

Recursive Theory of Mind in an Abstract Epistemic Environment: A Dynamic Epistemic Logic Approach

Baran Akkanat

Abstract

Theory of Mind (ToM) refers to the capacity to represent and reason about the mental states of other agents, including their beliefs, knowledge, and intentions, and to use these representations to predict and explain behavior. In artificial intelligence, this ability is especially important in multi-agent settings with incomplete and asymmetric information. This project implements the core epistemic machinery of Dynamic Epistemic Logic (DEL) in a simplified executable setting for recursive belief reasoning. The system depicts epistemic states using possible worlds and accessibility relations, supports bounded recursive belief queries up to a configurable theory-of-mind depth, and operates in an abstract event-driven environment with discrete enter, leave, and move events. Implementation is evaluated through scripted, deterministic scenarios that cover false-belief reasoning, first-, second-, and third-order observability-based belief revision, and bounded reasoning amid heterogeneous ToM depths. The results show that the model replicates core false-belief patterns in a controlled and reproducible setting, while also revealing limitations in higher-order belief repair and scalability. Overall, the project demonstrates that a simplified DEL-style implementation can provide an interpretable, reproducible framework for studying recursive Theory of Mind in artificial agents.

Contents

1	Introduction	3
2	Background and Related Work	3
3	Existing Framework and Project Novelty	4
3.1	Existing Framework Implemented	4
3.2	Explicit Departures from Standard DEL	4
3.3	Novel Environment	4
3.4	Expected Effects on Theory-of-Mind Reasoning	5
3.5	Project Contribution	5
4	Environment Design	5
4.1	Entities and World Facts	5
4.2	Event Types	6
4.3	Observability Assumption	6
4.4	Rationale for the Abstract Environment	7
5	Model and Implementation	7
5.1	Agent Representation	7
5.2	Theory-of-Mind Depth	7
5.3	Epistemic Models and Semantics	8
5.4	Bounded Recursion via ToM Depth	9
5.5	Observability-Gated Belief Updates	9
5.6	Implementation-Level View: Update and Query Procedures	9
6	Experimental Setup	10
6.1	Hypotheses	10
6.2	Scenario Design	11
6.3	Evaluation Protocol	11
6.4	Correctness Criteria	11
6.5	Repeatability	11
6.6	Multiple Tests Across Capabilities	11
7	Results	12
7.1	Overview of Scenario Outcomes	12
7.2	First-Order False Belief	13
7.3	Second-Order Belief Attribution	13
7.4	Higher-Order Recursive Beliefs	13
7.5	Belief Repair After Re-entry: S4a and S4b	13
7.6	Relation to Background Literature	13
8	Discussion	14
8.1	Implications for Theory-of-Mind Reasoning	14
8.2	Implications for Human-Agent Collaboration	14
8.3	Interpretability and Methodological Value	14
8.4	Comparison to Existing Literature	15
9	Limitations and Future Work	15
10	Conclusion	15

1 Introduction

Theory of Mind (ToM) refers to the ability to attribute mental states such as beliefs, knowledge, and perspectives to other agents and to reason about how these internal states influence behavior. In human cognition, this ability is essential for social interaction, communication, cooperation, and deception. In artificial intelligence, Theory of Mind has become increasingly important for agents operating in multi-agent or human-centered environments, where successful interaction often requires anticipating what other agents know, observe, or mistakenly believe [5].

A central difficulty in computationally modeling Theory of Mind is the representation of higher-order beliefs. While first-order beliefs are about the state of the world itself, higher-order beliefs involve reasoning about what an agent believes about the beliefs of another agent. These nested belief structures can grow rapidly in complexity, creating significant representational and computing challenges.

In artificial intelligence, false-belief reasoning is often studied within environments that also include perception noise, physical simulation, or learning dynamics [5, 6]. Although such components are important for realism, they can make it difficult to isolate the epistemic mechanisms responsible for belief attribution. To address this, the present project adopts an abstract epistemic approach in which belief states are explicitly represented and updated based on event observability.

The project makes three main contributions. First, it implements the core epistemic machinery of Dynamic Epistemic Logic in an executable setting for recursive belief reasoning. Second, it delivers a minimal event-driven environment in which observability is manipulated through agent entry, exit, and object movement. Third, it evaluates implementation using deterministic, scripted scenarios that cover first-, second-, and third-order false-belief reasoning. Together, these design choices provide a controlled and reproducible setting for studying recursive Theory of Mind under asymmetric observability.

The remainder of the report first situates the project within the literature on symbolic Theory-of-Mind modeling, then describes the framework and environment, and finally evaluates the system through deterministic scenario-based experiments.

2 Background and Related Work

Theory of Mind has long been studied in developmental psychology, where false-belief tasks are commonly used to assess whether individuals can reason about others’ beliefs [2]. Classical experiments such as the Sally-Anne task demonstrate that successful social reasoning requires representing beliefs that may differ from reality due to unequal access to information [2]. Extensions of these tasks have also been used to study higher-order belief reasoning, in which individuals must reason about what one agent believes about another agent’s belief.

Another important direction is the use of symbolic epistemic models, in which beliefs are explicitly represented. Dynamic Epistemic Logic (DEL) offers a formal framework for representing agents’ knowledge and beliefs using possible worlds and accessibility relations, and for modeling how these beliefs change when events occur [1, 3, 4]. In DEL-based models, actions are represented as structured updates to epistemic states, allowing the explicit representation of nested beliefs such as $B_A(B_B\varphi)$ [3, 4].

DEL has been used to formalize false-belief reasoning within multi-agent systems and to analyze how observability affects belief updates [4, 6]. In particular, DEL-based implementations have demonstrated that explicit epistemic representations can successfully reproduce classical false-belief scenarios, including higher-order belief tasks [4, 6]. These models offer strong interpretability because belief updates follow clearly defined logical rules rather than being implicitly learned from data.

More recently, learning-based approaches have attempted to infer other agents’ beliefs using neural networks and reinforcement learning. While such approaches have achieved promising results inside complex environments, they often offer limited transparency into agents’ internal belief structures. In contrast, symbolic epistemic models emphasize interpretability and correctness of belief reasoning.

Building on this symbolic tradition, the present work implements a DEL-style reasoning framework within a deliberately simplified abstract environment. By focusing on event-driven belief updates under regulated observability conditions, the project isolates the epistemic mechanisms responsible for recursive belief attribution while continuing a reproducible experimental setup.

3 Existing Framework and Project Novelty

3.1 Existing Framework Implemented

This project implements the core epistemic machinery of Dynamic Epistemic Logic (DEL) for recursive Theory-of-Mind reasoning in a simplified executable setting. In DEL-based models, agents’ beliefs are represented as epistemic states comprising possible worlds and accessibility relations. These structures capture which worlds each agent considers possible, given its observations. When events occur, the epistemic state is updated to reflect how different agents revise their beliefs depending on whether they observed the event [3, 4].

The present implementation follows this general DEL tradition by explicitly representing epistemic states, preserving agent-dependent uncertainty, and evaluating nested belief queries directly over the epistemic model. In this sense, the framework is not merely behaviorally inspired by DEL, but operationally grounded in its central representational ideas.

3.2 Explicit Departures from Standard DEL

At the same time, the implementation does not reproduce the full DEL event-model formalism in its most general form. Instead, it adopts a simplified event-driven update mechanism created for executable recursive belief reasoning. Each event specifies both a world update and the set of agents that observe the event. Agents that observe the event revise their beliefs accordingly, whereas agents that do not observe it retain their previous epistemic state.

This simplification was chosen deliberately. The aim of the project is not to present a new logical formalism, but to construct a controlled and interpretable implementation in which false beliefs and recursive belief attributions can be tested directly. The resulting model keeps the key epistemic distinction among observed and unobserved events while remaining computationally manageable and easy to inspect.

3.3 Novel Environment

Rather than modeling a physically realistic world with noisy perception, embodiment, or continuous dynamics, the framework functions in a minimal symbolic environment comprising agents, objects, and discrete locations. The environment is driven by a small set of events: agent entry, agent exit, and object movement. This abstraction isolates the epistemic reasoning process and avoids confounding effects caused by perception noise or physical simulation.

The environment is, therefore, novel not because it introduces new logical operators, but because it creates a controlled executable setting in which recursive Theory-of-Mind reasoning can be studied under uneven observation histories. This enables testing of false-belief persistence, higher-order belief attribution, and bounded recursive reasoning within a compact, reproducible environment.

3.4 Expected Effects on Theory-of-Mind Reasoning

The design of the environment is expected to produce several characteristic Theory-of-Mind effects.

- Agents that fail to observe an object movement should retain outdated first-order beliefs about the world. Without this mechanism, all agents would meet on the ground truth, and no epistemic asymmetry could emerge, making false-belief reasoning impossible to study.
- Agents that observed the event should be able to attribute false beliefs to agents who were absent. This second-order attribution is the core capability that distinguishes Theory-of-Mind reasoning from simple world-state tracking.
- Asymmetric observation histories should generate second-order and third-order belief asymmetries. These nested asymmetries arise inherently from the compositionality of the accessibility-relation structure and provide the basis for evaluating higher-order belief queries.
- Re-entry into the environment should enable belief revision only when new relevant observations become available. This constraint ensures that belief repair is grounded in perception rather than in implicit knowledge transfer, preserving the epistemic integrity of the model.
- Increasing the theory-of-mind depth parameter should allow the system to answer more deeply nested belief queries. This reflects the intuition that greater cognitive capacity enables richer social reasoning, at the cost of increased representational complexity.

3.5 Project Contribution

The main contribution of this project is an interpretable, reproducible implementation of recursive belief reasoning in a controlled, abstract epistemic environment. The system supports bounded recursive belief queries up to a configurable theory-of-mind depth and reproduces first-order and higher-order false-belief patterns through deterministic scripted scenarios. By combining explicit epistemic representation with a minimal event-driven environment, the framework delivers a clear testbed for studying recursive Theory of Mind in artificial agents.

4 Environment Design

4.1 Entities and World Facts

The environment consists of a finite set of agents $\mathcal{A}$, a set of objects $\mathcal{O}$, and a set of symbolic locations $\mathcal{L}$. The world state is described using propositional facts that encode the location of objects. For an object $o \in \mathcal{O}$ and a location $\ell \in \mathcal{L}$, the predicate

$$\text{In}(o, \ell)$$

denotes that object o is located in location ℓ . Agents do not have direct access to the ground truth; instead, they maintain internal epistemic states that encode their beliefs about the world. In the experiments used here, the environment contains a small number of agents (typically A , B , and optionally C), a single tracked object, and two possible locations. This minimal configuration is sufficient to construct false-belief scenarios as keeping the epistemic structure easy to analyze.

4.2 Event Types

Changes in the environment are triggered by discrete events. Rather than modeling a complex action space, the framework uses a small set of event types sufficient to generate belief updates and false-belief situations.

The following event types are used:

- $\text{Enter}(i)$: agent $i \in \mathcal{A}$ enters the environment.
- $\text{Exit}(i)$: agent i leaves the environment.
- $\text{Move}(i, o, \ell \rightarrow \ell')$: agent i moves object o from location ℓ to location ℓ' .

Each event may change the environment’s ground truth, the set of agents present in the environment, or both. Sequences of such events form the action histories used in the experimental scenarios.

4.3 Observability Assumption

The key mechanism enabling false beliefs in this framework is observability. Agents can only update their beliefs based on events that they observe. In the simplified environment used in this project, observability is determined by co-presence: agents that are present in the environment observe all events, while those absent observe none.

This assumption allows the construction of classical false-belief scenarios. When an agent leaves the environment, and an object is moved during its absence, that agent does not observe the move and therefore maintains an outdated belief about the object’s location. This mechanism directly produces the belief asymmetries required for Theory-of-Mind reasoning [4, 6].

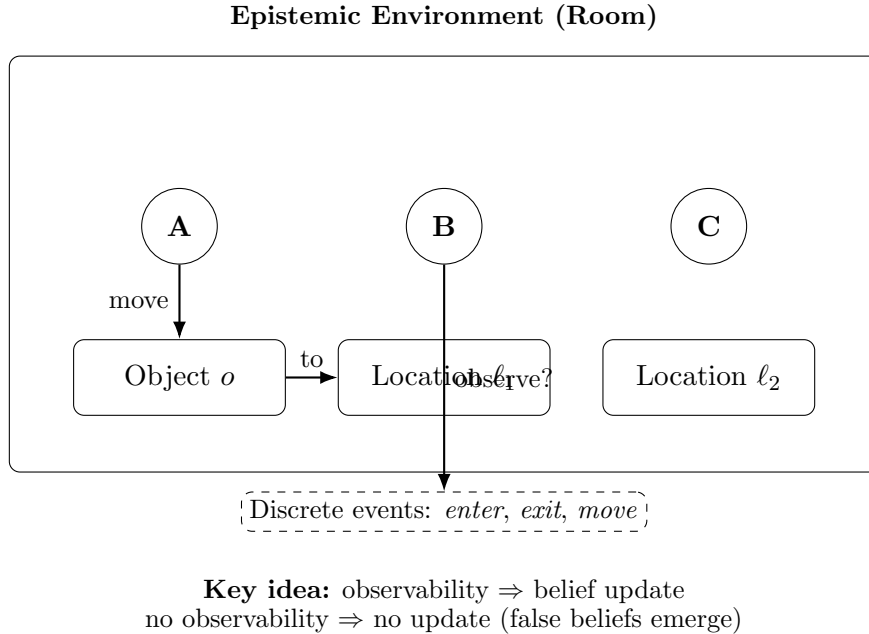


Figure 1: Conceptual overview of the abstract epistemic environment. Agents interact using discrete events (enter/exit/move). Observability determines which agents update their beliefs, enabling false beliefs amid asymmetric observation histories.

4.4 Rationale for the Abstract Environment

The environment is intentionally simplified to isolate the epistemic mechanisms underlying Theory-of-Mind reasoning. In more realistic environments, perception noise, continuous dynamics, and embodiment can make it difficult to determine whether incorrect beliefs arise from epistemic reasoning or from sensing limitations. By adopting a symbolic event-driven environment, the framework focuses directly on how belief states evolve under asymmetric observation histories. False beliefs therefore arise from differences in observation rather than uncertainty in the physical world [5, 6]. This design decision is directly motivated by the requirements of Theory-of-Mind research. In DEL-based models, the emergence of false beliefs depends entirely on the divergence between an agent’s observation history and the true world state. A symbolic, event-driven environment makes this divergence transparent and controllable: each false belief can be traced back to a specific unobserved event, and each belief update can be verified against the model’s formal semantics. This level of traceability would be difficult to achieve in physically realistic environments, where incorrect beliefs could arise from sensor noise or action failure rather than from epistemic asymmetry alone.

5 Model and Implementation

5.1 Agent Representation

Each agent maintains its own epistemic representation of the world. Agents do not have direct access to the true state of the environment; instead, they update their internal belief state based on the events they observe. As a result, different agents may have different beliefs about the world at the same time.

Internally, the system represents epistemic states using a possible-world structure. Each world corresponds to a possible configuration of object locations, while accessibility relations encode which worlds an agent considers possible [1, 4]. This representation allows the system to evaluate nested belief statements such as $B_A(\varphi)$ and $B_A(B_B(\varphi))$ directly within the epistemic model.

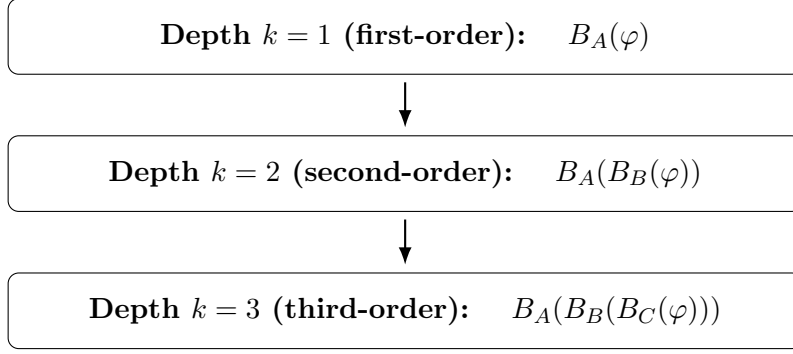
This agent design was chosen because Theory of Mind requires an explicit representation of unobservable mental content rather than observable behaviour alone. Alternative approaches, such as behaviour-based or reward-driven models, do not provide direct access to internal belief structures, making it difficult to verify whether nested belief attributions are correct. By contrast, the possible-world representation makes belief states directly inspectable and allows the evaluation of higher-order queries such as $B_A(B_B(\varphi))$, which is essential for testing recursive Theory-of-Mind reasoning in a controlled and reproducible way.

5.2 Theory-of-Mind Depth

Agents are parameterized by a theory-of-mind depth parameter $k \in \mathbb{N}$ that determines the maximum depth of recursive belief reasoning supported by the system. Informally:

- $k = 1$ supports first-order beliefs about the world.
- $k = 2$ supports second-order beliefs (beliefs about beliefs).
- $k \geq 3$ supports deeper recursive belief structures.

This parameter bounds the nesting depth of belief operators that the agent can evaluate. Queries exceeding the configured depth return **undefined**, representing limited recursive reasoning capacity.



Increasing k increases representational ability for recursive belief attribution.

Figure 2: Illustration of theory-of-mind depth as bounded recursive belief nesting. The parameter k determines the maximum nesting depth supported by the agent.

An important design aspect of the agent is that its internal epistemic state can be inspected by explicit belief queries [4, 6]. For example, a first-order query may ask whether an agent believes that an object is in a particular location, expressed as $B_i(\text{In}(o, \ell))$. Higher-order queries extend this structure by nesting belief operators, such as $B_i B_j(\text{In}(o, \ell))$. This query interface provides a direct and interpretable way to evaluate the agent’s reasoning against the observed action history.

Symbol	Meaning
$\mathcal{A}$	set of agents
$\mathcal{O}$	set of objects
$\mathcal{L}$	set of locations/containers
$\text{In}(o, \ell)$	object-location fact
$B_i \varphi$	agent i believes φ
k	theory-of-mind depth (max belief nesting)

Table 1: Core notation used throughout the model.

The depth parameter of the theory-of-mind directly affects the agent’s behavioral capabilities in collaborative settings. An agent with $k = 1$ can only reason about the true world state and its own first-order beliefs; it cannot anticipate what another agent mistakenly believes, and therefore cannot adapt its behavior to correct or exploit another agent’s false belief. In contrast, an agent with $k = 2$ can recognize that another agent has outdated beliefs and act accordingly — for example, redirecting a collaborator to the correct location of the object rather than assuming shared knowledge. An agent with $k \geq 3$ can model chains of mutual belief, permitting more sophisticated coordination and deception-detection in multi-agent settings. This behavioral stratification makes the depth parameter a practical tool for modeling heterogeneous cognitive capacities among agents.

5.3 Epistemic Models and Semantics

The system portrays the agents’ beliefs using a simplified epistemic model inspired by Dynamic Epistemic Logic (DEL). In this representation, the belief state of the system consists of a set of possible worlds, together with relations indicating which worlds each agent considers possible

[1, 3, 4]. Intuitively, each world represents a possible configuration of object locations, while accessibility relations capture each agent’s uncertainty about the true state of the environment.

When agents observe events, their belief relations are updated to eliminate worlds that are no longer consistent with the new observations. In contrast, agents that do not observe an event retain their previous belief relations and may therefore continue to consider outdated worlds possible. This mechanism allows the model to represent the divergence of beliefs among agents with different observation histories [6].

This representation permits the system to evaluate nested belief queries such as $B_i\varphi$ or $B_iB_j\varphi$ directly within the epistemic model. In practice, this means the agent can reason not only about the environment’s state but also about what other agents believe about it.

5.4 Bounded Recursion via ToM Depth

The theory-of-mind depth parameter k also determines the maximum nesting depth of belief queries that the system can answer. Queries within this bound are evaluated normally, whereas queries that exceed the configured depth return `undefined`. This bounded setup provides a simple operational model of limited recursive reasoning while still allowing systematic comparisons across levels of ToM depth.

5.5 Observability-Gated Belief Updates

The key mechanism that produces false beliefs in the system is observability. Agents update their beliefs only on the basis of events they observe. If an agent does not observe an event, its epistemic state remains unchanged and may therefore diverge from the true world state.

Each environment event specifies both a world update and a set of observing agents. When an event occurs, agents that observe the event update their belief state to reflect the new world configuration. Agents who do not observe the event retain their previous beliefs, which can lead to outdated or false beliefs about the environment.

This observability-driven update mechanism is sufficient to reproduce classical false-belief patterns. For example, if an agent leaves the environment and an object is moved during its absence, the absent agent will continue to believe that the object remains in its original location. Agents who remained present during the event will correctly update their beliefs.

Although this update mechanism is conceptually inspired by Dynamic Epistemic Logic, the implementation uses a simplified operational procedure rather than full constructions of the event-model [3, 4, 6]. This design keeps the system computationally tractable while preserving the core epistemic behavior required for Theory-of-Mind reasoning.

5.6 Implementation-Level View: Update and Query Procedures

From an implementation standpoint, the system maintains an epistemic state object that is updated event-by-event. The following two algorithms summarize the high-level procedure for (i) applying an event update under observability constraints and (ii) evaluating bounded-depth belief queries.

Procedure 1: Event-based belief update**Input:** current epistemic state, event e , observing agents O_e **Output:** updated epistemic state

1. Apply the event to update the true world state.
 2. For each agent:
 - (a) If the agent observed the event, update its belief state.
 - (b) Otherwise, keep the previous belief state unchanged.
 3. Return the updated epistemic state.
-

Procedure 2: Belief query evaluation**Input:** epistemic state, query φ , theory-of-mind depth k **Output:** $\{\text{true}, \text{false}, \text{undefined}\}$

1. If the nesting depth of φ exceeds k , return **undefined**.
 2. Otherwise, evaluate the query in the epistemic model.
 3. Return the result of the evaluation.
-

6 Experimental Setup

This section specifies the evaluation procedure used to assess recursive Theory-of-Mind reasoning. The evaluation is theory-driven rather than performance-driven: success is determined by whether belief queries match the epistemic state induced by the action history and observability constraints, rather than by aggregate numerical metrics. This is appropriate because the project’s central objectives are epistemic correctness and interpretability.

6.1 Hypotheses

We formulate hypotheses corresponding to progressively stronger ToM capabilities:

- **H1 (First-order false belief):** If an agent does not observe a world-changing event, it will maintain a false first-order belief about the world state.
- **H2 (Second-order belief attribution):** An agent with theory-of-mind depth $k \geq 2$ can correctly attribute the false first-order belief of another agent after asymmetric observation histories.
- **H3 (Higher-order recursive reasoning):** As theory-of-mind depth increases, the system can correctly answer nested belief queries up to the configured depth.
- **H4 (Observability-consistent revision):** Belief revision occurs only when relevant events become observable to the agent in question.
- **H5 (Bounded reasoning):** Queries whose nesting depth exceeds the configured theory-of-mind depth return **undefined** rather than an incorrect truth value.

6.2 Scenario Design

The evaluation consists of deterministic scripted scenarios that manipulate observability, agent presence, and theory-of-mind depth. Each scenario targets a specific Theory-of-Mind capability, including false-belief persistence, higher-order belief attribution, belief revision after re-entry, and bounded recursive reasoning.

Scenario	Target	Key mechanism
S1	1st-order FB	absence during move
S2	2nd-order FB	asymmetric observation history
S3	3rd-order chain	staged absences across 3 agents
S4a	First-order belief repair	re-entry enables first-order belief correction by observation
S4b	Higher-order belief repair limitation	re-entry over-updates second-order beliefs due to global ent
S5	Mixed depth	heterogeneous k across agents

Table 2: Overview of empirical scenarios and the ToM capability targeted by each.

6.3 Evaluation Protocol

The evaluation objective is epistemic correctness: whether the agent’s answers to belief queries match the expected epistemic state induced by the action history and observability conditions.

6.4 Correctness Criteria

For each scenario, we define a set of queries $\mathcal{Q} = \{\varphi_1, \dots, \varphi_m\}$. For an agent with ToM depth k , each query is evaluated with bounded-depth semantics:

$$\text{Ans}_k(\varphi) \in \{\text{true}, \text{false}, \text{undefined}\}.$$

A scenario is considered successful if:

- For all queries with $\text{depth}(\varphi) \leq k$, $\text{Ans}_k(\varphi)$ corresponds to the expected truth value in the epistemic state of the scenario.
- For all queries with $\text{depth}(\varphi) > k$, the answer is **undefined**.

6.5 Repeatability

Because each scenario is specified as a deterministic script, executing the same scenario with the same depth of theory-of-mind produces identical outcomes. This makes the experimental results repeatable and allows belief updates to be inspected consistently across runs. This determinism directly supports the interpretive goals of epistemic modeling.

6.6 Multiple Tests Across Capabilities

The set of scenarios is deliberately constructed to cover a range of cognitive capabilities. Scenario S1 focuses on first-order false beliefs, while S2 and S3 target progressively deeper levels of recursive belief attribution. Scenario S4 examines how changes in observability affect belief revision, and S5 introduces agents with heterogeneous theory-of-mind depths as a simple model of limited rationality.

7 Results

7.1 Overview of Scenario Outcomes

All scenarios were executed as deterministic event scripts in the simulator. For each scenario, belief queries were evaluated after the action sequence and compared with the expected epistemic state.

Table 3 summarises the outcome of each hypothesis against the experimental results.

Table 3: Hypothesis outcomes after scenario evaluation.

Hypothesis	Outcome	Supporting scenario
H1 — First-order false belief	Confirmed	S1
H2 — Second-order belief attribution	Confirmed	S2
H3 — Higher-order recursive reasoning	Confirmed	S3
H4 — Observability-consistent revision	Partially confirmed	S4
H5 — Bounded reasoning	Confirmed	S3, S5

The pattern of results directly reflects the model’s core design decisions. Scenarios S1–S3 succeed because the observability-gated update mechanism correctly partitions belief states according to agent presence: agents that observe an event, update their epistemic state, while agents that do not retain their previous beliefs unchanged. The partial success of S4 is directly attributable to the global treatment of re-entry: the implementation models entry as a publicly known event rather than an agent-dependent epistemic action, which causes second-order belief repair to propagate more broadly than intended. S5 succeeds because the bounded recursion mechanism returns *undefined* for queries exceeding the configured depth, rather than producing an incorrect truth value — preserving epistemic soundness at the cost of expressive completeness.

Scenario	Capability tested	Expected outcome	Observed outcome	Match
S1	First-order false belief	Absent agent retains outdated belief	False belief preserved as expected	Yes
S2	Second-order belief attribution	Observing agent attributes false belief to absent agent	Nested belief query evaluated correctly	Yes
S3	Third-order recursive reasoning	Depth-3 nested belief query succeeds when $k \geq 3$	Correct at depth 3; undefined below bound	Yes
S4a	First-order belief repair	Re-entry enables first-order revision by observation	First-order belief correctly revised upon re-entry	Yes
S4b	Higher-order belief repair limitation	Re-entry over-updates second-order beliefs	Higher-order repair propagates too broadly	Partial
S5	Bounded mixed-depth reasoning	Agents answer within bound and return undefined beyond bound	Correct within bound; undefined where expected	Yes

Table 4: Summary of expected and observed outcomes across the evaluation scenarios.

7.2 First-Order False Belief

In Scenario S1, agent A leaves the environment before agent B moves the object. Because A does not observe the movement event, its epistemic state remains unchanged. As expected, the system returns the result that A believes the object is still in the original position, while B correctly believes it is in the new location.

This behavior reproduces the classical Sally-Anne false-belief pattern. The result confirms that the observability-gated belief update mechanism correctly maintains outdated beliefs for agents who do not observe world changes.

7.3 Second-Order Belief Attribution

Scenario S2 evaluates whether agents can reason about the beliefs of other agents. In this scenario, agent B leaves the environment before an object is moved by agent A . When B returns later, it retains the outdated belief about the object's location.

Agent A , however, observed the move event and can therefore correctly infer that B holds an incorrect belief. When evaluating the nested query $B_A(B_B(\text{In}(o, \ell_1)))$, the system correctly returns that A believes that B believes that the object remains in the original location. This confirms that the model supports second-order belief attribution under asymmetric observation histories.

7.4 Higher-Order Recursive Beliefs

Scenario S3 deepens the reasoning by introducing three agents and a sequence of staged absences. In this configuration, agents must reason about what one agent believes about another agent's belief. The queries of depth three were evaluated, such as $B_A(B_B(B_C(\text{In}(o, \ell_1))))$.

The system correctly evaluates these nested queries when the theory-of-mind depth parameter is set to $k \geq 3$. When the depth parameter is smaller than the required nesting depth, the query evaluator returns an undefined result. This behavior confirms that recursive reasoning is explicitly bounded by the configured theory-of-mind depth.

7.5 Belief Repair After Re-entry: S4a and S4b

Scenario S4 is divided into two sub-cases to isolate the behaviour of first-order and higher-order belief repair separately.

In S4a, an agent that re-enters the environment after a world-changing event correctly revises its first-order belief to match the true world state. This confirms that the observability-gated update mechanism extends appropriately to re-entry: the returning agent observes the current configuration and eliminates worlds inconsistent with its new observation.

S4b, however, reveals a limitation of the current implementation. When an agent re-enters, other agents may update their beliefs about the returning agent's epistemic state as if the observation process were fully public. This causes higher-order belief repair to propagate more broadly than intended: agents that did not directly observe the re-entering agent's perceptual process may nonetheless revise their second-order beliefs, producing updates that are stronger than the observability constraints would warrant. This mismatch arises because entry is modelled as a globally known event rather than as an agent-dependent epistemic action. A more faithful treatment would represent re-entry as a private or partially observable event, consistent with the product-update semantics of standard DEL [3].

7.6 Relation to Background Literature

The results are consistent with the background literature on symbolic epistemic models of Theory of Mind. In particular, the first-order false-belief behavior observed in S1 corresponds

to classical false-belief tasks such as Sally–Anne, in which an agent’s outdated belief arises from asymmetric information access [2]. The successful evaluation of second- and third-order queries in S2 and S3 is also consistent with the DEL tradition of explicitly modeling nested beliefs using epistemic states and accessibility relations [3, 4, 6]. More broadly, the findings support the methodological claim that simplified symbolic environments can isolate epistemic reasoning more clearly than more behaviorally complex settings.

8 Discussion

8.1 Implications for Theory-of-Mind Reasoning

The scenario-based evaluation illustrates the central importance of observability in the emergence of false beliefs and the role of theory-of-mind depth in enabling higher-order belief attribution. As the depth parameter increases, agents gain the ability to represent and query more deeply nested belief structures that remain consistent with asymmetric observation histories. This confirms that recursive belief reasoning can be studied effectively in a simplified epistemic environment when the representational structure is made explicit.

8.2 Implications for Human-Agent Collaboration

These findings are relevant for human-agent and human-robot interaction. In collaborative settings, agents often need to reason about what others have observed, especially when information is incomplete, delayed, or distributed across participants. A system that can maintain false beliefs, attribute them to others, and revise them transparently is therefore useful not only as a theoretical model of Theory of Mind, but also as a component of socially aware artificial agents.

A simple example is joint search or object retrieval. If one agent moves an object while another agent is temporarily absent, a third agent must reason about which participants have observed the change and which still hold outdated beliefs. The ability to represent and query such asymmetric belief states is therefore directly relevant for coordination in shared environments. A further application domain is assistive robotics in care environments. A robot supporting an elderly person must reason not only about the current location of objects, but also about what the person believes regarding those locations. If the person left the room while a caregiver moved an object, the robot should be able to represent the person’s outdated belief and proactively provide corrective guidance, rather than assuming shared knowledge of the current world state. The bounded recursive belief framework developed in this project provides exactly this capability: by evaluating $B_{\text{robot}}(B_{\text{person}}(\text{In}(o, \ell)))$, the robot can assess whether a human collaborator holds a false belief and act to resolve the discrepancy before it leads to confusion or error.

8.3 Interpretability and Methodological Value

One of the main strengths of the proposed framework is its interpretability. Because beliefs are explicitly represented and accessible through direct queries, incorrect or unexpected outcomes can be examined at the epistemic-structure level. This differs from behavior-based or reward-driven evaluations, where the underlying belief state is often difficult to infer. Such transparency is especially valuable for advanced Theory-of-Mind reasoning, where behavior alone may not uniquely determine what an agent believes.

Abstracting away from physical simulation also serves a methodological purpose. Although physical realism matters in embodied systems, it brings extra uncertainty that can obscure epistemic reasoning. Working in an abstract epistemic environment makes it possible to connect action histories, observability, and resulting belief states more directly.

8.4 Comparison to Existing Literature

The results obtained in the experiments are consistent with previous work on symbolic models of Theory of Mind based on epistemic logic. Earlier research has shown that DEL-style models can reproduce classical false-belief tasks and higher-order belief reasoning by explicitly representing agents' knowledge states and updating them through epistemic events [5, 4, 6]. The behavior observed in the present implementation follows the same general pattern: agents maintain outdated beliefs when they fail to observe an event and revise them only when relevant new observations become available.

At the same time, the current implementation also makes its simplifications visible. In particular, the limitation observed in higher-order belief repair after re-entry shows that executable simplifications of DEL-style reasoning can preserve core false-belief behavior while still diverging from a more fully agent-dependent epistemic update treatment. This makes the framework useful both as a working model and as a methodological bridge between formal epistemic logic and implementable artificial agents.

9 Limitations and Future Work

A central limitation of this work is the absence of a large-scale empirical assessment. The framework is evaluated through controlled scenario-based experiments, which provide strong interpretability but do not capture performance under more realistic, noisy, or computationally demanding conditions. Future work could complement the present approach with quantitative analyses of robustness amid imperfect perception and with systematic measurements of computational cost as the number of agents, events, and levels of recursive belief increases. Such extensions would help clarify the trade-off between interpretability and scalability within epistemic Theory-of-Mind models.

The limitations identified in this project suggest several concrete directions for future work, each of which connects to open problems in the existing literature. The higher-order belief repair limitation observed in S4b directly motivates the adoption of private announcement operators, as formalised by [3], which would allow re-entry to be modelled as an agent-dependent epistemic event rather than a globally visible update. Similarly, the scalability constraints of the possible-world representation connect to ongoing work on compact epistemic representations [4], where factored or symbolic encodings have been proposed to reduce the cost of maintaining large sets of possible worlds. Dealing with these limitations would move the framework closer to deployment in embodied systems, where both epistemic accuracy and algorithmic effectiveness are required.

Promising directions for future work include:

- integrating intention and goal reasoning (beyond beliefs) to model richer social cognition, building on epistemic planning frameworks [4],
- adopting incomplete observability models more realistic than co-presence (e.g., occlusion, limited field of view), as explored in robot implementations of DEL [6],
- incorporating perception modules to connect symbolic epistemic states to sensory inputs,
- scaling scenario complexity to study the trade-off between epistemic expressivity and computational manageability.

10 Conclusion

This project implemented recursive Theory-of-Mind reasoning in an abstract epistemic environment grounded in the core representational ideas of Dynamic Epistemic Logic. By modeling

discrete events and belief updates under observability constraints, the framework enables the construction of false-belief scenarios in a controlled, interpretable, and reproducible way.

Across the evaluation scenarios, the implementation successfully reproduced first-, second-, and third-order false-belief reasoning, supported explicit nested belief queries, and demonstrated bounded recursive reasoning through the theory-of-mind depth parameter. At the same time, the experiments revealed an important limitation: higher-order belief repair after re-entry is not yet handled with full agent-dependent precision.

Overall, the project shows that a simplified executable DEL-style framework can capture the essential epistemic patterns required for recursive Theory of Mind while remaining transparent enough for detailed analysis. The framework therefore provides a useful basis for future work on more realistic observability models, richer social reasoning, and human-agent interaction. More broadly, the project demonstrates how symbolic epistemic modeling can provide interpretable foundations for recursive Theory-of-Mind reasoning in artificial agents.

References

- [1] Fagin, R., Halpern, J. Y., Moses, Y., and Vardi, M. Y. *Reasoning About Knowledge*. MIT Press, 1995.
- [2] Wimmer, H., and Perner, J. Beliefs about beliefs: Representation and constraining function of wrong beliefs in young children’s understanding of deception. *Cognition*, 13(1):103–128, 1983.
- [3] Baltag, A., Moss, L. S., and Solecki, S. The logic of public announcements, common knowledge, and private suspicions. In *Proceedings of TARK*, 1998.
- [4] Bolander, T., and Andersen, M. Epistemic planning for single- and multi-agent systems. *Journal of Applied Non-Classical Logics*, 21(1):9–34, 2011.
- [5] Scassellati, B. Theory of mind for a humanoid robot. *Autonomous Robots*, 12(1):13–24, 2002.
- [6] Dissing, L., and Bolander, T. Implementing theory of mind on a robot using dynamic epistemic logic. In *Proceedings of IJCAI*, pages 1869–1875, 2020.