-2 (Lecture):

Session 7-3 (Lecture):

8. Week 8

Session 8-1 (Lecture):

Session 8-2 (Lecture):

Session 8-3 (Lecture):