Reasoning about Probabilistic and Higher-order Beliefs

Von der Fakultät für Informatik der RWTH Aachen University
zur Erlangung des akademischen Grades
eines Doktors der Naturwissenschaften
genehmigte Dissertation

vorgelegt von

Qihui Feng, Master of Science

Berichter: Prof. Gerhard Lakemeyer, Ph.D.
Dr. Vaishak Belle

Tag der mündlichen Prüfung: 01. Juni 2026

Diese Dissertation ist auf den Internetseiten der Universitätsbibliothek verfügbar.

Acknowledgement

I would first like to express my sincere gratitude to my supervisor, Gerhard Lakemeyer, for his continuous support and invaluable guidance throughout my doctoral studies and in the preparation of this dissertation. I am also deeply grateful to Vaishak Belle for kindly serving as the second reviewer of this dissertation.

I am also indebted to my colleagues in the Knowledge-Based Systems Group: Mohamed Behery, Hayyan Helal, Till Hofmann, Daxin Liu, Matthew James Lynn, and Tarik Viehmann. Among them, I would especially like to thank Daxin, my Master's thesis advisor and former office mate, for his help during my first year of doctoral studies. I am likewise grateful to Matthew, my current office mate, with whom I have shared teaching duties and many discussions. Our exchanges of ideas have been a steady source of inspiration throughout my PhD.

I would also like to acknowledge those who have provided important academic input at various stages of my research. In particular, I thank Jürgen Giesl and Yongmei Liu for their thoughtful comments and invaluable suggestions. I am also grateful to my collaborators, Er Jin, Shakil M. Khan, Yongli Mou, and Hannah Louise Wilk, for the rewarding and productive collaborations we have shared.

I would like to express my heartfelt gratitude to my family and friends for their constant support and understanding throughout this journey. I am particularly grateful to my friend Ba Que Le for carefully reading the manuscript and offering language suggestions. Their encouragement, help, and trust have meant more to me than I can fully express.

Over these years, much of my research has been devoted to belief and cognition—to understanding how human mental states can be represented and reasoned about through logic. Yet I have come to realize, with regret, that the very first person I should have understood was the one who had once been closest to me. This dissertation is dedicated to Xuejun Zhong, my ex-wife. I wish her a future as bright and warm as ever.

Abstract

In Knowledge Representation and Reasoning, the notion of belief is used to model an agent's mental state about the world and other agents. Probabilistic beliefs extend this notion by capturing degrees of belief under uncertainty, while higher-order beliefs represent beliefs about the beliefs of others. These concepts play an important role in reasoning under uncertainty, multi-agent interaction, and decision-making. Developing logical frameworks that can faithfully represent such beliefs and support effective reasoning remains a fundamental challenge in the field.

The first contribution of this thesis is a measure-based semantics for probabilistic beliefs. By defining probability measures over spaces of possible worlds, the proposed logics support reasoning about arbitrary belief distributions, including continuous and mixed distributions, and allow reasoning about beliefs after actions via integration over such spaces. This provides a unification of logic and probability that goes beyond discrete probabilistic models commonly studied in the literature.

The second contribution concerns reasoning about higher-order beliefs in multi-agent settings, in both static and dynamic cases. For static reasoning, a translation-based inference method is developed for inference in modal logics with beliefs. For dynamic scenarios, the thesis proposes a syntactic and decidable approach to epistemic planning in the presence of higher-order beliefs.

The final contribution integrates probabilistic and higher-order beliefs into a unified logical account. By extending the notion of only-knowing to probabilistic and multi-agent settings, the resulting framework supports precise reasoning about agents' knowledge bases as well as their beliefs about others' beliefs.

Zusammenfassung

Im Bereich der Wissensrepräsentation wird der Begriff Glaube verwendet, um den mentalen Zustand eines Agenten in Bezug auf die Welt und andere Agenten zu modellieren. Probabilistischer Glaube erweitet dieses Konzept, indem er den Grad des Glaubens unter Unsicherheit erfasst, während höherstufiger Glaube „Glaube über den Glauben anderer Agenten“ erfasst. Diese Konzepte spielen eine zentrale Rolle beim Schließen unter Unsicherheit, in Multiagentensystemen sowie für Entscheidungen in komplexen, dynamischen Umgebungen. Eine grundlegende Herausforderung in diesem Forschungsbereich stellt dabei die Entwicklung logischer Systeme dar, die solche Strukturen angemessen abbilden und schlüssige Argumentationen unterstützen.

Der erste Beitrag dieser Dissertation ist eine maßbasierte Semantik für probabilistischen Glauben. Durch die Definition von Wahrscheinlichkeitsmaßen über Räumen möglicher Welten ermöglicht der vorgeschlagene Ansatz das Schließen über beliebige Verteilungen von Glauben, einschließlich kontinuierlicher und Mischverteilungen. Weiterhin ermöglicht der Ansatz durch Integration über die Räume eine Modellierung von Änderungen des Glaubens im Anschluss an Handlungen. Dies führt zu einer Verknüpfung von Logik und Wahrscheinlichkeit, die über die häufig betrachteten diskreten probabilistischen Modelle hinausgeht.

Der zweite Beitrag, den die vorliegende Arbeit leistet, befasst sich mit dem Schlussfolgern über höherstufigen Glauben in Multiagentensystemen, sowohl in statischen als auch in dynamischen Kontexten. Für statische Szenarien stellt diese Arbeit einen übersetzungsbasierten Ansatz für das Schlussfolgern in modalen Logiken mit Glauben vor. Für dynamische Szenarien wird ein syntaktischer und entscheidbarer Ansatz für die epistemische Planung in Systemen mit höherstufigem Glauben dargestellt.

Ein abschließender Beitrag integriert probabilistischen und höherstufigen Glauben in einem einheitlichen logischen Rahmen. Durch eine Erweiterung des Begriffs des

only-knowing auf probabilistische und multiagentenbezogene Kontexte ermöglicht das vorgeschlagene System präzise Schlussfolgerungen über die Wissensbasen von Agenten sowie deren Glauben über die Glauben anderer Agenten.

Contents

1 Introduction **1**

2 Relevant Literature **5**

2.1 Knowledge and Beliefs 5

2.1.1 Modal Logic for Knowledge 6

2.1.2 Logic for Probabilistic Beliefs 8

2.2 Actions and Changes 9

2.3 Discussion 11

3 Formal Foundations **13**

3.1 The Logic $\mathcal{OL}$ 13

3.1.1 Syntax 14

3.1.2 Semantics 15

3.1.3 Properties of $\mathcal{OL}$ 17

3.2 Measure and Integration Theory 18

3.2.1 Lebesgue Measure and Integral 19

3.2.2 Abstract Measure and Integration Theory 22

4 On the Unification of Logic and Probability **25**

4.1 Probability Spaces as Semantic Models 25

4.1.1 Syntax 26

4.1.2 Semantics 27

4.1.3 Properties of $\mathcal{OBL}_c$ 31

4.1.4 Reasoning about Distributions 39

4.2 Reasoning about Actions by Belief Integration 47
4.2.1 Syntax 47
4.2.2 Semantics of Static Beliefs 49
4.2.3 Semantics after Actions 51
4.2.4 Properties 64
4.2.5 Reasoning about Actions and Beliefs 70
4.3 Discussion: Relation to Levesque's Only-Knowing 76

5 Reasoning about Higher-order Beliefs **81**
5.1 Translation-based Modal Theorem Proving 82
5.1.1 The Logic $\mathcal{ML}_n$ 84
5.1.2 Comparison with Kripke Semantics 88
5.1.3 Adaptation for Other Modal Systems 92
5.1.4 Translation 94
5.1.5 First-order Translation 102
5.2 Planning with Higher-order Beliefs 108
5.2.1 The Logic $\mathcal{MS}$ 110
5.2.2 Multi-Agent Action Theories 115
5.2.3 Action Types 119
5.2.4 Regression 121
5.2.5 Examples 124
5.2.6 Multi-agent Epistemic Planning 127

6 Joining Higher-order and Probability **133**
6.1 A Logic for Multi-agent Only-believing 135
6.1.1 Syntax 135
6.1.2 Semantics 138
6.2 Properties of the Logic 144
6.2.1 Knowledge 144
6.2.2 Degree of Belief 147

6.3 Only-believing with Higher-order Beliefs 148
6.4 Discussion 152

7 Conclusion **155**

Bibliography **157**

A Appendix **173**
A.1 Long Proofs 173
A.1.1 Proofs about $\mathcal{KL}_n$ translation 173
A.1.2 Proofs about Regression 184

1 Introduction

As subjects situated within the lifeworld, we are inevitably confronted with two types of uncertainty. The first type originates from the objective physical environment, manifested in questions such as "Will it rain tomorrow?" or "Where will the robot eventually stop?" This form of uncertainty stems from partial observability, inherent stochasticity, and the seemingly chaotic nature of the world as a complex system. The second type arises from the coexistence of other subjects (or agents, as we follow the Computer Science naming convention). It concerns the uncertainty regarding the mental states of others, manifested in queries such as "Does he know I am here?", "What would she do when she is aware of his lies?"

While the existence of the first type of uncertainty is largely undisputed, philosophical skepticism has primarily targeted the second. Behaviorists might argue that "mental states" do not exist per se; since other agents are intrinsically part of the physical world, their behaviors are, in principle, predictable via physical laws or neurobiology. In response, Daniel Dennett offers a compelling counter-argument through the concept of the "*Intentional Stance*". He posits that even if a physical reduction is theoretically conceivable (the "*Physical Stance*"), predicting the behavior of complex systems like humans via atomic-level physics is practically impossible due to its immense complexity [Den89]. Consequently, treating others as rational agents possessing beliefs, desires, and intentions becomes not just a convenient heuristic, but the only pragmatically feasible strategy for explaining

and predicting social interactions. Furthermore, denying others' agency leads to a contradiction: by reducing them to mere physical mechanisms, we implicitly deny our own agency. If our reasoning is nothing more than the deterministic firing of neurons rather than a pursuit of truth based on reasons, then the very act of argumentation we are currently engaging in loses its epistemic validity.

To address these forms of uncertainty within a computational framework, Artificial Intelligence has predominantly adopted a data-driven paradigm, encompassing statistical inference and Deep Learning. While these sub-symbolic approaches have demonstrated remarkable efficacy in modelling uncertainty, they suffer from fundamental epistemological limitations. Inherently, these methods perform inference based on probabilistic distributions over atomic features, lacking an explicit logical structure to sustain systemic coherence or valid argumentation. Namely, they stand in contrast to the human cognitive process, often characterized by the *Hypothetico-Deductive Model*, as advocated by Karl Popper. Human cognition involves abstracting experiences into theories or conjectures through descriptive language, performing deductive reasoning based on these symbolic structures, and continuously updating or discarding theories in light of empirical reality. The absence of this cognitive loop in statistical learning is the root cause of its lack of explainability. Even though modern Large Language Models (LLMs) have shown the potential to mimic human reasoning via techniques such as Chain-of-Thought (CoT), their underlying probability-based next-token prediction mechanism does not support complex deductive reasoning and often leads to hallucinations.

Contrastively, logic-based AI, particularly the knowledge representation and reasoning (KRR), provides an alternative paradigm, where the two seemingly distinct forms of uncertainty are elegantly unified under the epistemic modalities of *Belief*: The stochasticity originating from the objective physical world (the first type) is characterized as the *probabilistic beliefs* (or degrees of belief), representing the agent's degree of confidence in factual propositions or statements. Meanwhile, the uncertainty arising from social interaction (the second type) is characterized

via *higher-order beliefs*—namely, beliefs about others' beliefs and non-beliefs. In this way, logic not only imposes structure upon uncertainty but also achieves a unified reasoning mechanism for both the natural and social domains within a family of formal systems.

Building on the foundational work of Hintikka and Kripke [Hin62, Kri63], extensive research on the notion of belief has been conducted over the past half-century. With the recent rise of Neurosymbolic AI, Knowledge Representation and Reasoning methods are gradually regaining the attention of the academic community. However, current KRR frameworks continue to face significant limitations, particularly in terms of expressiveness, availability of effective inference tools, and applicability to practical problems.

This thesis does not aim to resolve all conflicts in a single stroke. Instead, it seeks to make substantial breakthroughs in some specific key dimensions. Specifically, the thesis makes contributions in three aspects:

- Enhancing expressiveness with measure-based semantics:

  To overcome the limitations of existing logics in handling continuous and infinite domains, we propose a novel semantic framework based on probability spaces over possible worlds. We introduce the logic $\mathcal{OBL}_c$ and its dynamic extension $\mathcal{PS}$. By combining measure theory and integration theory with possible-world semantics, these formalisms enable the representation of arbitrary belief distributions—including continuous and mixed distributions—and reasoning about the belief after actions.

- Developing inference approaches for higher-order beliefs:

  To address the computational complexity of multi-agent reasoning, we develop inference methods for both static and dynamic settings. For static reasoning, we propose a translation-based approach that embeds multi-agent modal logics (e.g., $K45_n$, $S5_n$) into decidable fragments of first-order logic. For dynamic reasoning, we introduce the psychological concept of *Theory*

of Mind (ToM) and represent the *Multi-agent Epistemic Planning* (MEP) problem in an epistemic variant of the *Situation Calculus*. We prove that by bounding the order of ToM, the MEP problem becomes decidable without sacrificing essential expressiveness.

- Unifying probabilistic and higher-order beliefs:

  To unify probabilistic and higher-order beliefs, we propose the logic $\mathcal{OBL}_m$. This logic extends Levesque's notion of 'Only-knowing' to multi-agent probabilistic settings, allowing us to precisely capture agents' degrees of belief and their (probabilistic) higher-order beliefs about others.

The thesis is organized as follows: Chapter 2 examines related work. Chapter 3 reviews the formal foundations, particularly Levesque's logic $\mathcal{OL}$. Chapter 4 details the semantic extensions for probability and continuous domains ($\mathcal{OBL}_c$ and $\mathcal{PS}$). Chapter 5 investigates decidable reasoning methods for multi-agent beliefs in static and dynamic scenarios. Chapter 6 demonstrates the integration of higher-order beliefs and probabilistic reasoning within a unified semantic framework ($\mathcal{OBL}_m$). Finally, Chapter 7 concludes the thesis and outlines future work.

2 Relevant Literature

In this chapter, we review the literature across the relevant fields of this thesis.

We begin in Section 2.1 with the formalization and semantic models of knowledge and beliefs, including both the deterministic and probabilistic variants. Section 2.2 introduces accounts for reasoning about actions and changes, along with the corresponding reasoning approaches such as progression and regression.

We end this chapter with a discussion. In particular, we illustrate the relationships among the various logical accounts discussed and proposed in this thesis via a family tree.

2.1. Knowledge and Beliefs

The study of knowledge and beliefs as components of formal languages has a long history in philosophy, logic, cognitive science, and artificial intelligence, and it provides the symbolic basis for this thesis.

Over the decades, a variety of semantic models have been proposed—most notably Kripke's semantics by Hintikka and Kripke [Hin62, Kri63] and the "set-based" semantics proposed by Levesque [Lev82, Lev90] and later extended by Lakemeyer, Belle, and others [BL10, LL11, BL15a, LF23]. Although these formalisms can essentially be viewed as equivalent (as we will show in the thesis), they differ not only in how they model knowledge and beliefs but also in the

suitability for expressing richer epistemic concepts such as common knowledge, probabilistic beliefs, and only-knowing. In what follows, we review the core ideas behind these logical accounts as preparation for the more expressive languages and the reasoning methods developed later in the thesis.

2.1.1. Modal Logic for Knowledge

We begin by reviewing Kripke's semantics, also known as *relational* semantics. The central idea is to reason not about a single state of the world, but about a collection of *possible worlds*, each representing a way the world might be. A modal statement such as "the agent knows φ" is understood as saying that φ holds in all worlds the agent considers possible. This viewpoint, now known as possible-world semantics, has become the dominant framework in epistemic logic.

A key component of the semantics is the *accessibility relation*. This specifies, for each world, which alternative worlds are epistemically available to the agent. The structure of this relation determines the properties of the modal operators. For example, if the relation is *reflexive*-every state is accessible from itself, then knowledge satisfies the truthfulness property. Namely, what is known (believed) must be true. If the relation is transitive, the agent is *positively introspective*, i.e. it knows what it knows. Likewise, if it satisfies an additional "Euclidean" property, then the agent is negatively introspective, meaning it is aware of what it does not know. Different assumptions of these conditions correspond to different modal axiom systems, including K, T, $S4$, $S5$ and $K(D)45$[Che80].

Much of the early literature on epistemic logic focused on establishing these correspondences and understanding how the choice of accessibility conditions affects the expressive power and behaviour of the logic [Hin62, Kri63, HC68]. At the same time, a parallel line of research investigated the complexity of reasoning in these systems. Foundational results by Ladner [Lad77] and Halpern [HM92] prove the decidability and complexity for all normal modal logics. In particular, for the single-agent versions of knowledge ($S5$) and beliefs ($K45$ or $KD45$), the

SAT problem is NP-complete. This relies on the *small-model property*: In a nutshell, any satisfiable formula can be satisfied by a Kripke structure with very few states. For logics with many agents ($S5_n, KD45_n$), the SAT problem becomes PSPACE-complete.

The lower bound follows from a reduction of the PSPACE-complete QBF problem,[1] while the upper-bound is proved based on the *tableau* method [Kri63].

In 1982, Levesque proposed a simplified semantics for modal logic of knowledge and beliefs with only one agent [Lev82], where each model consists of "one distinguished state describes what is true in the 'real' world, and a set of states corresponds to the worlds that the agent thinks possible" [HM85]. The account from Levesque turned out to be suitable for characterising the knowledge base of an agent and was extended to the logical accounts $\mathcal{KL}$ and $\mathcal{OL}$ [Lev90]. The logic $\mathcal{OL}$ introduced an additional modality "only-knowing", which precisely captures the beliefs and non-beliefs of an agent. Intuitively, a formula $\boldsymbol{O}\varphi$ means "All the agent knows is φ", and any objective statement which cannot be entailed by φ is not the belief of the agent. This account has proved its suitability in expressing non-monotonic reasoning and was extended to $\mathcal{O}_3\mathcal{L}$ [LL05a, LL06], an account embedding default logic [Rei80] and different variants of autoepistemic logic [Moo85, Kon88] into a unified formalism.

The multi-agent extension of $\mathcal{OL}$ was first investigated by Halpern and Lakemeyer as an extended canonical Kripke model [HL01]. A more natural extension $\mathcal{OL}_n$ was done by Vaishak Belle and Gerhard Lakemeyer [BL10, BL15b].

Considering the property of models in the afore-mentioned accounts, we term them the family of "*set-based*" semantics. This thesis will show that set-based models provide an alternative to Kripke semantics. They also serve as an important tool for bridging the modal logics of knowledge and belief with first-order logic.

[1] The problem of deciding whether a QBF is true is PSPACE-complete [SM73].

2.1.2. Logic for Probabilistic Beliefs

In many logical accounts, an agent's subjective degree of belief is represented or modelled using probability. Such consideration allows the representation of formulas such as $P_{\geq r}\varphi$, read as "the probability that formula φ holds is at least r", or $\boldsymbol{B}_i(\varphi\colon 0.3)$, read as "the agent i believes φ with degree 0.3".

The following is a non-exhaustive list of logical accounts of probability and probabilistic beliefs. Bacchus developed a logical framework for formally representing and reasoning about uncertain information and integrating first-order language with probabilistic knowledge [Bac89]. In 1990, Fagin, Halpern, and Megiddo proposed a logic for reasoning about probability over propositional formulas [FHM90]. Around the same time, Halpern extended this work and provided semantics for first-order logic of probability [Hal90]. A key insight from Halpern is the distinction between two interpretations of probability: one defines the probability measure over the domain of discourse, while the other defines it over possible worlds. Fagin and Halpern enriched classical Kripke structures with probability spaces, thus combining knowledge, probabilistic beliefs, and probability in a unified framework.

A first-order account of probability and only-knowing $\mathcal{OBL}$ was proposed by Belle, Lakemeyer and Levesque [BLL16], where the epistemic state of the agent is defined as a set of discrete probability distributions[2] on the possible worlds. Their work was later extended to the dynamical account $\mathcal{DS}$ [BL17], which was used to reason about actions [LL21, LF21]. A central limitation of these accounts, including $\mathcal{OBL}$, $\mathcal{DS}$ and their variants, is that they are restricted to discrete probability distributions. As a result, they cannot express or manipulate continuous, mixed, or other more involved probabilistic terms. To accommodate such richer forms of uncertainty, a genuinely measure-theoretic treatment of probabilistic belief is re-

[2] More precisely, each element of the epistemic state is a value assignment that assigns weights to all possible worlds. Although there are uncountably many possible worlds, the semantics remains well-defined via assigning non-zero weight only to a countable subset.

quired. This need motivates the development of more general probabilistic logics capable of reasoning with arbitrary probability distributions.

2.2. Actions and Changes

Research on reasoning about actions and epistemic change has evolved along two paradigms. The first captures the dynamics of actions axiomatically, and proceeds reasoning by manipulating and updating logical formulae. We refer to this as the *syntactic approach*. Instances include the projection reasoning in the situation calculus [LR97] as well as the sequent calculi in the *Dynamic Epistemic Logic* (DEL) [Auc11, Auc12]. The second paradigm encodes both the states and their transformations directly within semantic models. Hence, we call it the *semantic approach*. An example is the DEL product update framework employed in epistemic planning[BA11, Bol17].

In this section, however, our exposition focuses on the situation calculus and its methods for *projection reasoning*. Since DEL is significantly different from the main content of this thesis, we do not dive into details and discuss DEL only in the corresponding sections.

The situation calculus traces back to the foundational work of McCarthy and Hayes [MH81]. Their framework provided a conceptual basis for formalizing changes and action effects.

A major refinement was done by Reiter [Rei91, Rei01]. In his work, Reiter introduced a precise axiomatization of actions using successor-state axioms, initial-state specifications, and action precondition axioms. Reiter's successor-state axioms provide a uniform solution to the *frame problem*, specifying both the effects and non-effects of actions in a single schema and forming the basis of what is now known as the *Basic Action Theory* (BAT) in the situation calculus. This formalization established the situation calculus as one of the dominant logical foundations for reasoning about action and change.

Building on this, Bacchus, Halpern, and Levesque (BHL) introduced a gen-

eral framework for representing probabilistic belief within the situation calculus [BHL99a]. Their account associates each situation with a weight and then takes the degree of belief to be the normalized weight. This framework allows expressing partial information, noisy sensors, and stochastic actions.

However, the original BHL framework is limited to discrete probability distributions. To express and reason about continuous probability, Belle and Levesque proposed a continuous extension [BL18]. The key insight was that while integration over situations is in general not well-defined, one can instead integrate over the values of a finite set of nullary fluents, enabling continuous beliefs over real-valued quantities.

Reasoning about knowledge was incorporated into the situation calculus in the work of Scherl and Levesque [SL03], which extends the basic formalism with epistemic fluents and sensing actions to model knowledge and information gathering.

To address certain semantic complexities of the purely axiomatic framework, Lakemeyer and Levesque proposed the logic $\mathcal{ES}$, a modal variant of the epistemic situation calculus [LL11]. By shifting from purely axiomatic definitions to a possible-world semantics, they provided a more straightforward account, while preserving much of the expressive power of the original situation calculus. This modal variant later served as the basis for further probabilistic extensions, including $\mathcal{DS}$ and its variants [BL17, LL21, LF23].

Projection reasoning is the mechanism in the situation calculus for determining what holds after actions. Two main approaches to projection have been studied: regression and progression. Regression reduces a query about a future situation to an equivalent query about the initial situation. The regression operator was first proposed by Reiter [Rei91, Rei01], and further developed by other researchers [LL04, LL11, BL20, LL21]. A key motivation for regression is that standard axiomatizations of the situation calculus involve second-order quantification, or, in modal variants, temporal modalities referring to action history. It is often difficult to reason directly about such formulae. Regression eliminates references

to situation terms or temporal modalities by rewriting formulas to concern only the initial state, thereby substantially simplifying the reasoning procedures.

However, as observed by Levesque and Lakemeyer and others, regression is not well-suited for long-lived or continuously operating agents. Regression results tend to grow exponentially with the length of the action history, and many practical queries concern the agent's current situation rather than the initial one. Maintaining the initial theory together with a growing action history is therefore inefficient in such settings. This motivates progression, which updates the initial state to a future state based on the effects of actions. The general formalism was introduced by Lin and Reiter [LR97], and further developed in [LL05b, VL$^+$07, VL13, BL14b].

Nevertheless, as proved by Lin and Reiter in [LR97], progression is in general not first-order definable. To preserve the simplicity in proving, it is necessary to investigate fragments of the language, where the progression results remain first-order [LL10, LC24, LC25].

2.3. Discussion

A number of logical accounts are involved in this thesis. Some have already been introduced in this chapter, including $\mathcal{KL}$, $\mathcal{OL}$, $\mathcal{ES}$, $\mathcal{OBL}$, and $\mathcal{DS}$, while others will be presented in later chapters. To avoid confusion and to provide a unifying perspective, we conclude this chapter with an overview of the languages discussed or proposed in the thesis, organized as a language family tree, shown in Fig. 2.1.

The languages are categorized along two dimensions. First, we distinguish between static and dynamic languages based on whether they can express and reason about actions and their effects. Second, we classify languages by their ability to handle probability: Deterministic languages cannot represent degrees of belief, while some languages support discrete probability. Additionally, there are languages grounded in measure-theoretic semantics and can therefore reason about arbitrary probability distributions.

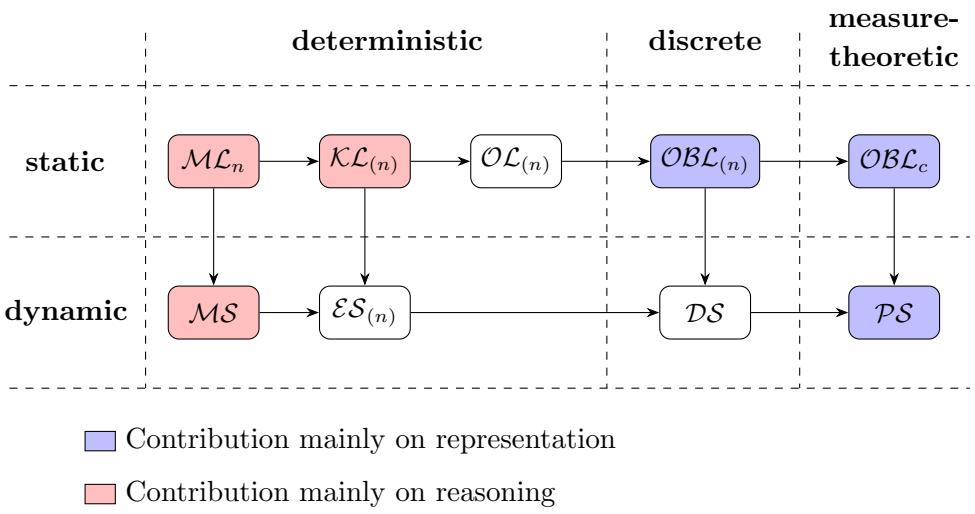


Figure 2.1.: Languages involved in the thesis, and their relation to the existing work. Arrows stand for extension relation between two accounts. Subscript (n) means that the language has a multi-agent extension.

3

Formal Foundations

This chapter introduces the formal, logical, and mathematical foundations that will be assumed throughout the remainder of the thesis.

In Section 3.1, we introduce Levesque's logic $\mathcal{OL}$. Although $\mathcal{OL}$ itself will not be directly employed in the subsequent chapters, it provides a unifying basis: all logical accounts considered later can be viewed either as extensions of $\mathcal{OL}$ or, at a minimum, as inheriting key properties from it. Presenting $\mathcal{OL}$ at this stage, therefore, clarifies formalisms developed in later chapters.

Section 3.2 reviews the necessary background from real analysis. In particular, we introduce the notions of measure theory, probability measures, the Lebesgue integral, and integration over abstract measure spaces. This material is included to prepare for the measure-theoretic semantics introduced in Chapter 4.

3.1. The Logic $\mathcal{OL}$

The logic $\mathcal{OL}$, originally proposed by Levesque, is a first-order logic capturing the knowledge and "*only-knowing*" of an agent [Lev90]. The notion of only-knowing has profound influence on knowledge-based systems [LL01], non-monotonic reasoning [LL05a, LL06] and reasoning about actions [LL04, LL11]. A series of studies investiages variants and extensions of $\mathcal{OL}$ [HL01, BL10, AB15, LL11, BL15a, BL17, BL18, LF21], including those presented in this thesis, such as $\mathcal{OBL}_c, \mathcal{PS}$,

and $\mathcal{OBL}_m$ [FLLB23, LFBL23, FL24].

3.1.1. Syntax

Compared with other systems of (first-order) modal logic, the logic $\mathcal{OL}$ introduces the notion of *standard names* to conveniently idenfy different elements and to enable substitutional quantification over first-order variables. The logic assumes a countably infinite set of names

$$\#1, \#2, \#3, \ldots$$

which are essentially ordinary constants. A bijective mapping is built between this set of names and the fixed domain of discourse. In this way, the logic avoids the use of variable maps and therefore interprets first-order quantification substitutionally.

The symbols of $\mathcal{OL}$ are taken from the following vocabulary:

- a countably infinite set of *variables*;
- a countably infinite set of standard names;
- the equality symbol $=$;
- the logical connectives $\forall, \wedge, \neg$ and parentheses;
- *predicate* and *function* symbols of each arity, in particular, *constants* are interpreted as functions without parameters, i.e. *nullary* functions;
- epistemic modalities $\boldsymbol{K}$ and $\boldsymbol{O}$.

Conventionally, we use the letters x, y and z for variables, and n, m for standard names, possibly with subscripts or superscripts. We use $\mathcal{N}$ to denote the set of all names. Other common logical connectives, such as $\exists$, $\supset$ and $\vee$, are treated as abbreviations.[1] We now describe the expression of $\mathcal{OL}$:

[1] In this thesis, we use both $\supset$ and $\rightarrow$ for implication.

Definition 3.1 ($\mathcal{OL}$ Term)**.** Terms in $\mathcal{OL}$ are the minimal set including the following elements:

- variables,
- standard names,
- $f(t_1, \ldots, t_k)$, where $t_1, \ldots, t_k$ are terms, and f is a k-ary function symbol.

A term that contains no variables is called a *ground term*. A ground term containing only a single function symbol is called *primitive*. Now we define the set of well-formed formulae:

Definition 3.2 ($\mathcal{OL}$ Formula)**.** The $\mathcal{OL}$ formulae are the minimal set containing:

- atom, i.e a formula of the form $P(t_1, \ldots, t_k)$ where P is a k-ary predicate symbol, and $t_1, \ldots, t_k$ are terms;
- $t = t'$, where t, t' are terms;
- $\neg\alpha$, where α is a formula;
- $\alpha \wedge \beta$, where α, β are formulae;
- $\forall x\ \alpha$, $\boldsymbol{K}\alpha$ and $\boldsymbol{O}\alpha$.

As with terms, a *ground atom* is an atom without variables, and a *primitive* atom is a ground atom where every t_i is a standard name. A *sentence* of $\mathcal{OL}$ is a formula with no free variables. A formula without $\boldsymbol{K}$ or $\boldsymbol{O}$ is called *objective*; and a formula is called *subjective* if every occurrence of predicates and functions is in the scope of $\boldsymbol{K}$ or $\boldsymbol{O}$.

3.1.2. Semantics

The semantics of $\mathcal{OL}$ is given in terms of possible worlds. Different from the classical interpretation of first-order sentences using Tarski structures, where one

specifies a *domain of discourse* and appropriates functions and relations for the corresponding symbols, the semantics of $\mathcal{OL}$ is much simpler: A fixed, countably infinite domain is considered, and a bijective correspondence between elements in the domains and the standard names is built. According to the well-known Löwenheim–Skolem Theorem, every satisfiable first-order formula is satisfiable in a countable domain. Thus, having a fixed, countable domain will not reduce the generality. Formally, a *world* w in $\mathcal{OL}$ is an assignment as follows:

- For each primitive term $f(n_1, \ldots, n_k)$, $w[f(n_1, \ldots, n_k)]$ is a (co-referring) standard name;
- For each primitive atom $P(n_1, \ldots, n_k)$, $w[P(n_1, \ldots, n_k)]$ is either 0 or 1, taken to be a specification of the atom's truth value.

We use W to denote the set of all worlds. The value assignment can be extended to every ground term t at world w, which we write as $w(t)$:

- When t is a standard name n, $w(t) := n$;
- When $t := f(t_1, \ldots, t_k)$, then $w(t) := w[f(n_1, \ldots, n_k)]$ where $n_i = w(t_i)$.

Now we are ready to define the models of $\mathcal{OL}$ as well as the satisfaction relation. By a model, we mean a tuple (e, w), where w is a single world specifying what is true in the real world, and e is a set of worlds specifying the world states, which are considered possible by the agent. Given a model (e, w),[2] terms $t, t_1, \ldots t_k$, and formulae α, β, the satisfaction relation is defined as follows:

- $e, w \models P(t_1, \ldots, t_k)$ iff $w[P(n_1, \ldots, n_k)] = 1$ where $n_i = w(t_i)$;
- $e, w \models t_1 = t_2$ iff $w(t_1)$ and $w(t_2)$ are identical;
- $e, w \models \neg\alpha$ iff it is not the case that $e, w \models \alpha$;
- $e, w \models \alpha \wedge \beta$ iff $e, w \models \alpha$ and $e, w \models \beta$;

[2]For simplicity, we often omit the parentheses.

- $e, w \models \forall x.\alpha$ iff $e, w \models \alpha_n^x$ for all $n \in \mathcal{N}$;
- $e, w \models \boldsymbol{K}\alpha$ iff for every w', $w' \in e$ implies $e, w' \models \alpha$;
- $e, w \models \boldsymbol{O}\alpha$ iff for every w', $w' \in e$ iff $e, w' \models \alpha$.

where α_n^x means to substitute every free occurrence of x with n. For simplicity, we write $e \models \alpha$ when α is subjective and $w \models \alpha$ when α is objective. A sentence α is *satisfiable* if and only if some model e, w exists such that $e, w \models \alpha$. Given a sentence α and a (possibly infinite) set of sentences Σ, we say that Σ *entails* α, written as $\Sigma \models \alpha$, if and only if for each model (e, w) such that $e, w \models \sigma$ for all $\sigma \in \Sigma$, it must hold that $e, w \models \alpha$. A sentence α is *valid* if and only if it is entailed by the empty set, i.e. $\{\} \models \alpha$ (also written as $\models \alpha$).

3.1.3. Properties of $\mathcal{OL}$

A detailed Discussion on the properties is in [LL01]. Here we present those that are the most relevant and those that will differ from the accounts we introduce later. For the knowledge modality $\boldsymbol{K}$, it satisfies the $K45$ axiom system [Kri63, Che80] and the *Barcan formula* [HC68]:

Theorem 3.3. For any formulae α and β,

- If $\models \alpha$ then $\models \boldsymbol{K}\alpha$;
- $\models \boldsymbol{K}\alpha \wedge \boldsymbol{K}(\alpha \wedge \beta) \supset \boldsymbol{K}\beta$;
- $\models \boldsymbol{K}\alpha \supset \boldsymbol{K}\boldsymbol{K}\alpha$;
- $\models \neg\boldsymbol{K}\alpha \supset \boldsymbol{K}\neg\boldsymbol{K}\alpha$;
- $\models \forall x\boldsymbol{K}\alpha \supset \boldsymbol{K}\forall x\alpha$.

For the properties of only knowing, we begin with the simplest case, where the argument is an objective sentence. In this case, the $\boldsymbol{O}$-formula is satisfied by a unique model:

Theorem 3.4. For any objective φ, there is a unique model e such that $e \models \boldsymbol{O}\varphi$.

Following the unique model theorem, we have

Corollary 3.5. If φ is objective and σ is subjective, then either
$\models (\boldsymbol{O}\varphi \supset \sigma)$ or $\models (\boldsymbol{O}\varphi \supset \neg\sigma)$.

Theorem 3.6. If φ, ψ are objective, then $\models \boldsymbol{O}\varphi \supset \boldsymbol{K}\psi$ iff $\models \varphi \supset \psi$.

According to Levesque's original design, only-knowing purely subjective sentences amounts to trivial cases. Noted that this property is different from the accounts we will consider in the thesis. A discussion is left in Section 4.3.

Theorem 3.7. For any state e and subjective formula σ, $e \models \boldsymbol{O}\sigma$ iff either $e \models \sigma$ and $e = W$, or $e \models \neg\sigma$ and $e = \varnothing$.

If we say "all that is known about the world is σ", yet σ is subjective and therefore does not describe the world itself, then it follows that "nothing is known about the world"[LL01]. In this case, either e considers all possibilities ($e = W$), or it is empty, in which case the agent's beliefs are inconsistent.

Corollary 3.8. For atomic φ, $\models \neg\boldsymbol{O}\boldsymbol{K}\varphi$.

3.2. Measure and Integration Theory

One of the central motivations behind the development of measure theory is the inherent limitation of the *Riemann integral* [SS09, Rud21]. It was observed that certain seemingly simple functions cannot be assigned an integral within the Riemann framework. A canonical example is the *Dirichlet function*, defined by

$$\delta(x) = \begin{cases} 1 & x \in \mathbb{Q}; \\ 0 & x \in \mathbb{R}\backslash\mathbb{Q}. \end{cases} \tag{3.1}$$

This function fails to be Riemann integrable on the interval $[0, 1]$ since its *upper Darboux integral* equals 1 while its *lower Darboux integral* equals 0.

The introduction of *Lebesgue measure* and the associated notion of the *Lebesgue integral* resolves this problem by shifting the partitions of the domain from intervals to the measures of preimages of function values.

While the Lebesgue integral is defined over subsets of $\mathbb{R}^d$ equipped with the Lebesgue measure, its essential ideas extend far beyond this setting. By abstracting away from the underlying Euclidean space and focusing instead on the abstract σ-algebra and its measures, one arrives at the general notion of measure spaces and integration over them. This abstraction not only subsumes the classical Lebesgue integral as a special case but also provides a uniform mathematical framework for defining integration over arbitrary domains. In particular, probability measure and integration with respect to measures play a central role in Chapter 4, where probability measures are defined over spaces of possible worlds rather than over numerical domains. The measure-theoretic perspective developed in this section thus forms the mathematical foundation for the probabilistic semantics of the logics introduced later in the thesis.

Note that most of the definitions and results within this section are directly from a textbook of real analysis [SS09]. For the sake of compactness, we do not give details of the proofs.

3.2.1. Lebesgue Measure and Integral

We begin with the definition of *outer measure*, also known as the exterior measure. Essentially, an outer measure describes the volume of a set E by approximating it from the outside by covering it with hypercubes.

Definition 3.9 (Outer Measure)**.** For any subset of an Euclidean space $E \subset \mathbb{R}^d$, the outer measure of E, written as $m_*(E)$, is defined as

$$m_*(E) = \inf_{\{Q_i\}} \sum_{j=1}^{\infty} |Q_j| \tag{3.2}$$

where $\{Q_i\}$ is an infinite set of closed hypercubes such that $E \subseteq \bigcup_{j=1}^{\infty} Q_j$, and inf stands for the infimum taken over all possible sets of $\{Q_i\}$.

The (Lebesgue) measurability of a set is defined as follows:[3]

Definition 3.10 (Lebesgue Measurable)**.** A set $E \subset \mathbb{R}^d$ is Lebesgue Measurable, if and only if for any $\varepsilon > 0$, there exists an open set $\mathcal{O}$ such that $E \subseteq \mathcal{O}$, and $m_*(\mathcal{O} \backslash E) < \varepsilon$.
For a measurable set, the Lebesgue measure is defined as its exterior measure, i.e. $m(E) = m_*(E)$.

We now turn to the measurability of functions.

Definition 3.11 (Measurable Function)**.** A function f defined on a measurable set E is measurable if for all $r \in \mathbb{R}$, the set $\{x \in E \,|\, f(x) < r\}$ is measurable.

There are two special forms of measurable functions: the *characteristic functions* and the *simple functions*, which are used in defining the integral of arbitrary measurable functions.

Definition 3.12 (Characteristic Function)**.** Given $E \subset \mathbb{R}^d$, the characteristic function of E, written as χ_E, is defined by

$$\chi_E(x) = \begin{cases} 1 & x \in E; \\ 0 & x \notin E. \end{cases} \tag{3.3}$$

Definition 3.13 (Simple Function)**.** A function is *simple* if it is of the form

$$f = \sum_{k=1}^{N} a_k \chi_{E_k} \tag{3.4}$$

where N is a natural number, each E_k is a measurable set of finite measure, and a_k are real numbers.

Any measurable function can be approximated via simple functions.

[3] There are multiple ways of defining Lebesgue measure, which are proved to be equivalent. Here we adopt the definition used in [SS09].

Theorem 3.14. Given a non-negative measurable function f on $\mathbb{R}^d$, there exists a non-decreasing sequence of non-negative simple functions $\{\varphi\}_{k=1}^{\infty}$ that converges to f, i.e. for all $x \in \mathbb{R}^d$,

$$\varphi_k(x) \leq \varphi_{k+1}(x) \quad \text{and} \quad \lim_{k\to\infty} \varphi_k(x) = f(x) \tag{3.5}$$

Then the Lebesgue integral on $\mathbb{R}^d$ can be defined in four steps:

1. Simple functions

We first consider the integral of simple functions, where each simple function can be viewed as a weighted sum of a finite set of characteristic functions. For a simple function of the form $\varphi(x) = \sum_{k=1}^{N} a_k \chi_{E_k}(x)$, the Lebesgue integral is defined as

$$\int_{\mathbb{R}^d} \varphi(x)dx = \sum_{k=1}^{N} a_k m(E_k) \tag{3.6}$$

2. Bounded functions with finite-measure support

A function is *bounded* if there exists a number M such that $|f(x)| < M$ for all x. The *support* of a function f is the preimage where f has a non-zero value, i.e.

$$supp(f) = \{x \,|\, f(x) \neq 0\} \tag{3.7}$$

Then we have the following result:

Theorem 3.15. Let f be a bounded function supported on a set E of finite measure, i.e. $m(E) < \infty$, and let $\{\varphi_k\}_{k=1}^{\infty}$ be an infinite sequence of simple functions which are bounded, supported on E, and converge to f *almost everywhere*[4], then

$$\lim_{k\to\infty} \int \varphi_k(x)dx \tag{3.8}$$

is well-defined. We define $\int f(x)dx$ as the above limit.

3. Non-negative functions

For every non-negative function f (not necessarily bounded), the Lebesgue integral is defined as

$$\int f(x)dx = \sup_g \int g(x)dx \tag{3.9}$$

[4]Meaning that $\varphi_k(x) \to f(x)$ for all x except a set of points with zero measure.

where g is taken over all bounded, measurable functions such that $0 \leq g \leq f$ and g is supported on a set with finite measure. We additionally require that $\int f(x)dx < \infty$.

Although the definition Eq. (3.9) is solid, it is unpleasant to take over an infinite set of functions, as it will make theorem proving extremely difficult. However, when the function f can be approximated by a monotone sequence of measurable functions $\{f_k\}_{k=1}^{\infty}$, i.e. for almost every x (a.e. x)

$$f_k(x) \leq f_{k+1}(x) \text{ for all } k \quad \text{and} \quad \lim_{k\to\infty} f_k(x) = f(x) \tag{3.10}$$

we denote the property as $f_k \nearrow f$. Then we have the following result:

Theorem 3.16. Suppose that $\{f_k\}_{k=1}^{\infty}$ is a sequence of non-negative measurable functions with $f_k \nearrow f$, then

$$\lim_{k\to\infty} \int f_k(x)dx = \int f(x)dx \tag{3.11}$$

4. General cases

For a real-valued measurable function f, we say f is integrable if and only if the non-negative function $|f|$ is integrable. To define the integral of f, we partition f into two parts:

$$f^+(x) = \max(f(x), 0) \quad \text{and} \quad f^-(x) = \max(-f(x), 0) \tag{3.12}$$

Then the Lebesgue integral of f is given by

$$\int f(x)dx = \int f^+(x)dx - \int f^-(x)dx \tag{3.13}$$

3.2.2. Abstract Measure and Integration Theory

In certain situation, considering integration only in Euclidean spaces could be deemed insufficient; a general form of integration based on abstract measure theory will be required. Following the insight of Carathéodory, a very general sense of measures, going far beyond the Lebesgue measure, can be constructed, and the integration based on abstract measure can be defined accordingly.

Definition 3.17 (σ-algebra)**.** Let X be a set, a σ-algebra $\mathcal{A}$ on X is a non-empty collection of subsets of X such that

- if $E \in \mathcal{A}$, then the complement $E^c \in \mathcal{A}$;
- if $\{E_i\}_{i=1}^{\infty} \subseteq \mathcal{A}$, then $\bigcup_i E_i \in \mathcal{A}$;
- if $\{E_i\}_{i=1}^{\infty} \subseteq \mathcal{A}$, then $\bigcap_i E_i \in \mathcal{A}$.

Namely, $\mathcal{A}$ is closed under complements, countable unions, and countable intersections.

Definition 3.18 (Measure Space)**.** A measure space is a tuple $(X, \mathcal{A}, \mu)$, where

- $\mathcal{A}$ is a σ-algebra;
- $\mu : \mathcal{A} \to [0, \infty]$ is a measure such that for a countable sequence of disjoint sets $E_1, E_2, \ldots$,

$$\mu(\bigcup_{i=1}^{\infty} E_i) = \sum_{i=1}^{\infty} \mu(E_i) \tag{3.14}$$

In particular, we call a measure space $(X, \mathcal{A}, \mu)$ a probability space, if $\mu(X) = 1$ and $\mu(\varnothing) = 0$.

Definition 3.19 (Measurable Function)**.** Given a measure $(X, \mathcal{A}, \mu)$, a real-valued function f on X is measurable if for all $r \in \mathbb{R}$,

$$\{x \in X \mid f(x) < r\} \in \mathcal{A} \tag{3.15}$$

With this definition, we define the simple functions on X, which take the form

$$\varphi(x) = \sum_{k=1}^{N} a_k \chi_{E_k} \tag{3.16}$$

where E_k are measurable sets with finite measures, i.e. $\mu(E_k) < \infty$. The integral of a simple function is defined similarly as in Lebesgue's setting:

$$\int \varphi d\mu = \sum_{k=1}^{N} a_k \mu(E_k) \tag{3.17}$$

Then the four-step construction of the Lebesgue integral can be carried out and adapted to general functions defined on measure spaces.

4 On the Unification of Logic and Probability

The unification of logic and probability has long been a central concern in philosophy and mathematical logic. Though many previous works exist, each of them has certain restrictions on expressiveness, as already discussed in Chapter 2. Building upon these analyses, we propose logical accounts that can represent arbitrary-discrete, continuous, or mixed-probability distributions in both static and dynamic settings. The main content of this chapter is based on our previous work [FLLB23, LFBL23] but several modifications have been made to the semantics. The revision avoids some undesired properties and paves the way to perform default reasoning over probability distributions.

The rest of this chapter is organized as follows. In Section 4.1, we introduce the logic $\mathcal{OBL}_c$, where logical models are defined based on probability spaces over possible worlds, and analyze its properties and expressive power concerning belief distributions. In Section 4.2 we extend the account to a dynamic probabilistic logic $\mathcal{PS}$, in which beliefs after actions are characterized via an integral calculus on possible worlds, analogous to Lebesgue integration.

4.1. Probability Spaces as Semantic Models

The logic $\mathcal{OBL}_c$ of only-believing with probabilistic degrees of belief is a first-order, many-sorted epistemic modal logic. For simplicity, we consider only function sym-

bols while predicates are understood as abbreviations of functions which take values from two reserved names $\{\top, \bot\}$. E.g. $P(x)$ means $f_P(x) = \top$, $\neg Q(a, b)$ means $f_Q(a, b) = \bot$ and $n \leq m$ means $f_{\leq}(n, m) = \top$. Constants are understood as nullary functions, i.e. functions without arguments.

4.1.1. Syntax

The language consider two sorts *object* and *number*. The vocabulary consists of *standard names*, *variables*, and *function symbols* of each sort. Standard names can be viewed as a fixed countable set in the domain with the *unique names assumption*. Conventionally, we use n with (sub-)superscripts for object standard names, e.g. $n_1, n_2, \ldots n', n'', \ldots$ etc. We use Arabic numbers both decimals and fractions for number standard names, e.g. 0.8, $\frac{1}{2}, \ldots$, etc. First-order (FO) variables are denoted by $x, y, u, v \ldots$, etc. Function symbols include:

- *rigid function symbols* of every arity, e.g. arithmetical functions $+, \times, e^x$.
- *fluent function symbols* of every arity, such as $HeartRate(x)$.

Here, we use the terminology from the situation calculus [Rei01]. By *rigid*, we mean that the meaning of the function is fixed, while by *fluent*, the meaning of the function might vary in different worlds.

Standard FO connectives $\wedge$, $\neg$, $\forall$ and modal operators $\{\boldsymbol{B}, \boldsymbol{O}\}$ are used to construct formulas.

Definition 4.1 ($\mathcal{OBL}_c$ Term)**.** Terms (for respective sort) of the language are the least set of expressions such that:

1. every standard name and first-order variable is a term;
2. if $t_1, ..., t_k$ are terms and f is a k-ary function symbol, then $f(t_1, ..., t_k)$ is a term of the same sort as f.

A term is said to be *rigid* if and only if it does not contain fluents. *Ground terms* are terms without variables.

The epistemic expression $\boldsymbol{B}(\alpha\colon r)$ should be read as "α is believed with a degree r". $\boldsymbol{K}\alpha$ means "α is known" and is an abbreviation for $\boldsymbol{B}(\alpha\colon 1)$. $\boldsymbol{O}(\alpha\colon r)$ is read as "all that believed is α with degree r". Similarly, $\boldsymbol{O}\alpha$ means "α is only known" and is an abbreviation for $\boldsymbol{O}(\alpha\colon 1)$.

As usual, we treat $\alpha \vee \beta$, $\alpha \supset \beta$, $\alpha \equiv \beta$, and $\exists v.\alpha$ as abbreviations. A *sentence* is a formula without free variables. We use TRUE as an abbreviation for $\forall x.(x = x)$, and FALSE for its negation. A formula without $\boldsymbol{B}$ and $\boldsymbol{O}$ is called *objective*. A formula with no fluent outside $\boldsymbol{B}$ or $\boldsymbol{O}$ is called *subjective*. An objective formula without fluent functions is called a *rigid formula*.

4.1.2. Semantics

The semantics is based on the notion of *possible worlds*, where a ($\mathcal{OBL}_c$) *world* is a Tarski-like structure. Formally, we assume a fixed domain of discourse $\mathcal{D} = \mathcal{D}_{obj} \cup \mathbb{R}$ where $\mathcal{D}_{obj}$ is a countable infinite set of objects and each element is assigned a *co-referring standard name*. $\mathbb{R}$ is the set of real numbers. The set of standard names $\mathcal{N} = \mathcal{N}_{obj} \cup \mathcal{N}_{num}$ is a countable subset of $\mathcal{D}$ and $\mathcal{N}_{obj} = \mathcal{D}_{obj}$. Lastly, we fix $\mathcal{N}_{num}$ to the set of *computable numbers* [Tur36] which is a countable subset of $\mathbb{R}$ but still includes important irrational numbers such as π, e.[1]

Objective Formula

We begin with the semantics of objective formulas. A world is a mapping from all function symbols to functions of the corresponding sorts.

Formally, a world w maps every k-ary object function symbol f_{obj} and number function symbol f_{num} to a function of the corresponding sort, i.e. $w[f_{obj}]\colon (\mathcal{D})^k \to \mathcal{D}_{obj}$ and $w[f_{num}]\colon (\mathcal{D})^k \to \mathbb{R}$, where the following constraints are satisfied:

1. **Rigidity:** if f is a rigid function symbol, then for all w, w' , $w[f] = w'[f]$;

[1] The domain is uncountable but we can only assign names to a countable subset of numbers.

2. **Arithmetical Correctness:** arithmetical function symbol (e.g. $+, \times, e^x$) are interpreted in the usual sense. For example, $w[+](1,1) = 2$ for any w.

Let W be the set of all such worlds. We denote the set of all (first-order) variables as $\mathcal{V}$. A **variable map** λ maps each element in $\mathcal{V}$ to $\mathcal{D}$ of the right sort. We write $\lambda \sim_v \lambda'$ to mean λ and λ' agree excepts perhaps on variable v. The *denotation* of terms is defined recursively:

Definition 4.2 (Denotation)**.** The *denotation* of a term t under world w and variable map λ, written as $\|t\|_{w,\lambda}$, is defined as:

- $\|t\|_{w,\lambda} = t$ if $t \in \mathcal{N}$;
- $\|t\|_{w,\lambda} = \lambda(t)$ if $t \in \mathcal{V}$;
- $\|t\|_{w,\lambda} = w[f](\|t_1\|_{w,\lambda}, ..., \|t_k\|_{w,\lambda})$ if term t is of the form $f(t_1, ..., t_k)$.

For simplicity, we write $\|t\|_\lambda$ when t is rigid, $\|t\|_w$ when t does not contain variables, and $\|t\|$ when t is both rigid and ground. By an *objective model* we mean a pair (w, λ). Truth of objective formulas is then given as:

- $w, \lambda \models t_1 = t_2$ iff $\|t_1\|_{w,\lambda}$ and $\|t_2\|_{w,\lambda}$ are identical;
- $w, \lambda \models \neg\alpha$ iff $w, \lambda \not\models \alpha$;
- $w, \lambda \models \alpha \wedge \beta$ iff $w, \lambda \models \alpha$ and $w, \lambda \models \beta$;
- $w, \lambda \models \forall v.\alpha$ iff $w, \lambda' \models \alpha$ for all $\lambda' \sim_v \lambda$.

Beliefs

To give the semantics of $\boldsymbol{B}$ and $\boldsymbol{O}$, we need the notion of epistemic state. We begin with a brief recap of some key concepts of probability theory. A measure space is a tuple $\langle X, \mathcal{X}, \mu \rangle$ where X is a set , $\mathcal{X}$ is a σ-algebra on the set X (i.e. a non-empty collection of subsets closed under complements and countable unions), and $\mu : \mathcal{X} \mapsto [0, +\infty]$ is a measure. More concretely, $\mathcal{X}$ is a non-empty subset of 2^X such that:

- for all $X_1, X_2, \ldots \in \mathcal{X}$, it holds $\bigcup_i X_i \in \mathcal{X}$ (closed under countable unions); and
- for all $X' \in \mathcal{X}$, $X'_c = X \backslash X' \in \mathcal{X}$ (closed under complements).

Moreover, μ is a mapping $\mu : \mathcal{X} \mapsto [0, +\infty]$ that satisfies:

- $\mu(\varnothing) = 0$;
- for all pairwise disjoint $X_1, X_2, \ldots \in \mathcal{X}$, it holds $\mu(\bigcup_i X_i) = \sum_i \mu(X_i)$.

Typical measure spaces include the Lebesgue measure spaces $\langle \mathbb{R}^k, \mathcal{M}, m \rangle$ where $\mathbb{R}^k$ is a k-dimensional Euclidean space, $\mathcal{M}$ is a σ-algebra on $\mathbb{R}^k$, and m is the Lebesgue measure. A probability space is a special measure space whose measure is normalized, i.e. $\mu(X) = 1$. For probability spaces, usually X is called the *sample space*, $\mathcal{X}$ the *event set*, μ the *probability measure*. A probability space $\langle X, \mathcal{X}, \mu \rangle$ is said to be *complete* if for all $B \in \mathcal{X}$ with $\mu(B) = 0$ and all $A \subseteq B$, one has $A \in \mathcal{X}$ and $\mu(A) = 0$. Intuitively, completeness means that if an event has zero probability, any subset of it is also an event and has zero probability; likewise, if an event has probability 1, all its supersets are events and have probability 1. We restrict ourselves to complete probability spaces since each probability space can be uniquely extended to a complete probability space.

An $\mathcal{OBL}_c$ epistemic state e is then defined as a set of $\langle \mathcal{W}, \mu \rangle$ pairs s.t. $\langle W, \mathcal{W}, \mu \rangle$ forms a complete probability space.[2] We use the letter s to denote a $\langle \mathcal{W}, \mu \rangle$ pair, and directly call s a probability space. We expand the model with the epistemic state. Namely, a model is now a triple (e, w, λ). For objective formulas, truth is given the same as before since the epistemic state e plays no role.

Let $W_\alpha^{e,\lambda} = \{w' \mid e, w', \lambda \models \alpha\}$, i.e. $W_\alpha^{e,\lambda}$ is the set of worlds which satisfy α, when the state and the map are fixed. Specifically, when e contains only one element, i.e. $e = \{s\}$, we write $W_\alpha^{s,\lambda}$ instead of $W_\alpha^{\{s\},\lambda}$. In case α has no free

[2] As observed in [GL07], using a single probability space as an epistemic state would result in that agents necessarily have *de re* knowledge about their degrees of belief, i.e. for all ϕ, $\exists x.\boldsymbol{KB}(\phi\colon x)$ is valid, which is counter-intuitive.

variables, we write W_α^e. Let t denote a rigid term. Truth for $\boldsymbol{B}$-formulae is given by:

- $e, w, \lambda \models \boldsymbol{B}(\alpha : t)$ iff for all $s = \langle \mathcal{W}, \mu \rangle$,
  $s \in e$ implies $W_\alpha^{e,\lambda} \in \mathcal{W}$ and $\mu(W_\alpha^{e,\lambda}) = \|t\|_\lambda$;

Namely, $e, w, \lambda \models \boldsymbol{B}(\alpha : t)$ if for all $\langle \mathcal{W}, \mu \rangle \in e$, the set of worlds that satisfy α, i.e. $W_\alpha^{e,\lambda}$, is an event in $\mathcal{W}$ and has probability measure $\|t\|_\lambda$. Unlike the conference proceedings version of $\mathcal{OBL}_c$, where only one probability space is involved in defining the event [FLLB23], here we adopt an **"add-one-in"** semantics for $\boldsymbol{O}$-formulae.

Suppose that an epistemic state e models an only-believing formula $\boldsymbol{O}(\alpha : r)$, given a probability space $s = \langle \mathcal{W}, \mu \rangle$, to decide whether e includes s, we consider the (possibly new) state $e' = e \cup \{s\}$ and test the measure of the set of world[3]

$$W_\alpha^{e',\lambda} = \{w' \mid e', w', \lambda \models \alpha\}$$

In this way, both e and s are involved in defining the set. Formally, the truth of $\boldsymbol{O}$-formulae is given by:

- $e, w, \lambda \models \boldsymbol{O}(\alpha : r)$ iff for all $s = \langle \mathcal{W}, \mu \rangle$,
  $s \in e$ *iff* $W_\alpha^{e+s,\lambda} \in \mathcal{W}$ and $\mu(W_\alpha^{e+s,\lambda}) = \|r\|_\lambda$.

where $e + s = e \cup \{s\}$. Intuitively, $e, w, \lambda \models \boldsymbol{O}(\alpha : r)$ if and only if e is a maximal set of such probability spaces. We will discuss the properties of this add-one-in semantics later and, by the end of this chapter, compare it with the traditional semantics of only-knowing. A similar design is also adopted by the accounts in Section 4.2 and Chapter 6.

For a sentence α, we write $e, w \models \alpha$ to mean $e, w, \lambda \models \alpha$ for all variable maps λ. When Σ is a set of sentences and α is a sentence, we write $\Sigma \models \alpha$ (read: Σ logically entails α) to mean that for every e and w, if $e, w \models \alpha'$ for every $\alpha' \in \Sigma$,

[3] Clearly, $e' = e$ if and only if $s \in e$.

then $e, w \models \alpha$. We say that α is valid ($\models \alpha$) if $\{\} \models \alpha$. Satisfiability is then defined in the standard way. If α is an objective sentence, we write $w \models \alpha$ instead of $e, w \models \alpha$. Similarly, we write $e \models \alpha$ instead of $e, w \models \alpha$ if α is subjective.

4.1.3. Properties of $\mathcal{OBL}_c$

In this subsection, we show that our logic has many reasonable properties. As we shall see, many properties in the logic $\mathcal{OBL}$ [BLL16] are retained.

To begin with, we have these properties in terms of validity and satisfiability:

- $\boldsymbol{B}(\textsc{True}\colon 1)$ is valid, i.e. $\models \boldsymbol{K}\textsc{True}$

  Proof. For any world $w \in W$, $w \models \textsc{True}$; for all epistemic state e and $\langle \mathcal{W}, \mu \rangle \in e$, $W^e_{\textsc{True}} = \{w \mid e, w \models \textsc{True}\} = W$. Since $\langle \mathcal{W}, \mu \rangle$ forms a probability space, $W \in \mathcal{W}$ and $\mu(W) = 1$. Hence $e \models \boldsymbol{B}(\textsc{True}\colon 1)$. □

- for negative r, $\boldsymbol{B}(\alpha\colon r)$ is satisfied only by the empty epistemic state $\varnothing$.

  Proof. Supposing $e \neq \varnothing$ and $e \models \boldsymbol{B}(\alpha\colon r)$ with $\|r\| < 0$, then for all $\langle \mathcal{W}, \mu \rangle \in e$, $\mu(W^e_\alpha) = \|r\| < 0$. This contradicts with the definition of μ. □

Additivity and Equivalence

Besides, $\boldsymbol{B}$ entertains the properties of probability as in [FH94]:

- if $\models \alpha \equiv \beta$, then $\models \boldsymbol{B}(\alpha\colon r) \equiv \boldsymbol{B}(\beta\colon r)$ for all r.

  Proof. For any e, and $\langle \mathcal{W}, \mu \rangle \in e$, since $W^e_\alpha = W^e_\beta$, if $\mu(W^e_\alpha) = \|r\|$, then $\mu(W^e_\beta) = \|r\|$ and vice versa. Thus $e \models \boldsymbol{B}(\alpha\colon r) = \boldsymbol{B}(\beta\colon r)$. □

- $\models \boldsymbol{B}(\alpha : r) \supset \boldsymbol{B}(\neg\alpha : 1 - r)$
  It means $\boldsymbol{B}$ satisfies the complement law of probability.

  Proof. Suppose $e \models \boldsymbol{B}(\alpha : r)$, then for all $\langle \mathcal{W}, \mu \rangle \in e$, $\mu(W^e_\alpha) = \|r\|$. Then $\mu(W^e_{\neg\alpha}) = \mu(W \backslash W^e_\alpha) = \mu(W) - \mu(W^e_\alpha) = 1 - \|r\|$. □

- $\models \boldsymbol{B}(\alpha \wedge \beta \colon r) \wedge \boldsymbol{B}(\alpha \wedge \neg\beta \colon r') \supset \boldsymbol{B}(\alpha \colon r + r')$

  Proof. Suppose $e \models \boldsymbol{B}(\alpha \wedge \beta \colon r)$ and $e \models \boldsymbol{B}(\alpha \wedge \neg\beta \colon r')$, this means for all $\langle \mathcal{W}, \mu \rangle \in e$, $\mu(W^e_{\alpha\wedge\beta}) = \|r\|$ and $\mu(W^e_{\alpha\wedge\neg\beta}) = \|r'\|$. Since $W^e_{\alpha\wedge\beta} \cap W^e_{\alpha\wedge\neg\beta} = \varnothing$ and $W^e_{\alpha\wedge\beta} \cup W^e_{\alpha\wedge\neg\beta} = W^e_\alpha$, we have $\mu(W^e_\alpha) = \mu(W^e_{\alpha\wedge\beta} \cup W^e_{\alpha\wedge\neg\beta}) = \mu(W^e_{\alpha\wedge\beta}) + \mu(W^e_{\alpha\wedge\neg\beta}) = \|r\| + \|r'\|$. □

- $\models \boldsymbol{B}(\alpha \colon r) \wedge \boldsymbol{B}(\beta \colon r') \wedge \boldsymbol{B}(\alpha \wedge \beta \colon r'') \supset \boldsymbol{B}(\alpha \vee \beta \colon r + r' - r'')$

  That is $\boldsymbol{B}$ satisfies the addition law of probability.

  Proof. Supposing that $\models \boldsymbol{B}(\alpha \colon r) \wedge \boldsymbol{B}(\beta \colon r') \wedge \boldsymbol{B}(\alpha \wedge \beta \colon r'')$, it means that for all $\langle \mathcal{W}, \mu \rangle \in e$, $\mu(W^e_\alpha) = \|r\|$, $\mu(W^e_\beta) = \|r'\|$ and $\mu(W^e_{\alpha\wedge\beta}) = \|r''\|$. Then,

$$\begin{aligned}
\mu(W^e_{\alpha\vee\beta}) &= \mu(W^e_{\alpha\wedge\beta} \cup W^e_{\alpha\wedge\neg\beta} \cup W^e_{\neg\alpha\wedge\beta}) \\
&= \mu(W^e_{\alpha\wedge\beta}) + \mu(W^e_{\alpha\wedge\neg\beta}) + \mu(W^e_{\neg\alpha\wedge\beta}) \\
&= \big(\mu(W^e_{\alpha\wedge\beta}) + \mu(W^e_{\alpha\wedge\neg\beta})\big) + \big(\mu(W^e_{\alpha\wedge\beta}) + \mu(W^e_{\neg\alpha\wedge\beta})\big) - \mu(W^e_{\alpha\wedge\beta}) \\
&= \mu(W^e_\alpha) + \mu(W^e_\beta) - \mu(W^e_{\alpha\wedge\beta}) \\
&= \|r + r' - r''\|
\end{aligned}$$

  □

All the above properties follow from the fact that $\boldsymbol{B}$ is essentially a probability over possible worlds.

Knowledge

Recall that $\boldsymbol{K}\alpha$ is an abbreviation for $\boldsymbol{B}(\alpha \colon 1)$. The knowledge modality $\boldsymbol{K}$ also satisfies many properties in the epistemic logic $\mathcal{KL}$ by Levesque and Lakemeyer [LL01], including universal and existential versions of the Barcan formula:

- $\models \boldsymbol{K}\alpha \supset \boldsymbol{K}(\alpha \vee \beta)$

Proof. Suppose that $e \models \boldsymbol{K}\alpha$, then for all $\langle \mathcal{W}, \mu \rangle \in e$, $\mu(W_\alpha^e) = 1$. Since $W_\alpha^e \subseteq W_{\alpha\vee\beta}^e$, the completeness of the probability space $\langle W, \mathcal{W}, \mu \rangle$ guarantees that $\mu(W_{\alpha\vee\beta}^e) = 1$. □

- $\models \boldsymbol{K}\alpha \wedge \boldsymbol{K}\beta \supset \boldsymbol{K}(\alpha \wedge \beta)$

  Proof. Suppose that $e \models \boldsymbol{K}\alpha \wedge \boldsymbol{K}\beta$, then by the complement law we have $e \models \boldsymbol{B}(\neg\alpha : 0) \wedge \boldsymbol{B}(\neg\beta : 0)$. Further, with the first property of knowledge and the complement law we have $e \models \boldsymbol{B}(\neg\alpha\wedge\neg\beta : 0)$. By the addition law of probability, i.e. $\models \boldsymbol{B}(\neg\alpha : 0)\wedge\boldsymbol{B}(\neg\beta : 0)\wedge\boldsymbol{B}(\neg\alpha\wedge\neg\beta : 0) \supset \boldsymbol{B}(\neg\alpha\vee\neg\beta : 0)$. Thus $e \models \boldsymbol{K}(\alpha \wedge \beta)$ must hold. □

- $\models \boldsymbol{K}\alpha \wedge \boldsymbol{K}(\alpha \supset \beta) \supset \boldsymbol{K}\beta$

  Proof. Suppose that $e \models \boldsymbol{K}\alpha \wedge \boldsymbol{K}(\alpha \supset \beta)$. To prove $e \models \boldsymbol{K}\beta$, it suffice to prove that $e \models \boldsymbol{K}((\alpha \vee \beta) \wedge (\neg\alpha \vee \beta))$. $\boldsymbol{K}(\alpha \vee \beta)$ can be proved by the first property of knowledge. $\boldsymbol{K}(\neg\alpha \vee \beta)$ is equivalent to $\boldsymbol{K}(\alpha \supset \beta)$. Based on the second property of knowledge $e \models \boldsymbol{K}\beta$. □

- $\models \exists x.\boldsymbol{K}\alpha \supset \boldsymbol{K}\exists x.\alpha$

  Proof. Suppose that $e, \lambda \models \exists x.\boldsymbol{K}\alpha$, then there exists λ' with $\lambda \sim_x \lambda'$ such that $e, \lambda' \models \boldsymbol{K}\alpha$. Namely, for all $\langle \mathcal{W}, \mu \rangle \in e$, $\mu(W_\alpha^{\langle \mathcal{W},\mu\rangle,\lambda'}) = 1$ where by definition
  $W_\alpha^{\langle \mathcal{W},\mu\rangle,\lambda'} = \{w | \langle \mathcal{W}, \mu \rangle, w, \lambda' \models \alpha\}$. On the other hand,

$$\begin{aligned} W_{\exists x.\alpha}^{\langle \mathcal{W},\mu\rangle,\lambda} &= \{w | \langle \mathcal{W}, \mu \rangle, w, \lambda \models \exists x.\alpha\} \\ &= \bigcup_{\{\lambda'' | \lambda \sim_x \lambda''\}} \{w | \langle \mathcal{W}, \mu \rangle, w, \lambda'' \models \alpha\} \end{aligned}$$

  Thus $W_\alpha^{\langle \mathcal{W},\mu\rangle,\lambda'} \subseteq W_{\exists x.\alpha}^{\langle \mathcal{W},\mu\rangle,\lambda}$. Since $\mu(W_\alpha^{\langle \mathcal{W},\mu\rangle,\lambda'}) = 1$, we have $\mu(W_{\exists x.\alpha}^{\langle \mathcal{W},\mu\rangle,\lambda}) = 1$ □

The converse of the above formula does not hold: knowing that α holds for someone does not imply knowing that individual. For universal quantification, we have

- $\models \boldsymbol{K}\forall x.\alpha \supset \forall x.\boldsymbol{K}\alpha$

  Proof. Suppose that $e, \lambda \models \boldsymbol{K}\forall x.\alpha$, then for any $\langle \mathcal{W}, \mu\rangle \in e$, $W^{e,\lambda}_{\forall x.\alpha} \in \mathcal{W}$ and $\mu(W^{e,\lambda}_{\forall x.\alpha}) = 1$.

  $$\begin{aligned} W^{e,\lambda}_{\forall x.\alpha} &= \{w | e, w, \lambda \models \forall x.\alpha\} \\ &= \{w | \text{for all } \lambda' \text{ s.t. } \lambda' \sim_x \lambda, e, w, \lambda' \models \alpha\} \\ &= \bigcap_{\{\lambda' | \lambda \sim_x \lambda'\}} W^{e,\lambda'}_{\alpha} \end{aligned}$$

  Thus, for all λ' with $\lambda \sim_x \lambda'$, $W^{e,\lambda}_{\forall x.\alpha} \subseteq W^{e,\lambda'}_{\alpha}$. Since $\mu(W^{e,\lambda}_{\forall x.\alpha}) = 1$, for all λ' with $\lambda' \sim_x \lambda$, $\mu(W^{e,\lambda'}_{\alpha}) = 1$. Thus $e, \lambda' \models \boldsymbol{K}\alpha$ for all $\lambda' \sim_x \lambda$, i.e. $e, \lambda \models \forall x.\boldsymbol{K}\alpha$ □

However, in general the converse of the above formula may not hold:

- $\not\models \forall x.\boldsymbol{K}\alpha \supset \boldsymbol{K}\forall x.\alpha$

  The reason is: in a probability space with uncountably many samples, there could be uncountably many distinct events where each of them has probability 1, the countable intersection of any of these events has probability 1, yet an uncountable intersection does not have probability 1.

  More concretely, we provide a counter-example as follows, assuming h is a nullary fluent taking values from $\mathbb{R}$, then
  $\not\models \forall x.\boldsymbol{K}(h \geq 0 \vee h \neq x) \supset \boldsymbol{K}\forall x.(h \geq 0 \vee h \neq x)$.

  Informally, given an epistemic state $e = \langle W, \mu\rangle$, the fluent h can be viewed as a random variables over $\mathbb{R}$. We further assume that the distribution of h is standard Gaussian in e. For this particular e, $e \models \forall x.\boldsymbol{K}(h \geq 0 \vee h \neq x)$ yet $e \not\models \boldsymbol{K}\forall x.(h \geq 0 \vee h \neq x)$.

For the former part: for all real $r < 0$, $h \geq 0 \vee h \neq r$ is equivalent to $h \neq r$. The probability of this event would be 1: it leaves out a single point ($h = r$) from the whole sample space of a Gaussian distribution. The remaining samples have a probability of 1. Additionally, for all real $r \geq 0$, $h \geq 0 \vee h \neq r$ is equivalent to $\textsc{True}$, hence the event $h \geq 0 \vee h \neq r$ has probability 1 trivially. Together, we have $e \models \forall x.\boldsymbol{K}(h \geq 0 \vee h \neq x)$.

For the later part: $\forall x.(h \geq 0 \vee h \neq x)$ is equivalent to $h \geq 0$, hence $e \models \boldsymbol{B}(\forall x.(h \geq 0 \vee h \neq x) : \frac{1}{2})$ (recall h is distributed as a standard Gaussian in e). Hence $e \not\models \boldsymbol{K}\forall x.(h \geq 0 \vee h \neq x)$.

Nevertheless, if we only consider the countable object domain, we still have:

- $\models \forall x.\boldsymbol{K}\alpha \supset \boldsymbol{K}\forall x.\alpha$ if x is a variable of sort object.

Proof. Suppose that $e, \lambda \models \forall x.\boldsymbol{K}\alpha$. By the semantics we have $e, \lambda' \models \boldsymbol{K}\alpha$ for all $\lambda' \sim_x \lambda$. For all $\langle \mathcal{W}, \mu \rangle \in e$ and $\lambda' \sim_x \lambda$, $\mu(W_\alpha^{e,\lambda'}) = 1$. Since x is of sort object, the set $\{\lambda' | \lambda' \sim_x \lambda\}$ is countable. Therefore

$$\begin{aligned}
\mu(W \backslash W_{\forall x.\alpha}^{e,\lambda}) &= \mu(W \backslash \bigcap_{\{\lambda' | \lambda' \sim_x \lambda\}} W_\alpha^{e,\lambda'}) \\
&= \mu(\bigcup_{\{\lambda' | \lambda' \sim_x \lambda\}} (W \backslash W_\alpha^{e,\lambda'})) \\
&\leq \sum_{\{\lambda' | \lambda' \sim_x \lambda\}} \mu(W \backslash W_\alpha^{e,\lambda'}) = 0
\end{aligned}$$

The inequality above is derived from the property of measure and based on the fact that $\{\lambda' | \lambda' \sim_x \lambda\}$ is countable. Thus $\mu(W_{\forall x.\alpha}^{e,\lambda}) = 1$ for all $\langle \mathcal{W}, \mu \rangle \in e$, i.e. $e, \lambda \models \boldsymbol{K}\forall x.\alpha$. □

E.g. $\models \forall x.\boldsymbol{K}(\mathit{father}(x) \neq \mathit{joe}) \supset \boldsymbol{K}\forall x.\mathit{father}(x) \neq \mathit{joe}$. Namely, if for all persons, it's known that the person's father is not *joe*, as a consequence, it's known that *joe* is not anyone's father.

Introspection

Lastly, let us turn to introspection. Formally, for positive introspection, we have:

- $\models \boldsymbol{B}(\alpha : r) \supset \boldsymbol{K}\boldsymbol{B}(\alpha : r)$

  Proof. Supposing $e \models \boldsymbol{B}(\alpha : r)$, by the semantics for all $\langle \mathcal{W}, \mu \rangle$ $\in e$, $\mu(W_\alpha^{\langle \mathcal{W},\mu \rangle}) = r$. On the other hand, $W^{\langle \mathcal{W},\mu \rangle}_{\boldsymbol{B}(\alpha : r)} = W$ if $\langle \mathcal{W}, \mu \rangle \models \boldsymbol{B}(\alpha : r)$. Hence for all $\langle \mathcal{W}, \mu \rangle \in e$, $\mu(W^{\langle \mathcal{W},\mu \rangle}_{\boldsymbol{B}(\alpha : r)})$ $= \mu(W) = 1$. By the semantics, $e \models \boldsymbol{K}\boldsymbol{B}(\alpha : r)$ □

Note that $\models \boldsymbol{K}\boldsymbol{B}(\alpha : r) \supset \boldsymbol{B}(\alpha : r)$ holds as well. The proof is in the same spirit as above. As a special case, we have the usual $\boldsymbol{K}$ properties [FHMV04]:

- $\models \boldsymbol{K}\alpha \supset \boldsymbol{K}\boldsymbol{K}\alpha$

Unlike the conference variant of $\mathcal{OBL}_c$ [FLLB23], negative introspection holds in the semantics presented in this thesis.

- $\models \neg\boldsymbol{B}(\alpha : r) \supset \boldsymbol{K}\neg\boldsymbol{B}(\alpha : r)$

Proof. Suppse that $e \models \neg\boldsymbol{B}(\alpha : r)$, by the semantics, it exists $\langle \mathcal{W}, \mu \rangle \in e$ such that either $W^e_\alpha \notin \mathcal{W}$ or $\mu(W^e_\alpha) \neq \|r\|$. Since $W^e_{\neg\boldsymbol{B}(\alpha : r)} = \{w \mid e, w \models \neg\boldsymbol{B}(\alpha : r)\} = W$, for all $\langle \mathcal{W}, \mu \rangle \in e$, $W^e_{\neg\boldsymbol{B}(\alpha : r)} \in \mathcal{W}$ and $\mu(W^e_{\neg\boldsymbol{B}(\alpha : r)}) = 1$, i.e. $e \models \boldsymbol{K}\neg\boldsymbol{B}(\alpha : r)$ □

The difference is that in [FLLB23], the truth of $\boldsymbol{B}(\alpha : r)$ is determined by the event $W_\alpha^{\langle W,\mu \rangle,\lambda}$, which is only related to individual space $\langle W, \mu \rangle$, while the version in this thesis defines the event $W_\alpha^{e,\lambda}$ considering the whole epistemic state.

Only-believing

Another important part of our formalism lies in the notion of only-knowing (or only-believing) [LL01]. Only-knowing captures the intuition that the beliefs and non-beliefs of an agent are precisely those that follow from its knowledge base. Hence it is useful to characterize a knowledge base. Here, we show that our

modality $\boldsymbol{O}$ faithfully captures the notion of only-believing by examining its properties. To begin with, the unique model theorem holds for only-believing as in the work [LL21].

Theorem 4.3 (Unique Model Theorem)**.** For any objective sentence α and rigid ground term r, there is an unique epistemic state e such that $e \models \boldsymbol{O}(\alpha : r)$

Proof. By the semantics of $\boldsymbol{O}$, we have $e \models \boldsymbol{O}(\alpha : r)$ iff

$$e = \{s = \langle \mathcal{W}, \mu \rangle \mid W_\alpha^{e+s} \in \mathcal{W} \text{ and } \mu(W_\alpha^{e+s}) = \|r\|\}.$$

Since α is objective, the set W_α^{e+s} is irrelevant to the state $e + s$ and therefore e is unique. □

While only-believing an objective formula is always uniquely satisfiable, it may be the case that e is empty. For example $\varnothing \models \boldsymbol{O}(\textsc{False}\colon 1)$. Likewise, the maximal epistemic state $e_{\textsc{True}} \models \boldsymbol{O}(\textsc{True}\colon 1)$ where $e_{\textsc{True}}$ is the set of all complete probability space of the form $\langle W, \mathcal{W}, \mu \rangle$.

Besides, as in the logic $\mathcal{OL}$ [LL01], only-knowing implies knowing and not knowing about what is not entailed by the knowledge base. Below, let α be an arbitrary sentence and ϕ, ψ objective sentences:

- $\models \boldsymbol{O}\alpha \supset \boldsymbol{K}\alpha$
- $\models \boldsymbol{O}\phi \supset \boldsymbol{K}\psi$ iff $\models \phi \supset \psi$

  Proof. Since ϕ is objective, for all epistemic state e and $\langle \mathcal{W}, \mu \rangle \in e$, the set W_ϕ^e is irrelevant to e and hence we write it as W_ϕ (respectively we write W_ψ).

  ($\Rightarrow$) Suppose that $\models \boldsymbol{O}\phi \supset \boldsymbol{K}\psi$, i.e. for all $\langle \mathcal{W}, \mu \rangle$, if $\mu(W_\phi) = 1$, then $\mu(W_\psi) = 1$.[4]

  Assuming that $\not\models \phi \supset \psi$, then there exists w such that $w \models \phi$ but $w \not\models \psi$. Let $\mathcal{W}'$ be any σ-algebra which contains $\{w\}$, W_ϕ and W_ψ as elements. Let

[4] Since ϕ is objective, $\mu(W_\phi) = 1$ already implies $\langle \mathcal{W}, \mu \rangle \in e$.

μ' be a probability measure such that $\mu'(\{w\}) = 0.1$ and $\mu'(W_\phi \backslash \{w\}) = 0.9$. Then $\mu'(W_\phi) = 1$ but $\mu'(W_\psi) \leq \mu'(W \backslash \{w\}) = 0.9$ which leads to contradiction. Thus $\models \phi \supset \psi$.

($\Leftarrow$) Suppose that $\models \phi \supset \psi$, for all $w \in W$, $w \models \phi$ implies $w \models \psi$. Thus $W_\phi \subseteq W_\psi$. For all $\langle \mathcal{W}, \mu \rangle$, if $\mu(W_\phi) = 1$, then $\mu(W_\psi) = 1$. Hence $\models \boldsymbol{O}\phi \supset \boldsymbol{K}\psi$ □

- $\models \boldsymbol{O}\phi \supset \neg \boldsymbol{K}\psi$ iff $\phi \not\models \psi$

  This is an easy consequence of the above property and Theorem 4.3.

- $\boldsymbol{O}\phi \not\models \boldsymbol{K}\psi$ iff $\phi \not\models \psi$

  Proof. Suppose that $e \models \boldsymbol{O}\phi$. By Theorem 4.3 e is unique. Hence $\boldsymbol{O}\phi \not\models \boldsymbol{K}\psi$ iff $e \not\models \boldsymbol{K}\psi$, i.e. $\boldsymbol{O}\phi \models \neg \boldsymbol{K}\psi$. □

More generally, we have properties for only-believing as follows:

- $\models \boldsymbol{O}(\alpha : r) \supset \boldsymbol{B}(\alpha \colon r)$
- $\neg \boldsymbol{K}\textsc{False} \models \boldsymbol{O}(\alpha : r) \supset \neg \boldsymbol{B}(h(\vec{n}) = m : r')$ for all r, r', where $\vec{n}, m$ are standard names and h is a fluent not in α.

E.g. $\boldsymbol{O}(1.7 \leq \mathit{heightOf}(A) \colon 0.9) \models \boldsymbol{B}(1.7 \leq \mathit{heightOf}(A) \colon 0.9)$ and $\boldsymbol{O}(1.7 \leq \mathit{heightOf}(A) \colon 0.9) \models \forall x. \neg \boldsymbol{B}(\mathit{weightOf}(A) \leq 50 \colon x)$
That is, only-believing the person A's height is greater than 1.7m with degree 0.9 entails believing the person's height is greater than 1.7m with degree 0.9, and also, not believing A's weight is no more than 50kg with any degree.

The following properties provide us with criteria under which sentences can be conjoined to what is only-believed without actually changing the epistemic state.

- $\models \boldsymbol{O}(\alpha \colon r) \wedge \boldsymbol{B}(\beta \colon r') \supset \boldsymbol{O}(\alpha \colon r, \beta \colon r')$[5]

[5]For simplicity, we write $\boldsymbol{O}(\boldsymbol{B}(\alpha_1 \colon r_1) \wedge \cdots \wedge \boldsymbol{B}(\alpha_k \colon r_k))$ as $\boldsymbol{O}(\alpha_1 \colon r_1, \ldots, \alpha_k \colon r_k)$.

Proof. Suppose that $e \models \boldsymbol{O}(\alpha\colon r) \wedge \boldsymbol{B}(\beta\colon r')$. To prove $e \models \boldsymbol{O}(\alpha\colon r, \beta\colon r')$, it is equivalent to prove that for all $s = \langle \mathcal{W}, \mu \rangle$, $\langle \mathcal{W}, \mu \rangle \in e$ iff $\mu(W_\alpha^{e+s}) = \|r\|$ and $\mu(W_\beta^{e+s}) = \|r'\|$.
($\Rightarrow$): For any $s \in e$, $e + s = e$. Since $e \models \boldsymbol{B}(\alpha\colon r) \wedge \boldsymbol{B}(\beta\colon r')$, it is obvious that the RHS holds.
($\Leftarrow$) Suppose that the RHS holds. Since $e \models \boldsymbol{O}(\alpha\colon r)$, $\mu(W_\alpha^{e+s}) = \|r\|$ implies $s \in e$. □

Analogously, we can extend the result to absorption of arbitrary subjective formulas:

- For any subjective σ, $\models \boldsymbol{O}(\alpha\colon r) \wedge \sigma \supset \boldsymbol{O}(\alpha\colon r, \sigma\colon 1)$

4.1.4. Reasoning about Distributions

We demonstrate how to represent probability distributions and how to reason about distributions in our logic.

The Nullary Fluents Fragment

In the works [LL21, LF21] (extensions to the logic $\mathcal{DS}$), a formula of the form $\forall \vec{x}.\boldsymbol{B}(\vec{h} = \vec{x}\colon f(\vec{x}))$ is used to express a joint *discrete* distribution, i.e. *belief distribution*, over a finite set of random variables, where $\vec{h}$ stands for a set of *nullary fluents* $h_1, h_2, \ldots h_k$. Each h_i is a random variable, and $\vec{h} = \vec{x}$ denotes $\bigwedge_{i=1}^{k} h_i = x_i$. f is a rigid mathematics function s.t. $\sum_{\vec{x}} f(\vec{x}) = 1$ describing the joint distribution.[6] In our logic, we can use a similar formula to express such belief distribution: a (potentially continuous) belief distribution over a finite set of nullary fluents $h_1, h_2, \ldots h_k$ is a formula of the form

$$\forall \vec{x}.\boldsymbol{B}(\bigwedge_{i=1}^{k} h_i \leq x_i\colon f(\vec{x})) \tag{4.1}$$

[6] It is also possible to use axioms to define such functions [LF21], for example, $\forall x \forall y.f_{=1}(x) = y \equiv x = 1 \wedge y = 1 \vee x \neq 1 \wedge y = 0$ defines the indicator function $f_{=1}(x)$ which checks if $x = 1$.

where f is a rigid mathematical function which is monotonic non-decreasing and satisfies $\lim_{\vec{x}\to-\infty} f(\vec{x}) = 0$ and $\lim_{\vec{x}\to\infty} f(\vec{x}) = 1$.[7] We use $\boldsymbol{B}^f$ for short. Intuitively, $f(\vec{x})$ represents the joint *cumulative distribution function* (CDF) of random variables $\vec{h}$.

Proposition 4.4. Given a belief distribution $\boldsymbol{B}^f$ over a finite set of nullary fluents $\vec{h}$, there exists a non-empty epistemic state e such that $e \models \boldsymbol{B}^f$.

The proof is based on the fact that there exists a subset $W_{\mathbb{R}^k}$ of W such that $W_{\mathbb{R}^k}$ is isomorphic to $\mathbb{R}^k$ in the sense that for all $\vec{x} \in \mathbb{R}^k$ there is a unique $w_{\vec{x}} \in W_{\mathbb{R}^k}$ s.t. $w_{\vec{x}} \models \bigwedge_i h_i = x_i$. Likewise, one can select a σ-algebra $\mathcal{W}_{\mathbb{R}^k}$ of $W_{\mathbb{R}^k}$ such that $\mathcal{W}_{\mathbb{R}^k}$ is isomorphic to the Borel σ-algebra of $\mathbb{R}^k$. Based on the Caratheodory's Extension Theorem [SS09], there exists a unique probability measure $\tilde{\mu}$ on $\mathcal{W}_{\mathbb{R}^k}$ such that $\tilde{\mu}(\{w \in W_{\mathbb{R}^k} | \bigwedge_i w[h_i] \leq x_i\}) = f(\vec{x})$ for any $\vec{x} \in \mathbb{R}^k$. Lastly, $\langle \mathcal{W}_{\mathbb{R}^k}, \tilde{\mu} \rangle$ can be extended to $\langle \mathcal{W}, \mu \rangle$, a pair of σ-algebra and measure on W, by assigning zero-measure on any subset of $W \backslash W_{\mathbb{R}^k}$, and $\{\langle \mathcal{W}, \mu \rangle\}$ is the desired epistemic state.

Example 4.5. Our notion of belief distribution can express both discrete and continuous distributions.

- The belief distribution $\forall x.\boldsymbol{B}(h \leq x\colon f_d(x))$ of the nullary fluent h with

$$f_d(x) = \begin{cases} 0 & x < 1 \\ 0.5 & 1 \leq x < 2 \\ 1 & 2 \leq x \end{cases}$$

  expresses a discrete uniform distribution over two points {1,2} and it is equivalent to $\boldsymbol{B}(h = 1\colon 0.5) \wedge \boldsymbol{B}(h = 2\colon 0.5)$.

- The belief distribution $\forall x.\boldsymbol{B}(h \leq x\colon f_l(x))$ expresses a continuous distribution where $f_l(x)$ expresses the CDF of a Logistic distribution with mean 75

[7] $\lim_{x\to\infty} f(x) = 1$ can be expressed as $\forall u.(u > 0 \supset \exists m.\forall v.(v > m) \supset |1 - f(v)| < u)$. There is a similar expression in our logic for multiple variables and for the limit when $\vec{x} \to -\infty$.

and scale 5.5:[8]

$$f_l(x) = \frac{1}{1 + e^{-(x-75)/5.5}} \tag{4.2}$$

Besides, it is also possible to express mixture distributions via our notion of belief distribution.

Example 4.6. Consider a belief distribution $\forall x.\boldsymbol{B}(h \leq x\colon f_m(x))$ of the nullary fluent h with

$$f_m(x) = \begin{cases} 0 & x < 0 \\ 0.2 & 0 \leq x \leq 1 \\ 0.1 + 0.1x & 1 < x \leq 9 \\ 1 & x > 9 \end{cases}$$

which expresses a mixture distribution, where the point $h = 0$ receives probability mass 0.2 and the rest 0.8 probability mass is distributed uniformly in the interval $[1, 9]$.

This is different from the formula $\boldsymbol{B}^{f_d} \vee \boldsymbol{B}^{f_c}$, which mixes a pure discrete distribution f_d with a pure continuous one f_c. Formulas like $\boldsymbol{B}^{f_d} \vee \boldsymbol{B}^{f_c}$ are possible in BL's formalism [BL18]; however, they cannot handle the real mixture distribution as our example above. The reason is that their notion of belief extends the BHL formalism, and belief is defined in terms of normalized sums or integral of weights of situations. Since it is unclear how to integrate over situations, they map situations to fluent values which are purely continuous, then belief is given in terms of integral over the continuous values of fluents. When the distribution is not purely continuous as in Example 4.6, where the fluent h has discrete probability at the point 0 and has continuous probability in the range $[1, 9]$, the integral over situations is not well-defined. Our formalism does not have such problems as we do not have explicit integrals over possible worlds.

[8] We use Logistic distributions, as they are approximations of normal distributions while their CDFs can be expressed in elementary functions.

Beyond Nullary Fluents

As mentioned earlier, another main advantage of our account is that the notion of belief (as well as only-believing) does not restrict possible worlds to only having finitely many nullary fluents. This, among other things, allows us to express probabilities involving potentially infinitely many random variables. For example:

Example 4.7. Let Σ be $\forall x.\neg Ath(x) \supset \forall y.\boldsymbol{B}(hr(x) \leq y\colon f_l(y))$ where $f_l(y)$ expresses the CDF of a Logistic distribution defined in Eq. (4.2). Formula Σ assigns a belief distribution of the heart rate for any non-athlete. For simplicity, we use $\boldsymbol{B}^{f_l}_x$ to denote $\forall y.\boldsymbol{B}(hr(x) \leq y\colon f_l(y))$. Then, we have

- Σ is satisfiable.

  Intuitively, Σ says that the heart rate of every non-athlete is subject to the same Logistic distribution. Clearly, there exists a probability space over possible worlds under which $hr(n)$ is distributed in this way where $Ath(n) = \bot$ holds. In the set of all such probability spaces, consider one $\langle \mathcal{W}, \mu \rangle$, which assigns non-zero probability only to worlds which satisfy

  $$\forall x, y.Ath(x) \equiv Ath(y) \supset hr(x) = hr(y) \tag{4.3}$$

  then $\{\langle \mathcal{W}, \mu \rangle\}$ is a desired epistemic state.

- $\boldsymbol{O}\Sigma$ is satisfiable.

  In fact, $\boldsymbol{O}\Sigma$ can be satisfied by multiple epistemic states: The simplest one is that $e \models \boldsymbol{O}(\forall x.Ath(x))$. For e' a model of $\boldsymbol{O}\Sigma$ and there is an $s = \langle \mathcal{W}, \mu \rangle$ such that $s \in e'$ and $\mu(W_{\neg Ath(Tom)}) > 0$, then by the Semantics $e' \models \boldsymbol{B}^{f_l}_{Tom}$ must hold, otherwise $\mu(W^{e'+s}_{\neg Ath(Tom) \supset \boldsymbol{B}^{F}_{Tom}}) \neq 1$ and therefore $e' \not\models \boldsymbol{O}\Sigma$. Thus we have the result as follows:

- $\boldsymbol{O}(\Sigma \wedge \neg Ath(Tom)) \models \forall y \boldsymbol{B}(hr(Tom) \leq y\colon f_l(y))$

- $\boldsymbol{O}(\Sigma \wedge Ath(Alice)) \models \neg\exists x.\boldsymbol{B}(hr(Alice) \leq 70\colon x)$

  Since $Ath(Alice)$ holds, $\mu(W^{e}_{\neg Ath(Alice) \supset \boldsymbol{B}^{f_l}_{Alice}}) = 1$ holds for any e. There-

fore, if e is a model of the premise, it includes all distributions of the athlete Alice's heart rate and hence the agent has no knowledge of the distribution.

Autoepistemic Reasoning about Distributions

The example in the last section already has a vibe of default reasoning about probability distributions. As described by Reiter [Rei80], a default rule should have the form $\alpha\colon \beta/\gamma$ and have an intuitive reading as "if α is believed and it is consistent to believe β, it infers that γ is true". For objective defaults, the consistency of β is interpreted as $\neg\boldsymbol{K}\neg\beta$ in Moore's autoepistemic logic (AEL) [Moo85]. To begin with, we show that our account preserves AEL's result on the flying bird example:

Example 4.8 ([LL01])**.** Assuming that $\mathsf{KB} = \{Bird(tweety)\}$ and

$$\delta = \forall x.\big(Bird(x) \land \neg\boldsymbol{K}\neg Fly(x)\big) \supset Fly(x) \tag{4.4}$$

then the following sentences are valid:

1. $\boldsymbol{O}(\mathsf{KB} \land \neg Fly(tweety) \land \delta) \supset \boldsymbol{K}\neg Fly(tweety)$
2. $\boldsymbol{O}(\mathsf{KB} \land Fly(tweety) \land \delta) \supset \boldsymbol{K}Fly(tweety)$
3. $\boldsymbol{O}(\mathsf{KB} \land \delta) \supset \boldsymbol{K}Fly(tweety)$

Proof. (1) and (2) are trivial since $\boldsymbol{O}\alpha \supset \boldsymbol{K}\alpha$ is valid. To prove (3), consider any state e such that $e \models \boldsymbol{O}(\mathsf{KB} \land \delta)$, we show that $e \models \neg\boldsymbol{K}\neg Fly(tweety)$. Let $W' = \{w \mid w \models \mathsf{KB} \land \forall x.Fly(x)\}$. We consider a space $s = \langle \mathcal{W}, \mu\rangle$ such that $\mu(W') = 1$. Since $\forall x.Fly(x) \models \delta$, $W' \subseteq W^{e+s}_{\mathsf{KB}\land\delta}$ and therefore $\mu(W^{e+s}_{\mathsf{KB}\land\delta}) \geq \mu(W') = 1$. By the semantics for $\boldsymbol{O}$, $s \in e$. Hence $e \models \neg\boldsymbol{K}\neg Fly(tweety)$ as $\mu(W_{\neg Fly(tweety)}) = 0$.
Now we prove that $e \models \boldsymbol{K}Fly(tweety)$. For all $\langle \mathcal{W}', \mu'\rangle \in e$, by the semantics, $\mu'(W^e_{\mathsf{KB}\land\delta}) = 1$. Since $e \models \neg\boldsymbol{K}\neg Fly(tweety)$,

$$\begin{aligned} W^e_{\mathsf{KB}\land\delta} =& \{w \mid e, w \models Bird(tweety) \land \forall x.(Bird(x) \land \neg\boldsymbol{K}\neg Fly(x) \supset Fly(x))\} \\ =& \{w \mid e, w \models Bird(tweety) \land \delta \land Fly(tweety)\} \\ \subseteq& W_{Fly(tweety)} \end{aligned}$$

Thus $\mu'(W_{Fly(tweety)}) = 1$ and hencefore $e \models \boldsymbol{K}Fly(tweety)$. □

However, for probabilistic defaults, consistency cannot be interpreted as "not know not". For instance, given a belief distribution $\boldsymbol{B}^f := \forall x.\boldsymbol{B}(t \leq x\colon f(x))$, $\neg\boldsymbol{K}\neg\boldsymbol{B}^f$ is equivalent to $\boldsymbol{B}^f$ itself due to the property of (negative) introspection. Instead, we express the consistency of probabilistic beliefs by introducing the following modality:

- $e, w, \lambda \models \boldsymbol{M}(\alpha\colon r)$ iff there exists $s \in e$ such that $s = \langle \mathcal{W}, \mu\rangle$, $\mu(W^s_\alpha) = \|r\|_\lambda$.

i.e. e contains an element s, which assigns degree $\|r\|$ to worlds which satisfy α. We write $\boldsymbol{M}\alpha$ when $r = 1$. We extend Example 4.7 to show how our account can handle autoepistemic reasoning about probabilistic distributions.

Example 4.9 (Heart Rate Distribution)**.** We define a default theory D with a single rule $\forall x.\big(\boldsymbol{K}(\neg Ath(x)) \wedge \boldsymbol{M}^{f_l}_x\big) \supset \boldsymbol{B}^{f_l}_x$, where f_l and $\boldsymbol{B}^{f_l}_x$ are defined as in Example 4.7 and $\boldsymbol{M}^{f_l}_x$ means $\boldsymbol{M}(\boldsymbol{B}^{f_l}_x\colon 1)$.
Let $F_1 = \{\boldsymbol{B}(hr(Tom) \leq 75\colon 0.6)\}$, $F_2 = \{\boldsymbol{B}(hr(Tom) \leq 75\colon 0.5)\}$ and $\mathsf{KB} = \{\neg Ath(Tom)\}$. Besides, we consider $\phi := \forall y \exists z.\boldsymbol{B}(hr(Tom) \leq y\colon z)$, i.e. there is a belief of the complete distribution of Tom's heart rate. Then the following sentences are valid:

1. $\boldsymbol{O}(\mathsf{KB} \wedge F_1 \wedge D) \wedge \phi \models \boldsymbol{B}(hr(Tom) \leq 75\colon 0.6)$
2. $\boldsymbol{O}(\mathsf{KB} \wedge F_2 \wedge D) \wedge \phi \models \boldsymbol{B}(hr(Tom) \leq 75\colon 0.5)$
3. $\boldsymbol{O}(\mathsf{KB} \wedge D) \wedge \phi \models \boldsymbol{B}(hr(Tom) \leq 75\colon 0.5)$

Analogous to the objective flying bird example, (1) and (2) are trivial.
To show (3), for any e such that $e \models \boldsymbol{O}(\mathsf{KB} \cup D) \wedge \phi$, let $s = \langle \mathcal{W}, \mu\rangle$ be a probability space. By the semantics, $s \in e$ iff $\mu(W^{e+s}_{\mathsf{KB}\cup D}) = 1$. Since $e \models \boldsymbol{K}\mathsf{KB}$ and D is subjective. $s \in e$ iff $\mu(W_{\mathsf{KB}}) = 1$ and $e + s \models \forall x.\big(\boldsymbol{K}(\neg Ath(x)) \wedge \boldsymbol{M}^{f_l}_x\big) \supset \boldsymbol{B}^{f_l}_x$.
We fix x to be Tom. Since $e \models \boldsymbol{K}\mathsf{KB}$, $e \models \boldsymbol{M}^{f_l}_{Tom} \supset \boldsymbol{B}^{f_l}_{Tom}$ must hold.
We prove that $e \models \boldsymbol{B}^{f_l}_{Tom}$. Assuming that $e \models \neg\boldsymbol{B}^{f_l}_{Tom}$, then $e \models \neg\boldsymbol{M}^{f_l}_{Tom}$ most

hold. By the semantics, $e+s \models \boldsymbol{M}^{fl}_{Tom}$ iff $\{s\} \models \boldsymbol{M}^{fl}_{Tom}$. Thus, for any $s = \langle \mathcal{W}, \mu \rangle$, $s \in e$ if and only if:

$$\mu(W_{\mathsf{KB}}) = 1,\ \{s\} \models \neg \boldsymbol{M}^{fl}_{Tom} \text{ and } e + s \models D^x_n \text{ for all } n \in \mathcal{N}, n \neq Tom. \tag{4.5}$$

We can construct two spaces s_1, s_2 such that $\{s_1\} \models \boldsymbol{B}(hr(Tom) \leq 75 \colon 0.6)$, $\{s_2\} \models \boldsymbol{B}(hr(Tom) \leq 75 \colon 0.7)$, and they both satisfy Eq.(4.5). Namely, $s_1, s_2 \in e$. Thus $e \models \neg\phi$ and it contradicts the assumption that $e \models \phi$.

We define $\boldsymbol{M}'\alpha \doteq \neg \boldsymbol{K} \neg \alpha$, i.e. $\boldsymbol{M}'$ is the modality for consistency in AEL. For objective α, we have the following result:

- $\models \boldsymbol{M}\alpha \supset \boldsymbol{M}'\alpha$

  Proof. Suppose that $e \models \boldsymbol{M}\alpha$, by the semantics, there is $\langle \mathcal{W}, \mu \rangle \in e$ such that $\mu(W_\alpha) = 1$, i.e. $\mu(W_{\neg\alpha}) = 0 \neq 1$. Thus $e \models \neg \boldsymbol{K} \neg \alpha$. □

However, the converse result does not necessarily hold, i.e. $\boldsymbol{M}'\alpha \supset \boldsymbol{M}\alpha$ might not be valid. When $e \models \neg \boldsymbol{K} \neg \alpha$, there is $\langle \mathcal{W}, \mu \rangle \in e$ such that $\mu(W_{\neg\alpha}) \neq 1$, i.e. $\mu(W_\alpha) \neq 0$. While $\boldsymbol{M}\alpha$ requires an $s \in e$ such that it assigns degree 1 to W_α.

To show that $\boldsymbol{M}$ is indeed an extension of $\boldsymbol{M}'$, we reconsider Example 4.8 by substituting $\neg \boldsymbol{K} \neg$ in δ by $\boldsymbol{M}$, i.e.

$$\delta' = \forall x. Bird(x) \wedge \boldsymbol{M} Fly(x) \supset Fly(x),$$

then the following sentences are valid:

1. $\boldsymbol{O}(\mathsf{KB} \wedge \neg Fly(tweety) \wedge \delta') \supset \boldsymbol{K} \neg Fly(tweety)$
2. $\boldsymbol{O}(\mathsf{KB} \wedge Fly(tweety) \wedge \delta') \supset \boldsymbol{K} Fly(tweety)$
3. $\boldsymbol{O}(\mathsf{KB} \wedge \delta') \supset \boldsymbol{K} Fly(tweety)$

A proof can be constructed similarly.

Proof. Let $e \models \boldsymbol{O}(\mathsf{KB} \wedge \delta')$, we show that $e \models \boldsymbol{M}Fly(tweety)$.
Let $W' = \{w \mid w \models \mathsf{KB} \wedge \forall x.Fly(x)\}$. We consider a space $s = \langle \mathcal{W}, \mu \rangle$ such that $\mu(W') = 1$. Since $W' \subseteq W^{e+s}_{\mathsf{KB}\wedge\delta'}$, $\mu(W^{e+s}_{\mathsf{KB}\wedge\delta'}) \geq \mu(W') = 1$. By the semantics for $\boldsymbol{O}$, $s \in e$. Thus $e \models \boldsymbol{M}Fly(tweety)$ as $\mu(W_{Fly(tweety)}) = 1$. Now we prove that $e \models \boldsymbol{K}Fly(tweety)$. For all $\langle \mathcal{W}', \mu' \rangle \in e$, by the semantics, $\mu'(W^{e}_{\mathsf{KB}\wedge\delta'}) = 1$. Since $e \models \boldsymbol{M}Fly(tweety)$,

$$\begin{aligned} W^{e}_{\mathsf{KB}\wedge\delta'} =& \{w \mid e, w \models Bird(tweety) \wedge \forall x.(Bird(x) \wedge \boldsymbol{M}Fly(x) \supset Fly(x))\} \\ =& \{w \mid e, w \models Bird(tweety) \wedge \delta' \wedge Fly(tweety)\} \\ \subseteq& W_{Fly(tweety)} \end{aligned}$$

Thus $\mu'(W_{Fly(tweety)}) \geq \mu'(W^{e}_{\mathsf{KB}\wedge\delta'}) = 1$ and hencefore $e \models \boldsymbol{K}Fly(tweety)$. □

To end this section, it is worth reflecting at this point that the proposal seems to be the most general framework for reasoning about first-order formulas with probabilistic beliefs. These examples can not be expressed in the logics of BHL [BHL99a], BL [BL18], $\mathcal{DS}$ [BL17] and its variants [LL21, LF21] as in these works, the nullary fluent assumption is essential to ensure that belief distributions are indeed probability distributions over finite-dimensional Euclidean space $\mathbb{R}^k$. Therefore, a belief distribution that corresponds to probability over infinite dimensional space is not well-defined in these works. To the best of our knowledge, we are not aware of any work in the literature that handles this level of generality.

4.2. Reasoning about Actions by Belief Integration

By considering probability spaces on possible worlds, the last section introduced the logic $\mathcal{OBL}_c$ which provides a general treatment of probabilistic beliefs and meta-beliefs in the presence of discrete, continuous, and possibly mixed distributions. However, the logic is restricted to a static setting. To reason about belief distributions in the dynamical cases, where action effects can be stochastic or noisy, we investigate the extension of $\mathcal{OBL}_c$ to an account for actions.

In this section, we introduce the logic $\mathcal{PS}$, which is a second-order, many-sorted modal variant of the situation calculus and inherits many characteristics from prior work such as $\mathcal{OBL}_c$ and $\mathcal{DS}$.[9] Inspired by the Lebesgue integration in the field of measure theory and real analysis, we formally define *belief integration* on the probability spaces of possible worlds, and define the semantics of beliefs after actions based on such integral. We show how to reason about actions w.r.t. basic action theories.

4.2.1. Syntax

Similar to the logic $\mathcal{OBL}_c$, the vocabulary of $\mathcal{PS}$ consists of standard names, variables, and function symbols. Predicates are treated as abbreviations. The language has three sorts: object, number (as in $\mathcal{OBL}_c$) and *action*. We follow the same convention on notations that letters $n, m \ldots$ are used for standard names and $x, y \ldots$ for FO variables of sort object or number. Second-order functional variables are labeled as F, G, etc. We use $a, a' \ldots$ for variables of sort action.

We consider rigid function symbols of each arity, including a special binary symbol oi, which describes the *observational-indistinguishability* (alternative choices) among actions. We require that all functions of sort action are rigid. Fluent function symbols of each arity are included. Particularly, we reserve an unary symbol l for action likelikhood. Connectives $\wedge$, $\neg$, $\forall$ and modal operators $\boldsymbol{B}$, $\boldsymbol{O}$, $[\cdot]$, $\Box$ are

[9] We consider second-order language, as the belief after actions potentially involves infinite summation or integration, which is not FO-definable.

used to construct formulas.

Definition 4.10 ($\mathcal{PS}$ Term)**.** The terms of the language $\mathcal{PS}$ are the least set of expressions such that:

- every standard name and first-order variable is a term;
- If $t_1, ..., t_k$ are terms and f is a k-ary function symbol, then $f(t_1, ..., t_k)$ is a term of the same sort as f;
- If $t_1, ..., t_k$ are terms and F is a k-ary second-order variable, then $F(t_1, ..., t_k)$ is a term of the same sort as F.

A term is said to be *rigid*, if and only if it does not contain fluents. *Ground terms* are terms without variables and while *SO ground terms* are terms without FO variables.

Definition 4.11 ($\mathcal{PS}$ formulas)**.** The $\mathcal{PS}$ formulas are the minimal set such that:

- If t, t' are terms, then $t = t'$ is a formula;
- If α and β are formulas, x is a variable, and t is a term of sort action, then $\neg\alpha$, $\alpha \wedge \beta$, $\forall x.\alpha$, $\forall F.\alpha$, $[t]\alpha$ and $\Box\alpha$ are also formulas;
- If α is a formula, r is a rigid term of sort number, then $\boldsymbol{B}(\alpha\colon r)$ and $\boldsymbol{O}(\alpha\colon r)$ are also formulas.

$[t]\alpha$ should be read as "α holds after action t" and $\Box\alpha$ as "α holds after any sequence of actions". The epistemic expression $\boldsymbol{B}(\alpha\colon r)$ should be read as "α is believed with a degree r". $\boldsymbol{K}\alpha$ means "α is known" and is an abbreviation for $\boldsymbol{B}(\alpha\colon 1)$. $\boldsymbol{O}(\alpha\colon r)$ is read as "all that believed is α with degree r". Similarly, $\boldsymbol{O}\alpha$ means "α is only known" and is an abbreviation for $\boldsymbol{O}(\alpha\colon 1)$. For action sequence $z = a_1 \cdots a_k$, we write $[z]\alpha$ to mean $[a_1] \cdots [a_k]\alpha$. As usual, we treat $\alpha \vee \beta$, $\alpha \supset \beta$, $\alpha \equiv \beta$, and $\exists v.\alpha$ as abbreviations.

A *sentence* is a formula without free variables. We define TRUE and FALSE identically as in $\mathcal{OBL}_c$. Namely, TRUE is an abbreviation for $\forall x(x = x)$, and FALSE is

its negation. A formula with no $\Box$ or $[t_a]$ is called *static*. A formula with no $\boldsymbol{B}$ or $\boldsymbol{O}$ is called *objective*. A formula with no fluent, $\Box$ or $[t_a]$ outside $\boldsymbol{B}$ or $\boldsymbol{O}$ is called *subjective*. A formula with no $\boldsymbol{B}$, $\boldsymbol{O}$, $\Box$, $[t_a]$, l, and oi is called a *fluent formula*. A fluent formula without fluent functions is called a *rigid formula*.

4.2.2. Semantics of Static Beliefs

The semantics is given in terms of possible worlds, where a *world* is a tree-like Tarski structure. Formally, we assume a fixed domain of discourse $\mathcal{D} = \mathcal{D}_{obj} \cup \mathcal{D}_{act} \cup \mathbb{R}$ where $\mathcal{D}_{obj}$ is a countable infinite set of objects, $\mathbb{R}$ the real numbers, and $\mathcal{D}_{act}$ an uncountable infinite set of actions. The set of standard names $\mathcal{N} = \mathcal{N}_{obj} \cup \mathcal{N}_{act} \cup \mathcal{N}_{num}$ is a countable subset of $\mathcal{D}$.[10] We require that $\mathcal{N}_{obj} = \mathcal{D}_{obj}$, i.e. object constants are all named. Additionally, we assume elements in $\mathcal{D}_{act}$ are of the form $f_{act}(d_1, d_2, \ldots, d_k)$ where f_{act} is an action function symbol and $d_i \in \mathcal{D}_{obj} \cup \mathbb{R}$. Lastly, we fix $\mathcal{N}_{num}$ to the set of *computable numbers* [Tur36], which is a countable subset of $\mathbb{R}$ but still includes important irrational numbers such as π, e. We denote the discrete sub-domain $\mathcal{N}_{obj} \cup \mathcal{N}_{num}$ as $\mathbb{D}$. Let Z refer to any finite *action sequences*, including the empty sequence $\langle\rangle$, i.e. $Z = (\mathcal{D}_{act})^*$. Formally, *worlds* in $\mathcal{PS}$ are defined as follows:

Definition 4.12 ($\mathcal{PS}$ World)**.** A world w is a mapping from all function symbols and Z to functions of the corresponding sorts, such that w satisfies

1. for all k-ary object (likewise for action and number) function symbol f_{obj} and action sequence $z \in Z$, $w[f_{obj}, z] \colon (\mathcal{D})^k \mapsto \mathcal{D}_{obj}$;

2. Rigidity: if f is rigid, then for all $(w, z), (w', z')$, $w[f, z] = w'[f, z']$;

3. Arithmetical correctness: elementary function symbols (e.g. $+, \times, e^x$) are rigid and interpreted in the usual sense. For example, $w[+, z](1, 1) = 2$ for any w and z.

[10] Even if the domain is uncountable, only a countable subset of it can have standard names.

Let W be the set of all such worlds. A **variable map** λ maps each FO variable $v \in \mathcal{V}_{FO}$ to an element in $\mathcal{D}$ of the right sort, and maps each SO variable in $\mathcal{V}_{SO}$ to a function of the same sort. We write $\lambda \sim_v \lambda'$ to mean λ and λ' agree except perhaps on variable v. The denotation of terms is defined recursively:

Definition 4.13 ($\mathcal{PS}$ Denotation)**.** The denotation of a term t under a tuple (w, z, λ) is given as (assuming t, t_i are terms):

- $\|t\|^z_{w,\lambda} = t$ if $t \in \mathcal{N}$;
- $\|t\|^z_{w,\lambda} = \lambda(t)$ if t is a first-order variable, i.e. $t \in \mathcal{V}_{FO}$;
- $\|t\|^z_{w,\lambda} = w[f, z](\|t_1\|^z_{w,\lambda}, ..., \|t_k\|^z_{w,\lambda})$ if t is of the form $f(t_1, ..., t_k)$ where f is a function symbol.
- $\|t\|^z_{w,\lambda} = \lambda(F)(\|t_1\|^z_{w,\lambda}, ..., \|t_k\|^z_{w,\lambda})$ if t is of the form $F(t_1, ..., t_k)$ where F is a SO variable.

For simplicity, we write $\|t\|_\lambda$ when t is rigid, $\|t\|^z_w$ when t does not contain variables, and $\|t\|$ when t is both rigid and ground.

Epistemic States in $\mathcal{PS}$ are defined based on probability spaces on W. We refer to Section 3.2 for an introduction to the relevant mathematical concepts. An epistemic state e is defined as a set of pairs of the form $\langle \mathcal{W}, \mu \rangle$ such that $\langle W, \mathcal{W}, \mu \rangle$ forms a complete probability space, where $\mathcal{W}$, a σ-algebra on W, is the domain of μ. Since the set of worlds W is fixed, we directly call such a pair a probability space (or a distribution), and denote it as $s = \langle \mathcal{W}_s, \mu_s \rangle$. By a model, we mean a 4-tuple (e, w, z, λ) with e an epistemic state, w a world, z an action sequence, and λ a variable map. For objective formulas and static subjective formulas, the satisfaction relation is defined as follows:

- $e, w, z, \lambda \models t_1 = t_2$ iff $\|t_1\|^z_{w,\lambda}$ and $\|t_2\|^z_{w,\lambda}$ are identical;
- $e, w, z, \lambda \models \neg\alpha$ iff $e, w, z, \lambda \not\models \alpha$;
- $e, w, z, \lambda \models \alpha \wedge \beta$ iff $e, w, z, \lambda \models \alpha$ and $e, w, z, \lambda \models \beta$;

- $e, w, z, \lambda \models \forall v.\alpha$ iff $e, w, z, \lambda' \models \alpha$ for all $\lambda' \sim_v \lambda$;
- $e, w, z, \lambda \models \forall F.\alpha$ iff $e, w, z, \lambda' \models \alpha$ for all $\lambda' \sim_F \lambda$;
- $e, w, z, \lambda \models [t_a]\alpha$ iff $e, w, z \cdot n, \lambda \models \alpha$ and $n = \|t_a\|^z_{w,\lambda}$;
- $e, w, z, \lambda \models \Box\alpha$ iff $e, w, z \cdot z', \lambda \models \alpha$ for all $z' \in Z$;
- $e, w, \langle\rangle, \lambda \models \boldsymbol{B}(\alpha : r)$ iff for all $s = \langle \mathcal{W}, \mu \rangle$, $s \in e$ implies $W^{e,\lambda}_\alpha \in \mathcal{W}$ and $\mu(W^{e,\lambda}_\alpha) = \|r\|_\lambda$;

where $W^{e,\lambda}_\alpha = \{w' \mid e, w', \langle\rangle, \lambda \models \alpha\}$. Essentially, $e, w, \langle\rangle, \lambda \models \boldsymbol{B}(\alpha : r)$ if for all the probability spaces $\langle \mathcal{W}, \mu \rangle \in e$, the set of worlds that satisfy α under e, i.e. $W^{e,\lambda}_\alpha$ is an event in $\mathcal{W}$ and has probability measure $\|r\|_\lambda$. Clearly, beliefs $\boldsymbol{B}(\alpha : r)$ defined in this way are indeed probabilities over possible worlds.

Similar to $\mathcal{OBL}_c$ in Section 4.1, we consider the add-one-in semantics for $\boldsymbol{O}$-formulae:

- $e, w, \langle\rangle, \lambda \models \boldsymbol{O}(\alpha : r)$ iff for all $s = \langle \mathcal{W}, \mu \rangle$, $s \in e$ iff $W^{e+s,\lambda}_\alpha \in \mathcal{W}$ and $\mu(W^{e+s,\lambda}_\alpha) = \|r\|_\lambda$.

where $e + s$ stands for $e \cup \{s\}$, and $W^{e+s,\lambda}_\alpha = \{w' \mid (e + s), w', \langle\rangle, \lambda \models \alpha\}$.

4.2.3. Semantics after Actions

In this subsection, we provide the mathematical foundations for the semantics of probabilistic beliefs after actions, particularly, when *stochastic actions* and *noisy sensing* are involved. Defining beliefs after (possibly noisy) actions is more involved than the static cases. There are three questions to be addressed:

1. how to specify the non-deterministic effects of stochastic actions?
2. how to specify the likelihood of stochastic actions or sensing?
3. how to integrate action likelihoods and world probabilities into beliefs?

The first question can be solved in the same way as in BHL [BHL99b]: Instead of saying that a stochastic action has non-deterministic effects, we view the stochastic action as a set of ground actions (mutual *alternatives*) that are observationally indistinguishable to the agent and each has a deterministic effect. This is done via the special rigid function symbol oi.[11] For example, let $fwd(x, y)$ denote an action that the agent intends to move forward a distance x but actually moves a distance y. Then $oi(fwd(1, 1), fwd(1, 0))$ states that the agent cannot distinguish a successful forward $fwd(1, 1)$ and a failed one $fwd(1, 0)$.

The solution to the second question is also inspired by the BHL[BHL99b] and the BL formalism[BL18]. To begin with, one would wish to express that some effects are more likely than others. This can be achieved by specifing the likelihood of an action a using oi together with a special fluent l. However, the problem is how to ensure l is indeed a probability among potentially uncountably many action alternatives for stochastic actions. Recall that the probability space consists of three ingredients: sample space, event set, and probability measures. It is challenging to use the single fluent l to express such a rich structure. Our solution is to project the likelihood of action into an Euclidean space. Specifically, we impose the following restriction:

1. there are finitely many action types, which are function symbols. Note that this does not imply a finite number of actions, as the parameters of actions can vary over an infinite domain, even the uncountable domain $\mathbb{R}$.

2. stochastic actions have at most 3 parameters, namely, they are of the form $sa(x, m, y)$ where x is *controllable* and *observable* while m and y are *uncontrollable* and *unobservable* to the agent. Additionally, m ranges over the discrete domain $\mathbb{D}$, and y ranges over the continuous domain $\mathbb{R}$. A formalism that allows more parameters such as $sa(x_1, x_2, m, y)$ or $sa(x, m, y_1, \ldots, y_k)$ is certainly possible, but we restrict to three parameters for simplicity of

[11] We set oi to be rigid. A fluent oi would cause counter-intuitive result, see [LL21]

presentation. Nevertheless, actions may contain less than 3 parameters, like the action $fwd(x, y)$ omits the uncontrollable discrete parameter m;

3. sensing actions do not have unobservable parameters m or y but only the observable parameters x, i.e. they are of the form $sen(x)$;

4. the alternative relation oi holds among two actions if and only if they have the same action symbol and the same controllable parameter.

Formally, the restriction above amounts to having the following sentences as a background theory:

$$\Box\big(\bigvee_i \exists x, m, y.\ a = sa_i(x, m, y)\big) \vee \big(\bigvee_j \exists x.a = sen_j(x)\big) \tag{4.6}$$

$$\begin{aligned}\Box oi(a, a') = \top \equiv &\bigvee_i \exists x, m, y, m', y'.a = sa_i(x, m, y) \wedge a' = sa_i(x, m', y') \\ &\bigvee_j \exists x.a = sen_j(x) \wedge a' = sen_j(x)\end{aligned} \tag{4.7}$$

The solution to the third question is the most involved. Before we develop the mathematics, we first recap the solution in the logic $\mathcal{DS}$ and its variants [BL17, LL21, LF21], where both the degree of beliefs and action likelihood are discrete:

In the logic $\mathcal{DS}$, epistemic states are defined based on discrete probabilistic distribution, where a distribution d assigns non-negative values to each worlds. A special fluent l specifies the likelihood of each action in a world. Essentially, the degree of belief of an objective formula α after an action a is defined as the normalized weight of worlds and alternatives actions which satisfies α. Namely,

$$\sum_{(w,\hat{a})\in W_\alpha} p(w) * l(w, \hat{a}) \tag{4.8}$$

where $p(w)$ is the degree of belief assigned to world w, $l(w, \hat{a})$ is the likelihood of action $\hat{a}$ under world w, and

$$W_\alpha = \{(w, \hat{a}) \,|\, \hat{a} \text{ s an alternative of } a, \text{ and } (w, \hat{a}) \text{ satisfies } \alpha\}$$

Namely, W_α is the set of world-action pairs which satisfy α and the action is an alternative of a. To lift the restriction on discreteness, degrees of belief after actions in $\mathcal{PS}$ should be defined as an integral. Ideally, it should have the form

$$\int_{(w,\hat{a})\in W_\alpha} p(w) * l(w,\hat{a})\ d\hat{a}dw \tag{4.9}$$

Apparently, the quasi-integral above is not well-defined: In the logic $\mathcal{PS}$, beliefs are assigned via probability spaces of the form $\langle \mathcal{W}, \mu \rangle$, which may not include singleton event $\{w\}$, and therefore $p(w)$ may not be defined. Additionally, both the sets of worlds and actions are complex sets, which are not - and in general are not isomorphic to - some Euclidean spaces $\mathbb{R}^k$. Therefore, integration in real analysis, including Riemann's and Lebesgue's, cannot be applied directly. We address the third question by constructing a rigorous integration method over the space of possible worlds. The "**belief integration**" is constructed via two steps:

1. For a fixed world w, we integrate the likelihood of alternative actions or action sequences.

2. Consider action likelihood as a function on worlds, i.e. $\nu\colon W \to \mathbb{R}$, we define the integral of the function w.r.t. a probabilistic measure.

For a sensing action $sen(x)$, since it has no observationally-indistinguishable alternatives, integration is not involved. To ensure that l indeed specifies a probability among potentially uncountably many alternatives for stochastic actions, we introduce the notion of *propriety*:[12]

Definition 4.14 (Local Propriety)**.** A world $w \in W$ is *locoally proper* if and only if for every stochastic aciton $a = sa(x, m, y)$ and action sequence $z \in Z$, $w[l, z](a) \geq 0$, and

$$\sum_{m'\in\mathbb{D}} \int_{y'\in\mathbb{R}} w[l, z](sa(x, m', y'))dy' = 1. \tag{4.10}$$

[12]More precisely, here we introduce the notion of local propriety. The *global* version will be introduced later when sequences of actions are considered.

Implicitly, Eq. (4.10) requires that the fluent l has to be a *Lebesgue integrable* function in terms of parameters y, i.e. for any $r \in \mathbb{R}$,

$$\{y' \in \mathbb{R} \,|\, w[l, z](sa(x, m', y')) < r\}$$

is a Lebesgue measurable set (See Section 3.2 for the definition). Essentially, a world w is proper if and only if the fluent l indeed specifies a probability distribution over alternatives for stochastic actions.

In the rest of the chapter, whenever we mention a world, we mean a proper one. Meanwhile, by an epistemic state, we mean a set of tuples of the form $\langle \mathcal{W}_p, \mu \rangle$ such that $\langle W_p, \mathcal{W}_p, \mu \rangle$ forms a probability space where W_p is the set of proper worlds, $\mathcal{W}_p$ is the event set, which forms a σ-algebra on W_p, and μ is a probability measure.

Truth of beliefs after stochastic actions

Given a stochastic action $a = sa(x, m, y)$ and a set of its alternatives $A' \subseteq A_a := \{a' \,|\, oi(a, a') = \top\}$, let

$$\mathbb{D}_{A'} = \{m' \in \mathbb{D} \,|\, \text{there exists } y' \in \mathbb{R} \text{ s.t. } sa(x, m', y') \in A'\}$$
$$\mathbb{R}_{A',m'} = \{y' \in \mathbb{R} \,|\, sa(x, m', y') \in A'\}.$$

Namely, $\mathbb{D}_{A'}$ is the set of all possible discrete values in A', and $\mathbb{R}_{A',m'}$ is the set of all possible real values when fixing m'. With these notions, we are able to build correspondence between the space of actions and the (Lebesgue) measurable sets:

Theorem 4.15. Given a proper world w and a stochastic action $a = sa(x, m, y)$, let $\mathcal{A}_a = \{A' \subseteq A_a \,|\, \text{for all } m' \in \mathbb{D}, \mathbb{R}_{A',m'} \text{ is Lebesgue measurable}\}$. For every action sequence $z \in Z$, we define $\nu^{w,z} \colon \mathcal{A}_a \mapsto \mathbb{R}^{\geq 0}$ as

$$\nu^{w,z}(A') = \sum_{m' \in \mathbb{D}_{A'}} \int_{y' \in \mathbb{R}_{A',m'}} w[l, z](sa(x, m', y'))dy' \tag{4.11}$$

then $\langle A_a, \mathcal{A}_a, \nu^{w,z} \rangle$ forms a probability space. We call a set of actions $A' \subseteq A_a$ *measurable* if $A' \in \mathcal{A}_a$.

Proof. $\mathcal{A}_a$ is a σ-algebra on A_a, as each element in $\mathcal{A}_a$ is an union of countably many Lebesgue measurable sets, and the set of Lebesgue measurable sets forms a σ-algebra. It suffices to prove that $\nu^{w,z}$ forms a probability measure:

- By the definition of *oi* and Eq. (4.10), $\nu^{w,z}(A_a) = 1$;
- Since $w[l,z](a) \geq 0$ for all $a \in A$, $\nu^{w,z}(A') \geq 0$ for all $A' \in \mathcal{A}_a$; and,
- for disjoint $\{A'_j\} \subseteq \mathcal{A}_a$, since the function to be integrated is non-negative,

$$
\begin{aligned}
&\nu^{w,z}(\bigcup_{j=1}^{\infty} A'_j) \\
&= \sum_{m' \in \mathbb{D}_{\bigcup A'_j}} \int_{y' \in \mathbb{R}_{\bigcup A'_j, m'}} w[l,z](sa(x,m',y'))dy' \\
&= \sum_{m' \in \mathbb{D}_{\bigcup A'_j}} \sum_{j=1}^{\infty} \int_{y' \in \mathbb{R}_{A'_j, m'}} w[l,z](sa(x,m',y'))dy' \qquad (\#) \\
&= \sum_{j=1}^{\infty} \Big(\sum_{m' \in \mathbb{D}_{\bigcup A'_j}} \int_{y' \in \mathbb{R}_{A'_j, m'}} w[l,z](sa(x,m',y'))dy' \Big) \\
&= \sum_{j=1}^{\infty} \Big(\sum_{m' \in \mathbb{D}_{A'_j}} \int_{y' \in \mathbb{R}_{A'_j, m'}} w[l,z](sa(x,m',y'))dy' \Big) \qquad (*) \\
&= \sum_{j=1}^{\infty} \nu^{w,z}(A'_j)
\end{aligned}
$$

To explain step (#): By Def. 4.14, $w[l,z](sa(x,m',y'))$ is Lebesgue integrable w.r.t. y'. Since $\{A'_j\}$ are disjoint, for a fixed m', $\mathbb{R}_{A'_j,m'}$ are disjoint for different j. As for $(*)$, by the definition of $\mathbb{R}_{A'_j,m'}$, for any $m' \notin \mathbb{D}_{A'_j}$, $\mathbb{R}_{A'_j,m'} = \varnothing$. □

The probability defined above can be extended from a single action to sequences:

Definition 4.16. We define:

1. $l^* : W \times Z \to \mathbb{R}^{\geq 0}$ as
   a) $l^*(w, \langle\rangle) = 1$, for every $w \in W$;
   b) $l^*(w, z \cdot a) = l^*(w,z) \times w[l,z](a)$.

2. $z \sim z'$ as

   a) $\langle\rangle \sim z'$ iff $z' = \langle\rangle$;

   b) $z \cdot a \sim z'$ iff $z' = z^* \cdot a^*, z \sim z^*$, and $oi(a, a^*) = \top$.

Let $\bigodot$ be the concatenation of actions. Given stochastic action sequence $z = \bigodot_{i=1}^{k} sa_i(x_i, m_i, y_i)$, we denote

- the set of alternatives of z as $Z_z = \{z' \,|\, z \sim z'\}$;
- the set of all possible discrete values in the subset Z' of z's alternatives as

$$\mathbb{D}_{Z'} = \{\vec{m} \,|\, \text{it exists } \vec{y}' \in \mathbb{R}^k \text{ s.t. } \bigodot_{i=1}^{k} a_i(x_i, m_i', y_i') \in Z' \subseteq Z_z\}$$

  where $\vec{m}' = (m_1', \ldots, m_k')$ and $\vec{y}' = (y_1', \ldots, y_k')$;

- the set of all possible real values of $\vec{y}'$ in Z' given $\vec{m}'$ as
  $\mathbb{R}_{Z',\vec{m}'} = \{\vec{y}' \in \mathbb{R}^k \mid \bigodot_{i=1}^{k} sa_i(x_i, m_i', y_i') \in Z' \subseteq Z_z\}$.

Similar to Theorem 4.15, we have:

Theorem 4.17. Given a proper world w and a sequence of stochastic actions $z = a_1 \cdots a_k$ where $a_i = sa_i(x_i, m_i, y_i)$. Let $\mathcal{Z}_z = \{Z' \subseteq Z_z \,|\, \text{for all } \vec{m}' \in \mathbb{D}^k, \mathbb{R}_{Z',\vec{m}'} \text{ is Lebesgue measurable}\}$, we define $\nu_{l^*}^w \colon \mathcal{Z}_z \mapsto \mathbb{R}^{\geq 0}$ as:

$$\nu_{l^*}^w(Z') = \sum_{\vec{m}' \in \mathbb{D}_{Z'}} \int_{\vec{y}' \in \mathbb{R}_{Z',\vec{m}'}} l^*(w, \bigodot_{i=1}^{k} sa_i'(x, m_i', y_i')) d\vec{y}' \tag{4.12}$$

then $\langle Z_z, \mathcal{Z}_z, \nu_{l^*}^w \rangle$ forms a probability space.

The proof is similar to the proof of Theorem 4.15. We call a set of action sequences $Z' \subseteq Z_z$ *measurable* if $Z' \in \mathcal{Z}_z$.

A remark is that the definition of $\nu^{w,z}$ in Theorem 4.15 requires an action sequence z as a parameter while the definition of $\nu_{l^*}^w$ does not. In fact, $\nu_{l^*}^w$ is implicitly defined wrt the empty sequence $\langle\rangle$.

Now, the task remains to incorporate the well-defined likelihood of actions into the model of belief. For that, we need two additional notations: the likelihood

for an unmeasurable set of action sequences and the integral over possible worlds. We begin by introducing the *Vitali sets* to illustrate the necessity of considering unmeasurable sets:

Definition 4.18 (Vitali Set[Vit05])**.** A Vitali set is a subset V of the interval $[0, 1]$ of real numbers such that, for each $r \in \mathbb{R}$, there is exactly one $v \in V$ such that $(v - r) \in \mathbb{Q}$, i.e. $v - r$ is a rational number.

It has been proved that Vitali sets are not measurable: Assuming that there is a measurable Vitali set V with $m(V) = \delta$, for each rational $q \in [-1, 1] \cap \mathbb{Q}$, we define $V_q = \{v+q \,|\, v \in V\}$. Apparently, $V_q \subseteq [-1, 2]$, and for $v, v' \in V$ with $v \neq v'$, $v - v' \notin \mathbb{Q}$ since $v - v = 0 \in \mathbb{Q}$. Then for $q \neq q'$, V_q and $V_{q'}$ must be disjoint, otherwise there are $v, v' \in V$ such that $v + q = v' + q'$, i.e. $v - v' = q' - q \in \mathbb{Q}$. Then $[0, 1] \subseteq \bigcup_{q\in[-1,1]\cap\mathbb{Q}} V_q$, as for each $r \in [0, 1]$, there is $v \in V$ s.t. $v - r \in \mathbb{Q}$. Since $v - r \in [-1, 1]$, let $q = v - r$, then $r \in V_q$.

- If $\delta > 0$, then $m(\bigcup_{q\in[-1,1]\cap\mathbb{Q}} V_q) = \sum_{q\in[-1,1]\cap\mathbb{Q}} m(V_q) = \sum_{q\in[-1,1]\cap\mathbb{Q}} \delta = \infty$, contradicting that $\bigcup_{q\in[-1,1]\cap\mathbb{Q}} V_q \subseteq [-1, 2]$.
- If $\delta = 0$, $m(\bigcup_{q\in[-1,1]\cap\mathbb{Q}} V_q) = 0$, contradicting that $[0, 1] \subseteq \bigcup_{q\in[-1,1]\cap\mathbb{Q}} V_q$.

Given a proper world w, even if only considering a simple atom, there might be an unmeasurable set of stochastic action sequences z (mutually alternative) such that the atom holds. We construct an artificial example as follows:

Example 4.19. Let V be a Vitali set. We consider a stochastic action $move(y_0)$,[13] a nullary fluent h, and a world w such that

- $w[h, \langle\rangle] = 0$;
- For an alternative action $move(y')$ with $y' \in V$, $w[h, move(y')] = 1$;
- For an alternative action $move(y')$ with $y' \notin V$, $w[h, move(y')] = 0$;

[13] We omit the controllable variable x and discrete unobservable variable m for simplicity.

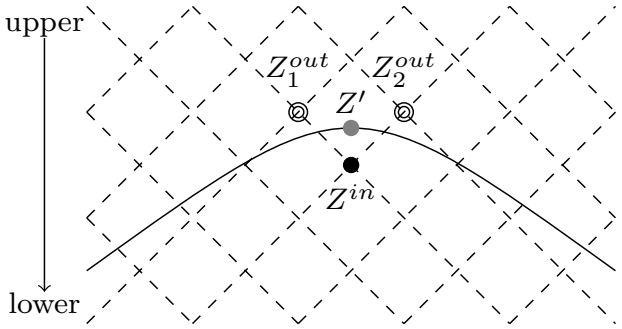


Figure 4.1.: Illustration of $\nu_{l*}^{w+}(Z')$ and $\nu_{l*}^{w-}(Z')$.

Then $A_{h=1} = \{a \,|\, a \sim move(y_0), w, a \models h = 1\}$ is an unmeasurable set.

Hence, inevitably, one needs to assign likelihoods to unmeasurable sets of action sequences. To do so, we borrow ideas from inner and outer Lebesgue measures. Formally,

Definition 4.20. Given a proper world w and a sequence of stochastic action z, let ν_{l*}^{w} be as in Theorem 4.17, for any $Z' \subset Z_z$, we define ν_{l*}^{w+} and ν_{l*}^{w-} as:

$$\nu_{l*}^{w+}(Z') = \inf\{\nu_{l*}^{w}(Z'') \,|\, Z'' \text{ measurable, } Z' \subseteq Z''\}$$
$$\nu_{l*}^{w-}(Z') = \sup\{\nu_{l*}^{w}(Z'') \,|\, Z^* \text{ measurable, } Z'' \subseteq Z'\}.$$

$\nu_{l*}^{w+}(Z')$ is called the *outer measure* of Z' while $\nu_{l*}^{w-}(Z')$ is called the *inner measure* of Z'. Fig. 4.1 provides a conceptual illustration of ν_{l*}^{w+} and ν_{l*}^{w-}. Dashed dots in the grid represent the measurable sets and the arrow on the left represents the direction of inclusion relations among sets. The set Z' (gray dot) is unmeasurable and its inner measure equals the measure of Z^{in} (black dot) while its outer measure equals the minimal of measures between Z_1^{out} and Z_2^{out} (hollow dots). A remark is that $\nu_{l*}^{w+}(Z') = \nu_{l*}^{w-}(Z')$ if and only if Z' is measurable.

The last ingredient is the notion of *integral over possible worlds*. Let $f_X(x) = 1$ *if* $x \in X$ *else* 0 be a *characteristic function* of the set X. Given a probability space $\langle X, \mathcal{X}, \mu \rangle$, the integral of a characteristic function of a measurable set $X' \subseteq X$ is defined in the standard way as: $\int_{x \in X} f_{X'}(x) d\mu = \mu(X')$. A

simple function $g(x) : X \mapsto \mathbb{R}$ over a probability space $\langle X, \mathcal{X}, \mu \rangle$ is a finite linear combination of characteristic functions of measurable sets in that probability space. Namely, $g(x) = \sum_i b_i f_{X_i}(x)$ for some characteristic functions $f_{X_i}(x)$ and X_i are measurable ($b_i \in \mathbb{R}$). The integral of simple function $g(x)$ is defined as $\int_{x \in X} g(x) d\mu = \sum_i b_i \mu(X_i)$. Given an epistemic state e, let $s \in e$ and $s = \langle \mathcal{W}, \mu \rangle$, we call a function $f : W_p \mapsto \mathbb{R}^{\geq 0}$ measurable on μ iff for all $b \in \mathbb{R}^{\geq 0}$, $\{w \in W_p \,|\, f(w) < b\} \in \mathcal{W}$.

Theorem 4.21. Given a probability space $\langle \mathcal{W}, \mu \rangle$, let $f : W \mapsto \mathbb{R}^{\geq 0}$ be a measurable function on $\mathcal{W}$. Then there exists a sequence of non-negative simple functions $\{g_0, g_1, \ldots g_n, \ldots\}$ such that for all $w \in W$ and $n \in \mathbb{N}$,

$$g_n(w) \leq g_{n+1}(w) \text{ and } \lim_{n \to \infty} g_n(w) = f(w)$$

The theorem suggests that any measurable function can be approximated as arbitrarily close by using a series of simple functions. Hence, we exploit the limit of the integral over such simple functions to define the integral over measurable functions. This is in the same spirit as the definition of Lebesgue integral [SS09]. [14]

Proof. W.l.o.g. we only consider functions f whose values are in $[0, 1]$. For any n, we define the simple function g_n as follows:

$$g_n(w) = \sum_{i=1}^{2^n} \frac{i-1}{2^n} f_{W_{n,i}}(w) \tag{4.13}$$

Where $f_{W_{n,i}}$ is the characteristic function of set $W_{n,i} = \{w \,|\, i/2^n \leq f(w) < (i+1)/2^n\}$, since f is measurable on μ, $W_{n,i}$ is measurable. Then $\{g_n\}_1^\infty$ will satisfy the required conditions. □

Intuitively, the theorem indicates that any non-negative measurable function is the limit of a sequence of simple functions. We exploit the limit of the simple

[14] The key to the proof is how to construct such $g_n(w)$. W.l.o.g. assume $f(w) \in [0, 1]$. For any n, let $f_{W_{n,i}}(w)$ be the characteristic function of the set $W_{n,i} = \{w \,|\, i/2^n \leq f(w) < (i+1)/2^n\}$. Then it is not hard to show $g_n(w) = \sum_{i=1}^{2^n} \frac{i-1}{2^n} f_{W_{n,i}}(w)$ is the desired simple function.

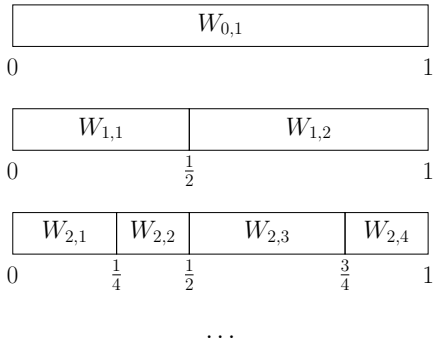


Figure 4.2.: The construction of simple functions in Theorem 4.21, where the length of blocks denotes the measure of each subset. For each block, the left and right subscripts denote lowest and highest values of $f(w)$ for worlds in this block.

function's integral to define the integral of measurable functions. For μ, f and g_n defined above, we define the integral of f as

$$\begin{aligned}\int_w f(w)d\mu &= \lim_{n\to\infty}\int_w g_n(w)d\mu \\ &= \lim_{n\to\infty}\sum_{i=1}^{2^n}\frac{i-1}{2^n}\mu(W_{n,i})\end{aligned}$$

$$\int_w f d\mu = \lim_{n\to\infty}\int_w g_n d\mu = \lim_{n\to\infty}\sum_{i=1}^{2^n}\frac{i-1}{2^n}\mu(W_{n,i}) \tag{4.14}$$

For any other monotonic sequence of simple functions converging to f, the limit of integral results in the same value.

Definition 4.22 (Integral over Possible Worlds)**.** The integral of a measurable function f on μ is defined as

$$\int_w f d\mu = \lim_{n\to\infty}\int_w g_n d\mu$$

where g_n are defined as in Theorem 4.21.

For any model (e, w, z, λ) and formula α, let $Z_\alpha^{e,w,z,\lambda} = \{z' \mid z' \sim z, e, w, z', \lambda \models \alpha\}$. If α is a sentence, we write $Z_{e,w,z}^{\alpha}$. If sentence α is objective or subjective, we write $Z_{w,z}^{\alpha}$ or $Z_{e,z}^{\alpha}$ respectively. Clearly, $Z_{\text{TRUE}}^{e,w,z,\lambda} = Z_z$ for all e, w, z, λ. Now we are ready to define the truth conditions of $\boldsymbol{B}$ and $\boldsymbol{O}$ after stochastic actions:

- $e, w, z, \lambda \models \boldsymbol{B}(\alpha : r)$ iff for all $s = \langle \mathcal{W}, \mu \rangle$, $s \in e$ implies
  $\int_w \nu_{l*}^{w+}(Z_\alpha^{e,w,z}) d\mu = \int_w \nu_{l*}^{w-}(Z_\alpha^{e,w,z}) d\mu = \|r\|_\lambda$;
- $e, w, z, \lambda \models \boldsymbol{O}(\alpha : r)$ iff for all $s = \langle \mathcal{W}, \mu \rangle$, $s \in e$ iff
  $\int_w \nu_{l*}^{w+}(Z_\alpha^{e+s,w,z}) d\mu = \int_w \nu_{l*}^{w-}(Z_\alpha^{e+s,w,z}) d\mu = \|r\|_\lambda$;

As in Section 4.1, $e + s$ means $e \cup \{s\}$. When the context is clear, we write Z_α or Z_α^e for $Z_\alpha^{e,w,z,\lambda}$. The semantics above imposes two implicit constraints:

1. the integral $\int_w \nu_{l*}^{w+}(Z_\alpha) d\mu$ and $\int_w \nu_{l*}^{w-}(Z_\alpha) d\mu$ requires that $\nu_{l*}^{w+}(Z_\alpha)$ and $\nu_{l*}^{w-}(Z_\alpha)$, interpreted as functions of w, have to be measurable;
2. the equal integrals of outer and inner measure of Z_α implies that the set $\{w \mid Z_\alpha^{e,w,z,\lambda}$ is unmeasurable$\}$ has zero-measure (i.e. zero-probability) under μ. This is crucial to ensure that $\boldsymbol{B}$ is indeed a well-defined probability (see Subsection 4.2.4).

Intuitively, given μ, $\nu_{l*}^{w+}(Z_\alpha)$ (resp. for ν_{l*}^{w-}) represents the superior (inferior) probability of α after z in the world w. Furthermore, the integrals, i.e. $\int_w \nu_{l*}^{w+}(Z_\alpha) d\mu$ and $\int_w \nu_{l*}^{w-}(Z_\alpha) d\mu$, are the probability of all worlds where α holds after z.

Truth of belief after sensing actions

The story for sensing actions is a bit different. To begin with, sensing outcomes are all observable to the agent. Hence, sensing actions do not have alternatives other than themselves, yet one would still wish to express that some outcomes are more likely than others. Besides, after observing some sensing outcomes, the possibilities of situations, which violate the sensing results, should be ruled out.

To address these problems, we extend the notion of propriety to the global cases:

Definition 4.23 (Global Propriety)**.** We call a world w *globally proper* if and only if for each action sequence $z = a_1 \cdots a_k$ where $a_i = sa_i(x_i, m_i, y_i)$ or $a_i = sen_i(x_i)$, it exists $r \in \mathbb{R}$ such that

$$\sum_{\vec{m}' \in \mathbb{D}^{k'}} \int_{\vec{y}' \in \mathbb{R}^{k'}} l^*(w, z') d\vec{y}' = r, \tag{4.15}$$

where $z' = a'_1 \cdots a'_k$. When $a_i = sa_i(x_i, m_i, y_i)$, then $a'_i = sa_i(x_i, m'_i, y'_i)$, when $a_i = sen_i(x_i)$ then $a'_i = a_i$. $\vec{m}'$ is the concatenation of all discrete (unobservable) variables m_i and $\vec{y}'$ be the concatenation of all continuous variables y_i.

Essentially, Eq. 4.15 ensures that the function $l^*(w, z)$ (z may contain sensing) like the one in Theorem 4.17 is integrable. With such a definition, we have a theorem that is similar to Theorem 4.17. Let the notion Z_z, $\mathbb{D}_{Z'}$, and $\mathbb{R}_{Z',\vec{y}'}$ be defined exactly the same as before but z could contain sensing of the form $sen_i(x)$, then we have:

Theorem 4.24. Given a proper world w and an action sequence $z = a_1 \cdots a_k$ where $a_i = sa_i(x_i, m_i, y_i)$ or $a_i = sen_i(x_i)$, let

$$\mathcal{Z}_z = \{Z' \subseteq Z_z \,|\, \text{for all } \vec{m}' \in \mathbb{D}^{|\vec{m}|}, \mathbb{R}_{Z',\vec{m}'} \text{ is Lebesgue measurable}\}$$

defining $\nu^w_{l^*} \colon \mathcal{Z}_z \mapsto \mathbb{R}^{\geq 0}$ as:

$$\nu^w_{l^*}(Z') = \sum_{\vec{m}' \in \mathbb{D}_{Z'}} \int_{\vec{y}' \in \mathbb{R}_{Z',\vec{m}'}} l^*(w, \bigodot_{i=1}^{k} a'_i) d\vec{y}' \tag{4.16}$$

where a'_i are actions obtained from a_i by replacing m_i, y_i respectively with m'_i, y'_i, then $\langle Z_z, \mathcal{Z}_z, \nu^w_{l^*} \rangle$ forms a measure space.

For the rest of this chapter, by a world we always mean a (globally) proper one. The remaining task is to update the agent's epistemic state after sensing. For this, we define the *compatibility* between action sequence and epistemic states. Formally, given μ and z, we require that there exists a value η^μ_z (or simply η) s.t.

$\eta = \int_w \nu_{l*}^w(Z_{\text{TRUE}})d\mu$. Intuitively, η represents the likelihood of action sequence z under μ.

We call μ *compatible* with action sequence z iff $0 < \eta_z^\mu < \infty$. Namely, the sequence z has a positive finite likelihood in μ. Given $z \in Z$, we denote $e_z = \{\langle \mathcal{W}, \mu\rangle \mid \langle \mathcal{W}, \mu\rangle \in e \text{ and } 0 < \eta_z^\mu < \infty\}$ as the compatible subset of e wrt z. For any set $Z' \subset Z_z$, the inner measure $\nu_{l*}^{w-}(Z')$ and outer measure $\nu_{l*}^{w+}(Z')$ can be defined similarly as in Def. 4.20. Now we can define the truth conditions of $\boldsymbol{B}$ and $\boldsymbol{O}$ after any action sequence :

- $e, w, z, \lambda \models \boldsymbol{B}(\alpha : r)$ if and only if for all $s = \langle \mathcal{W}, \mu\rangle$, $s \in e_z$ *implies* $\frac{1}{\eta}\int_w \nu_{l*}^{w+}(Z_\alpha^{e,w,z})d\mu = \frac{1}{\eta}\int_w \nu_{l*}^{w-}(Z_\alpha^{e,w,z})d\mu = \|r\|_\lambda$.
- $e, w, z, \lambda \models \boldsymbol{O}(\alpha : r)$ if and only if for all $s = \langle \mathcal{W}, \mu\rangle$, $s \in e_z$ *iff* $\frac{1}{\eta}\int_w \nu_{l*}^{w+}(Z_\alpha^{e+s,w,z})d\mu = \frac{1}{\eta}\int_w \nu_{l*}^{w-}(Z_\alpha^{e+s,w,z})d\mu = \|r\|_\lambda$.

For a sentence α, we write $e, w \models \alpha$ to mean $e, w, \langle\rangle, \lambda \models \alpha$ for all variable maps λ. When Σ is a set of sentences and α is a sentence, we write $\Sigma \models \alpha$ (read: Σ logically entails α) to mean that for every e and w, if $e, w \models \alpha'$ for every $\alpha' \in \Sigma$, then $e, w \models \alpha$. We say that α is valid ($\models \alpha$) if $\{\} \models \alpha$. Satisfiability is then defined in the usual way. If α is an objective formula, we write $w \models \alpha$ instead of $e, w \models \alpha$. Similarly, we write $e \models \alpha$ instead of $e, w \models \alpha$ if α is subjective.

4.2.4. Properties

Now that we have introduced the complete semantics, let us examine the properties of our logic. Despite the increasing expressiveness, our notions of belief and only-believing retain many of the properties known from its static predecessor, the logic $\mathcal{OBL}_c$, and its discrete predecessor, the logic $\mathcal{DS}$.

Additivity and Equivalence

We show that $\boldsymbol{B}$ satisfies the properties of probability as in $\mathcal{DS}$.

- If $\models \Box(\alpha \equiv \beta)$ then $\models \Box \boldsymbol{B}(\alpha : r) \equiv \boldsymbol{B}(\beta : r)$

  Namely, equivalent formulas have an equivalent degree of belief. The key of the proof is to show $Z_{\alpha}^{e,w,z} = Z_{\beta}^{e,w,z}$ for all e, w and z, which can be derived from the assumption that $\models \Box(\alpha \equiv \beta)$. We show $\models \Box \boldsymbol{B}(\alpha\colon r) \supset \boldsymbol{B}(\beta\colon r)$. The other direction can be proved likewise.

  Proof. It suffices to show that $e, z \models \boldsymbol{B}(\alpha\colon r)$ implies $e, z \models \boldsymbol{B}(\beta\colon r)$.

  Supposing that $e, z \models \boldsymbol{B}(\alpha\colon r)$, by the semantics, for all $\langle \mu, \mathcal{W} \rangle \in e_z$,

  $$\|r\| = \frac{1}{\eta} \int_w \nu_{l^*}^{w+}(Z_{\alpha}) d\mu = \frac{1}{\eta} \int_w \nu_{l^*}^{w-}(Z_{\alpha}) d\mu.$$

  On the other hand, since $\models \Box(\alpha \equiv \beta)$, $Z_{\alpha}^{e,w,z} = Z_{\beta}^{e,w,z}$ for all e, w, z. Hence,

  $$\|r\| = \frac{1}{\eta} \int_w \nu_{l^*}^{w+}(Z_{\beta}) d\mu = \frac{1}{\eta} \int_w \nu_{l^*}^{w-}(Z_{\beta}) d\mu.$$

  That is, $e, z \models \boldsymbol{B}(\alpha\colon r)$. □

- $\models \Box \boldsymbol{B}(\alpha\colon r) \supset \boldsymbol{B}(\neg\alpha\colon 1 - r)$

  This means $\boldsymbol{B}$ satisfies the complement law of probability. The proof is based on the fact that for all w, $\nu_{l^*}^{w+}(Z_{\neg\alpha}) = \mu_{l^*}^{w}(Z_{\mathrm{TRUE}}) - \nu_{l^*}^{w-}(Z_{\alpha})$, which can be derived from the definition of $\nu_{l^*}^{w+}$ and $\nu_{l^*}^{w-}$.

  Proof. Supposing $e, z \models \boldsymbol{B}(\alpha\colon r)$, by the semantics of $\boldsymbol{B}$, for all $\mu \in e_z$,

  $$\|r\| = \frac{1}{\eta} \int_w \nu_{l^*}^{w+}(Z_{\alpha}) d\mu = \frac{1}{\eta} \int_w \nu_{l^*}^{w-}(Z_{\alpha}) d\mu.$$

  By definition of $\nu_{l^*}^{w+}$ and $\nu_{l^*}^{w-}$,

  $$\begin{aligned}
  \nu_{l^*}^{w+}(Z_{\neg\alpha}) =& \inf\{\mu_{l^*}^{w}(Z^*) \mid Z^* \text{ measurable}, Z_{\neg\alpha} \subseteq Z^*\} \\
  =& \mu_{l^*}^{w}(Z_{\mathrm{TRUE}}) - \sup\{\mu_{l^*}^{w}(Z^*) \mid Z^* \text{ measurable}, Z^* \subseteq Z_{\alpha}\} \\
  =& \mu_{l^*}^{w}(Z_{\mathrm{TRUE}}) - \nu_{l^*}^{w-}(Z_{\alpha}).
  \end{aligned}$$

Therefore,

$$\int_w \nu_{l^*}^{w+}(Z_{\neg\alpha})d\mu = \int_w \mu_{l^*}^{w}(Z_{\text{TRUE}})d\mu - \int_w \nu_{l^*}^{w-}(Z_{\alpha})d\mu$$
$$= \eta - \|r\| \times \eta.$$

Likewise, $\int_w \nu_{l^*}^{w-}(Z_{\neg\alpha})d\mu = \eta - \|r\| \times \eta$. That is,

$$\frac{1}{\eta}\int_w \nu_{l^*}^{w+}(Z_{\neg\alpha})d\mu = \frac{1}{\eta}\int_w \nu_{l^*}^{w-}(Z_{\neg\alpha})d\mu = 1 - \|r\|.$$

By the semantics, $e, z \models \boldsymbol{B}(\neg\alpha\colon 1 - r)$. □

- $\models \Box\boldsymbol{B}(\alpha \wedge \beta : r) \wedge \boldsymbol{B}(\alpha \wedge \neg\beta : r') \supset \boldsymbol{B}(\alpha : r + r')$

  This can be proved in the same spirit as above with the fact that (likewise for $\nu_{l^*}^{w-}$) $\nu_{l^*}^{w+}(Z_\alpha) = \nu_{l^*}^{w+}(Z_{\alpha\wedge\beta}) + \nu_{l^*}^{w+}(Z_{\alpha\wedge\neg\beta})$.

- $\models \Box\boldsymbol{B}(\alpha : r) \wedge \boldsymbol{B}(\beta : r') \wedge \boldsymbol{B}(\alpha \wedge \beta : r'') \supset \boldsymbol{B}(\alpha \vee \beta : r + r' - r'')$

The above properties are stemming from the fact that $\boldsymbol{B}$ is indeed a probability over the possible world.

Knowledge

Now, let us turn to the properties of knowledge. Recall that $\boldsymbol{K}\alpha$ is an abbreviation for $\boldsymbol{B}(\alpha\colon 1)$. We show that our $\boldsymbol{K}$ shares many properties of its qualitative counterpart in the logic $\mathcal{ES}$ [LL11], its quantitative yet discrete counterpart in the logic $\mathcal{DS}$ [BL17], and its static counterpart $\mathcal{OBL}_c$ in Section 4.1.

- $\models \Box\boldsymbol{K}\alpha \supset \boldsymbol{K}(\alpha \vee \beta)$

  The proof is based on the fact that for all w, $Z_\alpha \subseteq Z_{\alpha\vee\beta}$.

  Proof. Suppose $e, z \models \boldsymbol{K}\alpha$. By the semantics, for all $\langle \mathcal{W}, \mu\rangle \in e_z$, $\eta = \int_w \nu_{l^*}^{w+}(Z_\alpha)d\mu = \int_w \nu_{l^*}^{w+}(Z_{\text{TRUE}})d\mu$. Meanwhile, $Z_\alpha \subseteq Z_{\alpha\vee\beta} \subseteq Z_{\text{TRUE}}$, which means $\nu_{l^*}^{w+}(Z_\alpha) \leq \nu_{l^*}^{w+}(Z_{\alpha\vee\beta}) \leq \nu_{l^*}^{w+}(Z_{\text{TRUE}})$. Hence,

$$\eta \leq \int_w \nu_{l^*}^{w+}(Z_{\alpha\vee\beta})d\mu \leq \int_w \nu_{l^*}^{w+}(Z_{\text{TRUE}})d\mu = \eta.$$

Therefore $\int_w \nu_{l*}^{w+}(Z_{\alpha\vee\beta})d\mu = \eta$, and likewise, $\int_w \nu_{l*}^{w-}(Z_{\alpha\vee\beta})d\mu = \eta$.
By the semantics $e, z \models \boldsymbol{K}(\alpha \vee \beta)$. □

- $\models \Box\boldsymbol{K}\alpha \wedge \boldsymbol{K}\beta \supset \boldsymbol{K}(\alpha \wedge \beta)$

  Proof. Suppose that $e, z \models \boldsymbol{K}\alpha \wedge \boldsymbol{K}\beta$. By the complement law $e, z \models \boldsymbol{B}(\neg\alpha\colon 0) \wedge \boldsymbol{B}(\neg\beta\colon 0)$. Additionally, $\models \Box\boldsymbol{K}\alpha \supset \boldsymbol{K}(\alpha \vee \beta)$, which means $e, z \models \boldsymbol{B}(\neg\alpha \wedge \neg\beta\colon 0)$. According to the addition law, $\models \boldsymbol{B}(\neg\alpha\colon 0) \wedge \boldsymbol{B}(\neg\beta\colon 0) \wedge \boldsymbol{B}(\neg\alpha \wedge \neg\beta\colon 0) \supset \boldsymbol{B}(\neg\alpha \vee \neg\beta\colon 0)$. Thus, $e, z \models \boldsymbol{K}(\alpha \wedge \beta)$. □

- $\models \Box\boldsymbol{K}\alpha \wedge \boldsymbol{K}(\alpha \supset \beta) \supset \boldsymbol{K}\beta$

  Proof. Suppose $e, z \models \boldsymbol{K}\alpha \wedge \boldsymbol{K}(\alpha \supset \beta)$. It suffice to prove $e, z \models \boldsymbol{K}((\alpha \vee \beta) \wedge (\neg\alpha \vee \beta))$. By the first property of knowledge, $e, z \models \boldsymbol{K}(\alpha \vee \beta)$. $\boldsymbol{K}(\neg\alpha \vee \beta)$ is equivalent to $\boldsymbol{K}(\alpha \supset \beta)$. Based on the second property of knowledge, $e, z \models \boldsymbol{K}\beta$. □

- $\models \Box\exists x.\boldsymbol{K}\alpha \supset \boldsymbol{K}\exists x.\alpha$

  Proof. For simplicity, we write Z_α^λ for $Z_\alpha^{e,w,z,\lambda}$.
  Suppose $e, z, \lambda \models \exists x.\boldsymbol{K}\alpha$. By the semantics, it exists λ' such that $\lambda' \sim_x \lambda$, and $e, z, \lambda' \models \boldsymbol{K}\alpha$. Namely, for all $\langle \mathcal{W}, \mu\rangle \in e_z$,

  $$\eta = \int_w \nu_{l*}^{w+}(Z_\alpha^{\lambda'})d\mu = \int_w \nu_{l*}^{w-}(Z_\alpha^{\lambda'})d\mu.$$

  Meanwhile, $Z_{\exists x.\alpha}^\lambda = \bigcup_{\{\lambda'' \mid \lambda'' \sim_x \lambda\}} Z_\alpha^{\lambda''}$. Therefore, $Z_\alpha^{\lambda'} \subseteq Z_{\exists x.\alpha}^\lambda$.

  $$\eta = \int_w \nu_{l*}^{w+}(Z_\alpha^{\lambda'})d\mu \leq \int_w \nu_{l*}^{w+}(Z_{\exists x.\alpha}^\lambda)d\mu \leq \int_w \nu_{l*}^{w+}(Z_{\mathrm{TRUE}}^\lambda)d\mu = \eta;$$
  $$\eta = \int_w \nu_{l*}^{w-}(Z_\alpha^{\lambda'})d\mu \leq \int_w \nu_{l*}^{w-}(Z_{\exists x.\alpha}^\lambda)d\mu \leq \int_w \nu_{l*}^{w-}(Z_{\mathrm{TRUE}}^\lambda)d\mu = \eta.$$

  Thus $\eta = \int_w \nu_{l*}^{w+}(Z_{\exists x.\alpha}^\lambda)d\mu = \int_w \nu_{l*}^{w-}(Z_{\exists x.\alpha}^\lambda)d\mu$. □

The converse of the above formula does not hold: knowing that α holds for someone does not imply knowing that individual. For the universal version, it holds

- $\models \Box \boldsymbol{K} \forall x.\alpha \supset \forall x.\boldsymbol{K}\alpha$

  This can be proved in a similar spirit as above and based on the fact that $Z_{\forall x.\alpha} = \bigcap_\lambda Z^\lambda_\alpha \subseteq Z^\lambda_\alpha$.

  Proof. Suppose $e, z \models \boldsymbol{K}\forall x.\alpha$. By the semantics, for all $\langle \mathcal{W}, \mu \rangle \in e_z$,

  $$\eta = \int_w \nu^{w+}_{l^*}(Z_{\forall x.\alpha})d\mu = \int_w \nu^{w-}_{l^*}(Z_{\forall x.\alpha})d\mu.$$

  Since $Z_{\forall x.\alpha} = \bigcap_\lambda Z^\lambda_\alpha \subseteq Z^\lambda_\alpha$,

  $$\eta = \int_w \nu^{w+}_{l^*}(Z_{\forall x.\alpha})d\mu \leq \int_w \nu^{w+}_{l^*}(Z^\lambda_\alpha)d\mu \leq \int_w \nu^{w+}_{l^*}(Z^\lambda_{\text{TRUE}})d\mu = \eta$$
  $$\eta = \int_w \nu^{w-}_{l^*}(Z_{\forall x.\alpha})d\mu \leq \int_w \nu^{w-}_{l^*}(Z^\lambda_\alpha)d\mu \leq \int_w \nu^{w-}_{l^*}(Z^\lambda_{\text{TRUE}})d\mu = \eta$$

  Hence, $e, z, \lambda \models \boldsymbol{K}\alpha$ for all λ, and $e, z \models \forall x.\boldsymbol{K}\alpha$. □

The converse of the above formula, i.e. $\Box\forall x.\boldsymbol{K}\alpha \supset \boldsymbol{K}\forall x.\alpha$, is not valid as already shown in [FLBL23] and Section 4.1 for the static cases. The reason is that in a probability space with uncountably many samples, there could be uncountably many distinct events where each of them has probability 1 (those satisfying $\boldsymbol{K}\alpha$), any countable intersection of them has probability 1, yet an uncountable intersection does not have probability 1.

We have the following properties for introspection:

- $\models \Box \boldsymbol{K}\alpha \supset \boldsymbol{K}\boldsymbol{K}\alpha$
- $\models \Box \neg \boldsymbol{K}\alpha \supset \boldsymbol{K}\neg\boldsymbol{K}\alpha$
- $\models \Box \boldsymbol{B}(\alpha : r) \supset \boldsymbol{K}\boldsymbol{B}(\alpha : r)$
- $\models \Box \neg \boldsymbol{B}(\alpha : r) \supset \boldsymbol{K}\neg\boldsymbol{B}(\alpha : r)$

That is, the agent has knowledge about what is known or believed, as well as what is not known or believed. We only prove the last two items.

Proof. Suppose that $e, z \models \boldsymbol{B}(\alpha\colon r)$, we prove that $e, z \models \boldsymbol{K}\boldsymbol{B}(\alpha : r)$. For any $\langle \mathcal{W}, \mu \rangle \in e_z$, from definition we know that for any w and z' with $z' \sim z$, $Z_\alpha^{e,w,z} = Z_\alpha^{e,w,z'}$. Thus $e, w, z \models \boldsymbol{B}(\alpha : r)$ iff $e, w, z' \models \boldsymbol{B}(\alpha : r)$. Then

$$\begin{aligned} Z_{\boldsymbol{B}(\alpha:r)}^{e,w,z} &= \{z' | z' \sim z, e, w, z' \models \boldsymbol{B}(\alpha : r)\} \\ &= \{z' | z' \sim z, e, w, z \models \boldsymbol{B}(\alpha : r)\} = Z_z \end{aligned}$$

Thus for any $\langle \mathcal{W}, \mu \rangle \in e_z$, the integral of ν_{l*}^{w-} and ν_{l*}^{w+} will both equal to η.
Now, suppose that $e, z \models \neg\boldsymbol{B}(\alpha\colon r)$. Similarly, for any w, $Z_{\neg\boldsymbol{B}(\alpha:r)}^{e,w,z} = Z_z$.
Thus, for all $\langle \mathcal{W}, \mu \rangle \in e_z$, the integral of ν_{l*}^{w-} and ν_{l*}^{w+} will both equal to η. □

Only-Believing

Lastly, we examine the properties of $\boldsymbol{O}$. Recall that only-knowing (or only-believing) captures the intuition that the beliefs and non-beliefs of an agent are precisely those that follow from its knowledge base. Hence it is useful to characterize a knowledge base. Similar properties and proof technique can be found in Section 4.1.

To begin with, the unique model theorem holds for only-believing an objective formula:

Theorem 4.25 (Unique Model Theorem)**.** For any objective sentence α and rigid ground term r, there is a unique epistemic state e such that $e \models \boldsymbol{O}(\alpha\colon r)$.

As in the logic $\mathcal{OL}$ [LL01], only-knowing implies knowing and not knowing about what is not entailed by the knowledge base. Below, let α be an arbitrary sentence and ϕ, ψ objective sentences:

- $\models \boldsymbol{O}(\alpha\colon r) \supset \boldsymbol{B}(\alpha\colon r)$
- $\models \boldsymbol{O}\phi \supset \boldsymbol{K}\psi$ iff $\phi \models \psi$
- $\models \boldsymbol{O}(\alpha\colon r) \supset \neg\boldsymbol{B}(h(\vec{n}) = m\colon r')$ for all r', where $\vec{n}, m$ are standard names and h is a fluent not in α.

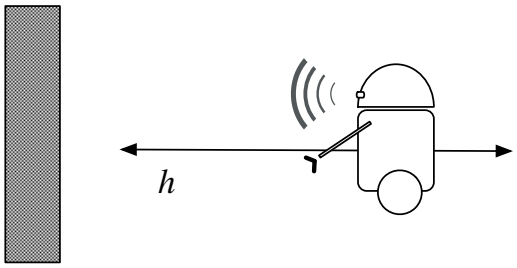


Figure 4.3.: A robot moving towards a wall.

E.g. let $\Sigma := \boldsymbol{O}(\mathit{weightOf}(A) \leq 50\colon 0.9)$, then we have $\Sigma \models \boldsymbol{B}(\mathit{weightOf}(A) \leq 50\colon 0.9)$ and $\Sigma \models \forall x.\neg\boldsymbol{B}(1.7 \leq \mathit{heightOf}(A)\colon x)$. That is, only-believing the person A's weight is less than 50kg with a degree 0.9 entails believing the person's weight is less than 50kg and also, not believing A's height is greater than 1.7m with any degree.

4.2.5. Reasoning about Actions and Beliefs

We show the expressiveness and flexibility of our logic through examples.

Expressing Belief Distributions

To reason about beliefs, one needs to specify what is believed initially. Let us consider the robot moving example in Fig. 4.3 from [BL17]. In the (discrete) logic $\mathcal{DS}$ and its extensions, a nullary fluent h is used to indicate the robot's horizontal distance toward the wall. Moreover, the robot's uncertain belief about h is specified via a formula of the form $\forall x.\boldsymbol{B}(h = x\colon f(x))$. Namely, the rigid mathematical function $f(x)$ expresses a *belief distribution* of the random variable h where $\sum_x f(x) = 1$. This can be easily extended for multiple but finite nullary fluents.

Similar to the logic $\mathcal{DS}$, we first consider distributions with a finite set of nullary fluents. However, we will later show that, in our logic, expressiveness is not restricted to these cases. Given a finite set of nullary fluents $\vec{h} = (h_1, h_2, \ldots h_k)$,

similar to Eq.(4.1) in Section 4.1, we represent a (potentially continuous) belief distribution over $\vec{h}$ via formulae of the form

$$\forall \vec{x}.\boldsymbol{B}(\bigwedge_{i=1}^{k} h_i \leq x_i \colon f(\vec{x})),$$

where f is a rigid mathematical function satisfying $f(-\infty) = 0$ and $f(+\infty) = 1$. [15] We write $\boldsymbol{B}^f$ for short. Intuitively, $f(\vec{x})$ represents the *joint cumulative distribution function* of random variables $\vec{h}$. For example, the formula $\forall x.\boldsymbol{B}(h \leq x \colon \mathcal{U}(x; 10, 11))$ expresses the robot believes its distance is distributed uniformly among the interval $[10, 11]$, where

$$\mathcal{U}(x; 10, 11) = \begin{cases} 0 & x < 10 \\ x - 10 & x \in [10, 11) \\ 1 & 11 \leq x \end{cases} \tag{4.17}$$

Note that $\boldsymbol{B}^f$ could also be a discrete or mixture distribution. The belief distributions $\forall x.\boldsymbol{B}(h \leq x \colon f'(x))$ and $\forall x.\boldsymbol{B}(h \leq x \colon f''(x))$ below respectively express such distributions.

$$f'(x) = \begin{cases} 0 & x < 1 \\ 0.5 & x \in [1, 2) \\ 1 & 2 \leq x \end{cases} \qquad f''(x) = \begin{cases} 0 & x < 0 \\ 0.5 & x = 0 \\ x + 0.5 & x \in (0, 0.5] \\ 1 & 1 \leq x \end{cases}$$

The latter is rather useful since the robot's belief after actions is usually a mixture distribution even if it starts with a pure continuous one. Imagine that the robot cannot get across the wall and its initial belief distortion is given as $\mathcal{U}(x; 10, 11)$, then after a deterministic action $fwd(10.5)$, it will end up in $\boldsymbol{B}^{f''}$. This is because some probability mass will accumulate at the point $h = 0$.

[15] Here $f(+\infty) = 1$ means $\lim_{\vec{x} \to \infty} f(\vec{x}) = 1$. When there is only one argument, $\lim_{x \to \infty} f(x) = y$ is defined as $\forall u.\big(u > 0 \supset \exists m.\forall v.((v > m) \supset |y - f(v)| < u)\big)$. Cases with multiple arguments can be represented similarly.

We comment that expressing such mixture distribution is impossible, both in the logic $\mathcal{DS}$ and in the BL framework. The reason is that their notion of belief extends the BHL formalism and is defined in terms of normalized sums or integral of weights of situations. Since it is unclear how to integrate over situations, they map situations to fluent values which are purely continuous, then belief is given in terms of integral over the continuous values of fluents that are either pure continuous or pure discrete.

Basic Action Theories

To infer beliefs after actions, one needs to specify a theory of action besides the initial beliefs. In our logic, the basic action theory includes a set of successor state axioms, one for each fluent, incorporating Reiter's [Rei01] solution to the frame problem, and a set of likelihood axioms, one for each action symbol.[16]

The following is a possible basic action theory for the robot moving example

- the successor state axiom Σ_{post} for fluent h as:

$$\begin{aligned}\Box[a]h = v \equiv &\exists x, y.a = fwd(x,y) \wedge v = h - y \\ &\vee \forall x, y.a \neq fwd(x,y) \wedge v = h\end{aligned}$$

That is, the robot's new location is always determined by the actual outcome y of the action $fwd(x, y)$;

- the likelihood axioms Σ_l for stochastic action $fwd(x, y)$ and noisy sensing $sonar(x)$

$$\Box l(fwd(x,y)) = \mathcal{N}(y; x, 1)$$
$$\Box l(sonar(x)) = \mathcal{N}(x; h, 0.25)$$

Namely, always, the likelihood of $fwd(x, y)$ is given by a Gaussian density for the outcome y centered on the intended value x with a variance of 1. Likewise,

[16] For simplicity, we do not include precondition axioms for actions and assume all actions are executable yet the usual "impossible actions" will have a zero likelihood.

the sensing model is given by a Gaussian density for the value read x centered on the true value h (the fluent that it is measuring) with a variance of 0.25. Let $\Sigma_{dyn} = \Sigma_{post} \cup \Sigma_l$ and Σ be as:

$$\forall x. \boldsymbol{B}(h \leq x \colon \mathcal{U}(x; 10, 11)) \wedge \boldsymbol{K}\Sigma_{dyn}$$

then Σ entails the following:

- $[fwd(2, 2.1)]\boldsymbol{B}(h \leq 9 \colon 0.6844)$

Intuitively, when the robot intends to forward 2 units, even if nature selects 2.1 as an outcome, the robot has no observation of this and considers $fwd(2, y)$ possible for all y with y distributed as $\mathcal{N}(y; 2, 1)$. The posterior of $h \leq 9$ amounts to the product of the prior of h and the probability of y subject to $h - y \leq 9$, this is because Σ_{post} says the locations after action $fwd(x, y)$ is given by $h - y$ where h is the initial location. Hence the posterior amounts to

$$\int_{\mathbb{R}} \int_{\mathbb{R}} \left(\begin{cases} \mathcal{N}(y; 2, 1) & \psi \\ 0 & \text{otherwise} \end{cases} \right) dydx = 0.6844 \tag{4.18}$$

here ψ is the conjunction of three conditions: 1) $h = x$; 2) $x \in [10, 11]$ since h distributes uniformly in [10,11]; 3) $x - 9 \leq y$. Hence, $\Sigma \models [fwd(2, 2.1)]\boldsymbol{B}(h \leq 9 \colon 0.6844)$. Formally,

Proof. Suppose $e \models \Sigma$. For all $\mu \in e$. By definition,

$$\begin{aligned} Z_{h\leq 9}^{w, fwd(2,2.1)} &= \{z' | z' \sim fwd(2, 2.1), w, z' \models h \leq 9)\} \\ &= \{a' | \exists y. a' = fwd(2, y), \|h\|_w - y \leq 9\} \end{aligned}$$

Hence $\mathbb{R}_{Z_{h\leq 9}} = \{y \,|\, \|h\|_w - 9 \leq y\}$. By Def. of $\nu_{l^*}^{w+}$,

$$\begin{aligned} \nu_{l^*}^{w+}(Z_{h\leq 9}) &= \int_{y \in \mathbb{R}_{Z_{h\leq 9}}} l^*(w, fwd(2, y)) dy \\ &= \int_{\mathbb{R}} \left(\begin{cases} \mathcal{N}(y; 2, 1) & \|h\|_w - y \leq 9 \\ 0 & \text{otherwise} \end{cases} \right) dy. \end{aligned}$$

Consequently $\int_w \nu_{l^*}^{w+}(Z_{h\leq 9})d\mu$ amounts to

$$\int_w \int_{\mathbb{R}} \left(\begin{cases} \mathcal{N}(y;2,1) & \|h\|_w - y \leq 9 \\ 0 & \text{otherwise} \end{cases} \right) dyd\mu \tag{4.19}$$

which is essentially the same as Eq. (4.18) and equal to 0.6844. The same result holds for $\nu_{l^*}^{w-}$, hence, $e \models [fwd(2,2.1)]\boldsymbol{B}(h \leq 11 \colon 0.6844)$. □

- $[fwd(2,2.1) \cdot sonar(8)]\boldsymbol{B}(h \leq 9 \colon 0.9778)$

After the sonar reads a value less than 9, i.e. $sonar(8)$, the robot's belief of being less than 9 is enhanced. After the action sequence $fwd(2,2.1) \cdot sonar(8)$, the robot considers the set of sequences $fwd(2,y) \cdot sonar(8)$ possible for all y. Moreover, the likelihood of such a sequence is given by $\mathcal{N}(y;2,1) \times \mathcal{N}(8;h,0.25)$, the posterior of $h \leq 9$ is

$$\int_{\mathbb{R}} \int_{\mathbb{R}} \left(\begin{cases} \mathcal{N}(y;2,1)\mathcal{N}(8;h,0.25) & \psi' \\ 0 & \text{otherwise} \end{cases} \right) dydx \tag{4.20}$$

where ψ' is the conjunction of:

1. $x = h + y$;
2. $x \in [10,11]$;
3. $x - 9 \leq y$.

Condition 1. is because h here represents the location after the sequence $fwd(2,y) \cdot sonar(8)$, hence, the initial value would be $h + y$ as sensing does not change the world. After normalization, the result is 0.9778. Thus,

$$\Sigma \models [fwd(2,2.1)]\boldsymbol{B}(h \leq 9 \colon 0.9778)$$

Beyond Nullary Fluent

Another appealing point of our logic is that when it comes to expressing degrees of belief, fluents do not have to be nullary, which goes beyond previous proposals such as BHL, $\mathcal{DS}$, and $\mathcal{XS}$ [Bel23] and their variants.

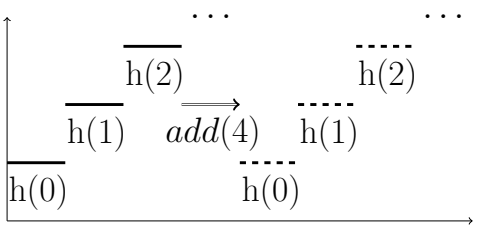


Figure 4.4.: Infinitely many random variables $h(x)$.

As a concrete example, imagine there are infinitely many random variables $h(x)$ with $x \in \mathbb{N}^{\geq 0}$ distributed uniformly among the interval $[x, x+1]$ which looks like an infinite ladder in Fig. 4.4. Supposing deterministic action $increase(b)$ will increase the values of $h(x)$ by b, then $h(x)$ should be distributed uniformly among the interval $[x+b, x+b+1]$ after the action $increase(b)$. Formally,

Example 4.26. let Σ_{dyn} be as:

$$\begin{aligned} &\Box[a]h(x) = v \equiv \exists x'.a = increase(x') \wedge v = h(x) + x' \\ &\qquad\qquad \vee \forall x'.a \neq increase(x') \wedge v = h(x) \\ &\Box l(increase(x')) = 1. \end{aligned} \tag{4.21}$$

Then, we have that

- $\models \forall x, y.\boldsymbol{B}(h(x) \leq y \colon \mathcal{U}(y, x, x+1)) \wedge \boldsymbol{K}\Sigma_{dyn}$
  $\supset [increase(4)]\forall x, y.\boldsymbol{B}(h(x) \leq y \colon \mathcal{U}(y, x+4, x+5))$

We comment that the result does not have to limit to uniform distributions and deterministic actions. A more complex example could be as follows: suppose there are infinitely many boxes and each has infinitely many balls, yet the number of balls $numOfBall(x)$ ($x \in \mathbb{N}^{\geq 0}$) is distributed as Poisson with an expectation of 6. Now the robot observes some balls are added to the boxes. It cannot tell the exact number of balls that are added to each box, yet the number of new balls for each box is another Poisson distribution with an expectation of 3. As a result, the robot would believe that $numOfBall(x)$ is distributed as Poisson with an expectation of 9. This can be formulated as:

Example 4.27. Let Σ_{dyn} be as:

$$\Box[a]numOfBall(x) = v \equiv \exists y.a = add(y) \wedge v = numOfBall(x) + y \vee \forall y.a \neq add(y) \wedge v = h(x)$$

$$\Box l(add(y)) = \begin{cases} 3^y e^{-3}/y! & y \in \mathbb{N}^{\geq 0} \\ 0 & \text{otherwise} \end{cases}$$

Then we have:

- $\models \forall x, y.\boldsymbol{B}(h(x) = y\colon 6^y e^{-6}/y!) \wedge \boldsymbol{K}\Sigma_{dyn}$
  $\supset [add(4)]\forall x, y.\boldsymbol{B}(h(x) = y\colon 9^y e^{-9}/y!)$

To sum up, our logic is capable of capturing interesting results such as the sum of two Poisson is another Poisson, the sum of two Gaussian is another Gaussian, and the product of two log-norms is another log-norm, making it more expressive than the existing accounts.

4.3. Discussion: Relation to Levesque's Only-Knowing

Readers who are familiar with Levesque's logic $\mathcal{OL}$ may notice that, in the logics $\mathcal{OBL}_c$ and $\mathcal{PS}$, the add-one-in semantics deviates from the standard semantics of only-knowing. We end this chapter by presenting a variant of $\mathcal{OL}$ that corresponds to the add-one-in semantics, and by comparing our semantics with the standard one. An introduction to $\mathcal{OL}$ is given in Section 3.1; see also [LL01] for further details. As a reminder of the $\mathcal{OL}$ semantics, given a model (e, w), terms $t, t_1, \ldots t_k$, and formulae α, β, the satisfaction relation is defined as follows:

- $e, w \models P(t_1, \ldots, t_k)$ iff $w[P(n_1, \ldots, n_k)] = 1$ where $n_i = w(t_i)$;
- $e, w \models t_1 = t_2$ iff $w(t_1)$ and $w(t_2)$ are identical;
- $e, w \models \neg\alpha$ iff it is not the case that $e, w \models \alpha$;

- $e, w \models \alpha \wedge \beta$ iff $e, w \models \alpha$ and $e, w \models \beta$;
- $e, w \models \forall x.\alpha$ iff $e, w \models \alpha^x_n$ for all $n \in \mathcal{N}$;
- $e, w \models \boldsymbol{K}\alpha$ iff for every w', $w' \in e$ implies $e, w' \models \alpha$;
- $e, w \models \boldsymbol{O}\alpha$ iff for every w', $w' \in e$ iff $e, w' \models \alpha$.

The variant we consider differs from Levesque's original design only in the interpretation of $\boldsymbol{O}$-formulae. To distinguish the two, we use $\boldsymbol{O}'$ to denote the only-knowing modality subject to the add-one-in semantics, and extend Levesque's $\mathcal{OL}$ with an additional item as follows:

- $e, w \models \boldsymbol{O}'\alpha$ iff for every w', $w' \in e$ iff $(e + w'), w' \models \alpha$.

where $(e + w')$ denotes the epistemic state $e' = e \cup \{w'\}$. Clearly, $e' = e$ iff $w' \in e$. The following results can be proved directly by the semantics:

Proposition 4.28. For any α, $\models \boldsymbol{O}'\alpha \supset \boldsymbol{K}\alpha$.

Theorem 4.29. For any objective ϕ, $\models \boldsymbol{O}\phi \equiv \boldsymbol{O}'\phi$

Namely, $\boldsymbol{O}$ and $\boldsymbol{O}'$ behave identically for objective formulae. When ϕ is not objective, however, $\boldsymbol{O}$ and $\boldsymbol{O}'$ shows difference properties. For atomic ψ, $\boldsymbol{O}\boldsymbol{K}\psi$ is not satisfiable (See Cor. 3.8). In contrast, the add-one-in semantics behaves differently that only-knowing about a subjective formula can be reduced to only-knowing the corresponding objective formula:

Proposition 4.30. Suppose that ψ is objective. Then $\models \boldsymbol{O}'\boldsymbol{K}\psi \equiv \boldsymbol{O}'\psi$

Proof. Consider an atom ψ, then for any e

$$
\begin{aligned}
& e \models \boldsymbol{O'K}\psi \\
\iff & \text{for all } w, w \in e \text{ iff } e + w, w \models \boldsymbol{K}\psi \\
\iff & \text{for all } w, w \in e \text{ iff } e + w \models \boldsymbol{K}\psi \\
\iff & \text{for all } w, w \in e \text{ iff for any } w' \in (e + w), e + w, w' \models \psi \\
\iff & \text{for all } w, w \in e \text{ iff for any } w' \in (e + w), w' \models \psi \\
\iff & \text{for all } w, w \in e \text{ iff } w \models \psi \text{ and for all } w' \in e, w' \models \psi \qquad (\#)
\end{aligned}
$$

Clearly, $e_\psi = \{w \mid w \models \psi\}$ is the only epistemic state which satisfies (#). Thus $\models \boldsymbol{O'K}\psi \supset \boldsymbol{O'}\psi$. Additionally, e_ψ is also the unique state which satisfies $\boldsymbol{O'}\psi$, thus $\models \boldsymbol{O'}\psi \supset \boldsymbol{O'K}\psi$.

□

Our semantics also preserves the property of conjoining knowledge to what is only-known without changing the state.

Theorem 4.31. $\models (\boldsymbol{O'}\alpha \wedge \boldsymbol{K}\beta \supset \boldsymbol{O'}(\alpha \wedge \beta))$

Proof. Suppose that $e \models \boldsymbol{O'}\alpha \wedge \boldsymbol{K}\beta$, then $e \models \boldsymbol{K}\alpha \wedge \boldsymbol{K}\beta$ and thus $e, w \models (\alpha \wedge \beta)$ for all $w \in e$. For $w \notin e$, since $e \models \boldsymbol{O'}\alpha$, $e + w, w \not\models \alpha$, so $e + w, w \not\models \alpha \wedge \beta$.

□

The motivation to adopt this non-standard semantics arises when considering the probabilistic extension of only-knowing: as we have already seen in this chapter, beliefs about a probability distribution will, in general, be represented by a subjective formula ψ. As in Thm. 3.7, "only-knowing" a (subjective) distribution ψ cannot be straightforwardly expressed as $\boldsymbol{O}\psi$. To handle the problem, Belle, Lakemeyer, and Levesque proposed two solutions:

- In the logic $\mathcal{OBL}$ [BLL16], they considered an extended modality of the form $\boldsymbol{O}(\alpha_1 : r_1, \ldots, \alpha_k : r_k)$, read as "All I believe is α_1 with degree r_1, ..., and α_k with degree r_k". However, this solution limits the expressiveness.

- In the logic $\mathcal{DS}$[BL17], they modified the semantics and treated elements in the epistemic state independently. However, under this consideration, the knowledge and beliefs of the agent no longer follow the property of negative introspection.

Our treatment of only-knowing addresses both concerns: while preserving most of the expressive power, our account ensures the $K45$ properties of $\mathcal{OL}$ and lifts them to the probabilistic cases.

5

Reasoning about Higher-order Beliefs

The previous chapter addressed the integration of logic and probability, focusing on the development of semantic models and the expressiveness of the language. In contrast, the present chapter turns to a different but equally fundamental aspect of epistemic logic: reasoning in multi-agent systems. In particular, reasoning about higher-order beliefs. In a scenario with many agents, an agent's knowledge and beliefs are not only about the world but also about one another's mental state. By extending the single-agent modality $\boldsymbol{K}$ for knowledge to a family of indexed modalities $\boldsymbol{K}_i$, multi-agent epistemic logics allow the nesting of modalities such as $\boldsymbol{K}_i\neg\boldsymbol{K}_j\varphi$, expressing that agent i knows that agent j does not know φ. However, as the number of agents and the depth of nesting increase, reasoning in these systems quickly becomes complicated or even undecidable.

While the representation of higher-order beliefs has been well investigated since the early works of [Che80, HM92], the main challenge lies not in representation but in reasoning: how to perform inference efficiently, or how to reduce complex multi-agent reasoning problems to simpler problems in which sophisticated reasoning tools exist.

This chapter includes our work on addressing the reasoning problems. In the static cases, we investigate how the theorem proving for multi-agent modal logics of knowledge and belief—such as $K45_n$ and $S5_n$—can be translated into decidable

fragments of first-order logic. In the dynamic cases, we propose a situation calculus approach for multi-agent epistemic planning. Inspired by the cognitive notion of Theory of Mind (ToM), our framework bounds the order of mental recursion, which ensures the decidability of planning while retaining most of the expressive power of existing accounts such as Dynamic Epistemic Logic (DEL).

5.1. Translation-based Modal Theorem Proving

Modal logics, particularly epistemic modal logics, extend ordinary logical systems by introducing modalities for knowledge and belief. These enable the expression and reasoning about the subjective, epistemic attitude of agents. Over the past several decades, epistemic modal logics have attracted significant attention and have been widely applied in various domains, including protocol analysis in distributed systems [HM90], knowledge-based systems [LL01] and epistemic planning [Bol17].

The conventional semantics of modal logic is defined via a set of possible-worlds and their relations. Such an approach, commonly known as Kripke semantics, was initially proposed by Hintikka and Kripke [Hin62, Kri63]. It was later extended by numerous scholars (for a detailed overview, see [VDvdHHK15]). One of the advantages of Kripke semantics is that it provides a graphical interpretation for many characteristics of agent reasoning. It has been shown that normal modal systems such as K45 for belief and S5 for knowledge can be precisely captured by Kripke frames satisfying certain relational properties [Che80]. Based on Kripke semantics, Ladner [Lad77], Halpern and Moses [HM92] proved that the satisfiability problem for modal logics of knowledge and belief is decidable, yet of a high complexity especially when multiple agents are involved. Therefore, it is essential to explore modal inference approaches, which are efficient in practice.

Traditionally, two paradigms in modal reasoning can be distinguished. The first is to develop proving techniques directly for modal logic, such as modal resolution [Sch99, DNSH00, SH13] or tableaux calculi [Cat91]. The second is translating

the modal formulae into an ordinary system, such as first-order logic. With the second approach, efficient theorem provers and SAT-solvers developed for classical logic, such as VAMPIRE[KV13], Prover9 and MACE4[McC10], can be applied for modal reasoning.

The idea of translating modal logic into first-order logic was first introduced by Fine, van Benthem, and Morgan [Fin75, Mor76, vB77], known as the *standard translation*. As shown in [Var96], this approach can encode the simplest normal modal logic K into the first-order fragment GF^2, which is the intersection of the guarded fragment GF and the two-variables fragment FO^2 [GO99, GMV99]. Since GF^2 has an EXPTIME complexity upper bound, the translation for K is decidability-preserving.[1] However, for agent-based modal logics with introspection, i.e. accounts where agents can reflect on their own knowledge and lack of knowledge, the translation is out of the decidable fragment, mainly because frame properties such as transitivity or Euclideanity cannot be expressed in GF or FO^2. To translate modal systems with introspection, many variants of the standard translation have been studied: Auffray and Enjalbert proposed a translation into path logic, where a sub-domain of action is introduced to handle modal operators and their nesting [AE89]. Nonnengart combined relational translation with functional translation [Non93]. Schmidt and Hustadt associate (sub-)formulae with separate predicate symbols and propose an axiomatic translation [SH03, SH07], Demri and de Nivelle studied the translation for regular grammar logic with converse[DDN05], which can be used to embed some modal systems. While some of the mentioned approaches including [DDN05, SH07] can preserve decidability, these accounts mostly focus on single-agent languages. A clear formalism for multi-agent logics of knowledge and belief, such as $K45_n$, $S5_n$, cannot be found. Additionally, it is unclear how these methods can be extended to first-order modal logics. We notice that all of the existing translation approaches are based on simulation and axiomatization of Kripke structures, and the undecidability usually

[1] In fact, it translates into an even more restricted fragment of GF^2 [StC13, TCS11].

comes from expressing the frame conditions. It raises the question of whether axiomatizing Kripke models is the optimal choice to represent modal properties in a first-order language.

This section is organized as follows: In 5.1.1, a set-based semantics for propositional modal logic is introduced. We analyze its properties and compare it with ordinary Kripke semantics. Based on explicitly specifying the inclusion relation of the models in this semantics, we introduce a new translation approach in 5.1.4. We prove that this approach is decidability-preserving, and the translation has linear complexity with a small linear scaling factor. Furthermore, we extend the method to translating first-order modal formulae and discover a decidable first-order modal fragment.

5.1.1. The Logic $\mathcal{ML}_n$

We introduce a dialect of modal logic called $\mathcal{ML}_n$. The language to be expressed is identical to ordinary modal logic [Che80]. The semantics can be considered as a propositional fragment of the logic proposed in [BL15b, BL10].

Syntax

Let Ap be the set of atomic propositions, which could potentially be countably infinite. The logic includes the usual Boolean connectives $\neg$ and $\wedge$. Others like $\vee, \supset, \equiv$ are interpreted as abbreviations.

Let $\mathsf{Ag} = \{1, \ldots, n\}$ denote a finite set of agents. For each $i \in \mathsf{Ag}$, an epistemic operator $\boldsymbol{K}_i$ is used to denote knowledge or belief of agent i. The language of $\mathcal{ML}_n$ is defined as follows:

$$\varphi := p \,|\, \neg\varphi \,|\, \varphi \wedge \varphi \,|\, \boldsymbol{K}_i\varphi$$

Where $p \in \mathsf{Ap}$ and $i \in \mathsf{Ag}$. We say a formula is *objective* if it mentions no $\boldsymbol{K}$-operators and a formula is *subjective* if every proposition occurs in the scope of a $\boldsymbol{K}$-operator. A formula is *i-objective* if every occurrence of $\boldsymbol{K}_i$ is in the scope of a

$\boldsymbol{K}_j$ where $j \neq i$, and a formula is *i-subjective* if every occurrence of propositions or $\boldsymbol{K}_j$ for $j \neq i$ is in the scope of a $\boldsymbol{K}_i$. In addition, we use $\textsc{True} := p \vee \neg p$ and $\textsc{False} := p \wedge \neg p$ to represent tautologies and falsity.

As Halpern pointed out in [Hal95], the complexity of modal logic is affected by the depth of nesting of modal operators. To evaluate the depth of a formula, we define:

Definition 5.1 (i-depth)**.** For $i \in \mathsf{Ag}$, the i-depth of a formula α, written as $dep[\alpha, i]$, is recursively defined as

- $dep[p, i] = 0$ for $p \in \mathsf{Ap}$
- $dep[\neg\alpha, i] = dep[\alpha, i]$
- $dep[\alpha \wedge \beta, i] = max(dep[\alpha, i], dep[\beta, i])$
- $dep[\boldsymbol{K}_i\alpha, i] = \max\{\max\{dep[\alpha, j] \mid j \neq i\} + 1, dep[\alpha, i]\}$
- $dep[\boldsymbol{K}_j\alpha, i] = 0$ for $j \neq i$

Intuitively, $dep[\alpha, i]$ stands for the depth of agent i's knowledge in formula α. The depth of a formula α is defined as the maximal i-depth, written as $dep[\alpha] = \max_{i \in \mathsf{Ag}} dep[\alpha, i]$.

Example 5.2. Let $\varphi := \neg p \wedge \neg \boldsymbol{K}_2(\neg p)$. Suppose that p means the coin is fair. Let $\alpha := \boldsymbol{K}_1\varphi$, then α means that agent 1 knows that the coin is not fair, and he knows that agent 2 doesn't know the unfairness.

$$\begin{aligned} dep[\alpha, 1] &= \max\{dep[\varphi, 2] + 1, dep[\varphi, 1]\} \\ &= \max\{1 + 1, 0\} = 2 \\ dep[\alpha, 2] &= dep[\boldsymbol{K}_1\varphi, 2] = 0 \end{aligned}$$

Semantics

The logic $\mathcal{ML}_n$ has a set-based possible-worlds semantics, where introspection of agents have been embedded. It was first used by Levesque as the semantics for

$K45$ [Lev82]. In Levesque's design, a model consists of a distinguished state that describes what is "true" in the real world and a set of states that the agent thinks possible. Levesque's model was further developed as the semantics to formalize the notion of "only-knowing" [Lev90, LL01]. Belle and Lakemeyer extended the semantics to multi-agent cases by defining a nested set structure [BL10, BL15b] and we adopt similar structures in our semantics. By a *world* we mean a set of atomic propositions considered as true, i.e. $w \subseteq \mathsf{Ap}$. Let W be the set of all possible-worlds. For $k \geq 0$, we inductively define the epistemic states (also written as states for short) and structures as follows:

- $\mathcal{E}^0 = \{\{\varnothing\}\}$;
- $\mathcal{S}^{k+1} = \{(w, e_1, \ldots, e_{n-1}) \mid w \in W,\ e_j \in \mathcal{E}^k \text{ for all } j\}$;
- $\mathcal{E}^{k+1} = \{e^{k+1} \mid e^{k+1} \subseteq \mathcal{S}^{k+1}\}$.

We call $s^k \in \mathcal{S}^k$ a k-structure and $e^k \in \mathcal{E}^k$ a k-state. We call k the depth of s^k and e^k. Intuitively, a structure consists of what an agent considers possible about the world and about other agents' beliefs. When the context is clear, we omit the superscript k. By a model, we mean a tuple $(w, e_1, \ldots e_n)$ (also written as $(w, \vec{e})$ for simplicity), where w is a possible-world and e_i is the epistemic state of agent i. We use $\mathbb{M}$ to denote the set of all models. Note the difference between a structure and a model: a structure consists of $n-1$ states of the same depth, while a model consists of n states whose depths can be different. For state $\hat{e}$ and structure $s = (w, e_1, \ldots, e_{n-1})$, we use $s \cup_i \hat{e}$ to denote the model obtained by inserting $\hat{e}$ at the i-th position, i.e. $(w, e_1, \ldots e_{i-1}, \hat{e}, e_i, \ldots e_{n-1})$. The satisfaction relation is defined as follows:

- $(w, \vec{e}) \models p$ iff $p \in w$;
- $(w, \vec{e}) \models \neg\alpha$ iff it is not the case that $(w, \vec{e}) \models \alpha$;
- $(w, \vec{e}) \models \alpha \wedge \beta$ iff $(w, \vec{e}) \models \alpha$ and $(w, \vec{e}) \models \beta$;

- $(w, \vec{e}) \models \boldsymbol{K}_i\varphi$ iff for all $s \in e_i$, $s \cup_i e_i \models \varphi$.

We say a model and a formula α is *compatible* if, for any $i \in \mathsf{Ag}$, the i-depth of α is not deeper than the epistemic state of agent i. Given a finite set $\Sigma \subseteq \mathcal{ML}_n$,[2] $\alpha \in \mathcal{ML}_n$, we say Σ entails α (written as $\Sigma \models \alpha$) iff for any model $(w, \vec{e})$ compatible with each formula in $\Sigma \cup \{\alpha\}$, if $(w, \vec{e}) \models \varphi$ for all $\varphi \in \Sigma$, then $(w, \vec{e}) \models \alpha$. We say α is *valid* (written as $\models \alpha$) iff $\{\} \models \alpha$. We say a formula α is *satisfiable* if there is a compatible model $(w, \vec{e})$ s.t. $(w, \vec{e}) \models \alpha$.
When α is objective, we write $w \models \alpha$ instead of $(w, \vec{e}) \models \alpha$. When α is i-subjective, we write $e_i \models \alpha$. Note that this account cannot handle unbounded depth of formulae. When Σ is an infinite set of formulae with unbounded depth, i.e. for any k it exists $\psi \in \Sigma$ s.t. $dep[\psi] > k$, then $\Sigma \models \alpha$ is not well-defined.

Example 5.3. Suppose that $\mathsf{Ag} = \{1, 2\}$, let $p, q, r \in \mathsf{Ap}$, $w_0 = \{r\}$, $w_1 = \{p, q\}$. We define two states of depth 1 for agent 2:

$$e_2 = \{(w_0, \{\varnothing\}), (w_1, \{\varnothing\})\} \qquad e_2' = \{(w_1, \{\varnothing\})\}$$

and a state of depth 2 for agent 1: $e_1 = \{(w_0, e_2), (w_0, e_2')\}$. Consider the formula in Example 5.2, we have $e_1 \models \boldsymbol{K}_1(\neg p \wedge \neg \boldsymbol{K}_2(\neg p))$.

A model in $\mathcal{ML}_n$ cannot assign truth values to all formulae, mainly because the depth of the model is bounded. For instance, in the above example the state $e_1' = \{(w_0, \{\varnothing\})\}$ can not determine the truth of formula $\boldsymbol{K}_1(\neg p \wedge \neg \boldsymbol{K}_2(\neg p))$. We argue that this is not a problem because for any formula α, any model $(w, \vec{e})$ can be extended to a deeper one that is compatible with α while preserving the truth value of all compatible formulae. To understand the relation among models of different depths, we introduce the notion of regression:

Definition 5.4 (Regression)**.** The regression of structures and epistemic states is inductively defined as follows:

[2] Σ can also be an infinite set, provided that the depth of formulae is bounded, i.e. it exists $k \in \mathbb{N}$ s.t. $dep[\varphi] \leq k$ for all $\varphi \in \Sigma$.

- For $s \in \mathcal{S}^1$, i.e. $s = (w, \{\varnothing\}, \ldots, \{\varnothing\})$, the regression of s, written as $s\downarrow$, is $\varnothing$;
- For $e \in \mathcal{E}^{k+1}$, $e' \in \mathcal{E}^k$, we say e' is the regression of e, written as $e' = e\downarrow$, if and only if $e' = \{s\downarrow \mid s \in e\}$;
- For $s \in \mathcal{S}^{k+1}$, $s' \in \mathcal{S}^k$, s' is the regression of s, written as $s' = s\downarrow$, if and only if $s = (w, \vec{e})$ and $s' = (w, \vec{e}')$ where $e'_j = e_j\downarrow$.

As shown in the following lemma, regression preserves the truth values of compatible formulae.

Lemma 5.5. For $i \in \mathsf{Ag}$, $\vec{e}$ and $\vec{e}'$ s.t. $e'_i = e_i\downarrow$ and $e_j = e'_j$ for any $j \neq i$. For any $w \in W$ and α compatible with $\vec{e}'$ and $\vec{e}$,

$$(w, \vec{e}) \models \alpha \text{ iff } (w, \vec{e}') \models \alpha$$

A proof is given in Lem. 8 of [BL15b].

5.1.2. Comparison with Kripke Semantics

We briefly recap Kripke semantics to compare with ours. See [Lad77, HM92, HM90] for more details on Kripke semantics.

Definition 5.6 (Kripke structure)**.** A Kripke structure for a set of agents Ag is a tuple $M = (\Omega, R_1, \ldots, R_n, \pi)$, where Ω is a set of Kripke worlds.[3] For $i \in \mathsf{Ag}$, R_i is a binary relation on the worlds in Ω. π assigns each $\omega \in \Omega$ a set of propositions considered to be true under ω, i.e. $\pi(\omega) \subseteq \mathsf{Ap}$.

To distinguish from the set-based semantics, we use $\models_K$ to denote the satisfaction relation in Kripke's semantics. Given a Kripke model M and a world $\omega \in \Omega$, the satisfaction relation is defined as follows:

- $(M, \omega) \models_K p$ iff $p \in \pi(\omega)$;

[3] Note the difference between Kripke's worlds and Levesque's. To avoid confusion, we use the Latin letter w for Levesque's world and the Greek letter ω for Kripke's.

- $(M, \omega) \models_K \neg\alpha$ iff it is not the case that $(M, \omega) \models_K \alpha$;
- $(M, \omega) \models_K \alpha \wedge \beta$ iff $(M, \omega) \models_K \alpha$ and $(M, \omega) \models_K \beta$;
- $(M, \omega) \models_K \boldsymbol{K}_i\alpha$ iff $(M, \omega') \models_K \alpha$ for all ω' s.t. $\omega R_i \omega'$.

The relation between frame conditions and the axiom systems of different modal logics are investigated in [Kri63, Che80], for example. We only introduce some conditions related to our account:

- *Transitivity*: If $\omega R_i \omega'$ and $\omega' R_i \omega''$, then $\omega R_i \omega''$.
- *Euclideanity*: If $\omega R_i \omega'$ and $\omega R_i \omega''$, then $\omega' R_i \omega''$.

Theorem 5.7 (Halpern [HM92])**.** A Kripke structure M is a model of $K45_n$ iff all relations R_i are Euclidean and transitive.

Intuitively, $K45_n$ axiom system requires both *positive* and *negative introspection*. Namely, for any α and agent $i \in \mathsf{Ag}$,

- $\models_K \boldsymbol{K}_i\alpha \supset \boldsymbol{K}_i\boldsymbol{K}_i\alpha$
- $\models_K \neg\boldsymbol{K}_i\alpha \supset \boldsymbol{K}_i\neg\boldsymbol{K}_i\alpha$

Given a $\mathcal{ML}_n$ model $(w, \vec{e})$, we show that it amounts to a tree-like exploration of a Kripke model up to a finite depth. Therefore it does not utilize the global transitivity or Euclideanity of all worlds. This is part of the reason why the translation can avoid axioms on frame conditions. For Kripke model $M = (\Omega, R_1, \ldots, R_n, \pi)$, we define a set of mappings from Ω to structures and models in $\mathcal{ML}_n$:

Definition 5.8. Given $M = (\Omega, R_1, \ldots, R_n, \pi)$, $\sigma \in \mathsf{Ag}^*$, $\omega \in \Omega$,

- For $|\sigma| = k$, $\phi_{M,k}^{(\sigma \cdot i)}(\omega) = \varnothing$;
- For $|\sigma| < k$, $\phi_{M,k}^{(\sigma \cdot i)}(\omega) = (w, e_1, \ldots, e_{i-1}, e_{i+1}, \ldots, e_n)$, where $w = \pi(\omega)$, $e_j = \{\phi_{M,k}^{(\sigma \cdot i \cdot j)}(\omega') \mid \text{ for any } \omega' \text{ s.t. } \omega R_j \omega'\}$;
- $\Phi_{M,k}(\omega) = (\pi(\omega), \vec{e})$, where $e_i = \{\phi_{M,k}^{(i)}(\omega') \mid \omega R_i \omega'\}$.

When the context is clear, we fix the Kripke model and omit M for short. Here σ is a finite word which uses Ag as the alphabet.[4] We use $start(\sigma)$ and $end(\sigma)$ to denote the first and the last letter in σ. The length of σ is denoted as $|\sigma|$. For any σ, e.g. $\sigma = 1 \cdot 2$, the mapping $\phi_k^{(\sigma)}(\omega)$ returns a structure, which agent 1 believes that agent 2 considers as possible. Clearly, the mappings are defined recursively from longer words to their prefix.

Lemma 5.9. For any $\omega \in \Omega$, $k \in \mathbb{N}$ and $\sigma \in \mathsf{Ag}^+$,

- For any $j \neq start(\sigma)$, $\phi_k^{(\sigma)}(\omega) = \phi_{k+1}^{(j\cdot\sigma)}(\omega)$
- $\phi_{k+1}^{(\sigma)}(\omega) \downarrow= \phi_k^{(\sigma)}(\omega)$

Proof. When $k - |\sigma| = 0$, the proof is trivial. Suppose that the lemma holds for any k' and $|\sigma'|$ s.t. $k' - |\sigma'| < m$. For k, σ s.t. $k - |\sigma| = m$, let $\phi_k^{(\sigma)}(\omega) = (w, e_1, \ldots, e_{i-1}, e_{i+1}, \ldots, e_n)$, then for all $l \neq end(\sigma)$, by the induction hypothesis we have

$$e_l = \{\phi_k^{(\sigma\cdot l)}(\omega') \mid \omega R_l \omega'\} = \{\phi_{k+1}^{(j\cdot\sigma\cdot l)}(\omega') \mid \omega R_l \omega'\}$$
$$e_l = \{\phi_k^{(\sigma\cdot l)}(\omega') \mid \omega R_l \omega'\} = \{\phi_{k+1}^{(\sigma\cdot l)}(\omega') \downarrow \mid \omega R_l \omega'\}$$

Thus $\phi_k^{(\sigma)}(\omega) = \phi_{k+1}^{(j\cdot\sigma)}(\omega)$ and $\phi_{k+1}^{(\sigma)}(\omega) \downarrow= \phi_k^{(\sigma)}(\omega)$. □

With the above mappings, we show the bi-simulation between Kripke models and $\mathcal{ML}_n$ models up to any finite depth.

Lemma 5.10. For any $K45_n$ Kripke model M, $\omega \in \Omega$ and $k \in \mathbb{N}^+$, let $\Phi_k(\omega)$ be defined as in Def. 5.8. For any formula α s.t. $dep[\alpha] \leq k$,

$$(M, \omega) \models_K \alpha \text{ iff } \Phi_k(\omega) \models \alpha \tag{5.1}$$

Proof. We prove it via induction on the structure of α.

[4] We only consider σ, where the adjacent letters are distinct i.e. $\sigma \notin (\mathsf{Ag}^* \cdot i \cdot i \cdot \mathsf{Ag}^*)$ for any $i \in \mathsf{Ag}$.

- Base case: Given $\alpha := p$ for $p \in \mathsf{Ap}$, then for any k

$$\begin{aligned}
(M, \omega) \models_K p &\iff p \in \pi(\omega) && \text{(Kripke semantics)} \\
&\iff p \in w \text{ for } (w, \vec{e}) = \Phi_k(\omega) && \text{(Def. } \Phi_k\text{)} \\
&\iff \Phi_k(\omega) \models p && (\mathcal{ML}_n \text{ semantics})
\end{aligned}$$

- Induction on $\wedge$ and $\neg$ is straightforward.

- Induction on $\boldsymbol{K}_i$: suppose that Eq. 5.1 holds for all ω and k. Let $(w, \vec{e}) = \Phi_{k+1}(\omega)$,

$$\begin{aligned}
& (M, \omega) \models_K \boldsymbol{K}_i \alpha \\
\iff & (M, \omega') \models_K \alpha \text{ for all } \omega' \text{ s.t. } \omega R_i \omega' && \text{(Kripke semantics)} \\
\iff & \Phi_k(\omega') \models \alpha \text{ for all } \omega' \text{ s.t. } \omega R_i \omega' && \text{(I.H.)} \\
\iff & \text{for all } s \in e_i, s \cup_i e'_i \models \alpha && (\#) \\
\iff & \text{for all } s \in e_i, s \cup_i e_i \models \alpha && (e'_i = e_i \downarrow, \text{Lem.5.5}) \\
\iff & (w, \vec{e}) \models \boldsymbol{K}_i \alpha && (\mathcal{ML}_n \text{ semantics})
\end{aligned}$$

To explain step (#):
($\Rightarrow$) For each $s \in e_i$, $s = \phi^{(i)}_{k+1}(\omega')$ with some ω' s.t. $\omega R_i \omega'$. By Lem.5.9, $s \cup_i e'_i = \Phi_k(\omega')$.
($\Leftarrow$) Consider a fixed ω' s.t. $\omega R_i \omega'$ and $(w', \vec{e}') = \Phi_k(\omega')$. By definition, $e'_i = \{\phi^{(i)}_k(\omega'') \mid \omega' R_i \omega''\}$. Since R_i is transitive and Euclidean, for any $\tilde{\omega}$, $\phi^{(i)}_{k+1}(\tilde{\omega}) \in e_i$ iff $\phi^{(i)}_k(\tilde{\omega}) \in e'_i$. Thus e'_i is the regression of e_i, i.e. $e_i \downarrow = e'_i$. For $j \neq i$, by Lem. 5.9 we have $e'_j = \{\phi^{(i \cdot j)}_{k+1}(\omega'') \mid \omega' R_j \omega''\}$. Let $s = (w', e'_1, \ldots, e'_{i-1}, e'_{i+1}, \ldots e'_n)$, we have $s = \phi^{(i)}_{k+1}(\omega')$ and by definition of e_i, $s \in e_i$. Namely, $\Phi_k(\omega') = s \cup_i e'_i$ for some $s \in e_i$. Thus $\Phi_k(\omega') \models \alpha$. □

Properties of the Logic

$\mathcal{ML}_n$ precisely captures the $K45_n$ properties in the sense that $K45_n$ modal system forms a sound and complete axiomatization for $\mathcal{ML}_n$. The soundness is implied

by the following theorem:

Theorem 5.11. For any $\alpha, \beta \in \mathcal{ML}_n$ and $i \in \mathsf{Ag}$,

- (Prop) If α is a propositional tautology, then $\models \alpha$
- (MP) If $\models \alpha$ and $\models \alpha \supset \beta$ then $\models \beta$
- (Dist) $\models (\boldsymbol{K}_i\alpha \wedge \boldsymbol{K}_i(\alpha \supset \beta)) \supset \boldsymbol{K}_i\beta$
- (Nec) If $\models \alpha$ then $\models \boldsymbol{K}_i\alpha$
- (4) $\models \boldsymbol{K}_i\alpha \supset \boldsymbol{K}_i\boldsymbol{K}_i\alpha$
- (5) $\models \neg\boldsymbol{K}_i\alpha \supset \boldsymbol{K}_i\neg\boldsymbol{K}_i\alpha$

Proving completeness is equivalent to proving that every $K45_n$-consistent formula is $\mathcal{ML}_n$-satisfiable. By [HM92] every $K45_n$-consistent formula α is satisfiable in some Kripke structure with transitive, Euclidean relations. Then by Lem. 5.10 we prove that α is $\mathcal{ML}_n$-satisfiable. With the soundness and completeness, we have the following result:

Theorem 5.12. α is $K45_n$-consistent iff α is $\mathcal{ML}_n$-satisfiable .

5.1.3. Adaptation for Other Modal Systems

In Kripke's semantics, the consistency of belief and the truthfulness of knowledge can be interpreted as the seriality and the reflexivity of the accessible relations. We show that in our set-based semantics, they amount to considering a subset of $\mathcal{ML}_n$ models.

The consistency of beliefs, also known as the D-property, requires that for any $i \in \mathsf{Ag}$ and formula α in the language,

$$\models \boldsymbol{K}_i\alpha \supset \neg\boldsymbol{K}_i\neg\alpha$$

It can be achieved by ruling out models with empty (sub-)structures.

Definition 5.13 (D-state)**.** The set of D-states is defined as follows:

- If $e \in \mathcal{E}^1$, then e is a D-state iff $e \neq \{\}$.
- If $e \in \mathcal{E}^k$, where $k > 1$, then e is a D-state iff $e \neq \{\}$ and for all $(w', e'_1, \ldots, e'_{i-1}, e'_{i+1}, \ldots, e'_n) \in e$ and $j \neq i$, e'_j is a D-state.

We call $(w, \vec{e})$ a D-model if for all $i \in \mathsf{Ag}$, e_i is a D-state. We say a formula α is $\mathcal{ML}_n^D$-satisfiable iff α is satisfied in some D-models.

The truthfulness of knowledge requires that for $i \in \mathsf{Ag}$ and α,

$$\models \boldsymbol{K}_i\alpha \supset \alpha$$

Namely, the actual world is always considered possible, and the actual epistemic state of an agent is always considered possible by any other agents. We say a model $(w, \vec{e})$ is *homogeneous* if all e_i are of the same depth, i.e. there is a k s.t. $e_i \in \mathcal{E}^k$ for all $i \in \mathsf{Ag}$.

Definition 5.14. A T-model is a homogeneous model $(w, \vec{e})$ s.t.

- $e_i = \{\varnothing\}$ for all $i \in \mathsf{Ag}$, or
- F.a. $i \in \mathsf{Ag}$, $s \in e_i$, $s \cup_i (e_i \downarrow)$ is a T-model and $(w, e_1 \downarrow, \ldots, e_{i-1} \downarrow, e_{i+1} \downarrow, \ldots, e_n \downarrow) \in e_i$

We say α is $\mathcal{ML}_n^T$-satisfiable iff α is satisfied in some T-models.

Theorem 5.15. For any $\alpha \in \mathcal{ML}_n$,

- α is $KD45_n$-consistent iff α is $\mathcal{ML}_n^D$-satisfiable;
- α is $S5_n$-consistent iff α is $\mathcal{ML}_n^T$-satisfiable.

The proof is analogous to Thm. 5.12. Soundness is proved by the validity of axioms (similar to Thm. 5.11) and completeness is proved by constructing a D-model (T-model) given a Kripke frame and the formula to be satisfied (similar to Lem. 5.10).

5.1.4. Translation

In this section, we propose a new approach to translate modal formulae into first-order ones. The translation is mainly based on explicitly specifying the inclusion relation of $\mathcal{ML}_n$ structures. The form is similar to the standard translation, yet some essential differences exist: Binary predicates of the form E_σ are used to represent the inclusion relation instead of the accessibility of Kripke states. For $\sigma \in \mathsf{Ag}^*$ s.t. $end(\sigma) = i$, $E_\sigma(u, v)$ means that the epistemic state at the i-th position of structure v contains structure u. We prove that such a translation preserves satisfiability, without explicitly expressing the frame conditions as the standard translation does. For any $\mathcal{ML}_n$ formula, the translation results in a first-order formula in the GF^2 fragment.

Translation Function

We define a recursive translation function $\mathcal{R}[\alpha, \sigma, u, v]$. Here α is the $\mathcal{ML}_n$ formula to be translated, $\sigma \in \mathsf{Ag}^*$ is a finite word, u, v are variables used for the translation.

Definition 5.16 (translation function)**.** Given $\sigma \in \mathsf{Ag}^*$,

1. For $p \in \mathsf{Ap}$, $\mathcal{R}[p, \sigma, u, v] := P(u)$
2. $\mathcal{R}[\neg\alpha, \sigma, u, v] := \neg\mathcal{R}[\alpha, \sigma, u, v]$
3. For $\odot \in \{\wedge, \vee, \supset, \equiv\}$,
   $\mathcal{R}[\alpha_1 \odot \alpha_2, \sigma, u, v] := \mathcal{R}[\alpha_1, \sigma, u, v] \odot \mathcal{R}[\alpha_2, \sigma, u, v]$
4. If $end(\sigma) = i$,
   $\mathcal{R}[\boldsymbol{K}_i\alpha, \sigma, u, v] := \forall u\ E_\sigma(u, v) \supset (\mathcal{R}[\alpha, \sigma, u, v])$
5. If $end(\sigma) \neq i$,
   $\mathcal{R}[\boldsymbol{K}_i\alpha, \sigma, u, v] := \forall v\ E_{\sigma \cdot i}(v, u) \supset (\mathcal{R}[\alpha, \sigma \cdot i, v, u])$

Where P is a unary predicate associated with proposition p.

Example 5.17. Let $\alpha := \boldsymbol{K}_1(\neg p \wedge \neg \boldsymbol{K}_2(\neg p))$ The translation of α is

$$\begin{aligned}
&\mathcal{R}[\alpha, \epsilon, u, v] \\
=&\mathcal{R}[\boldsymbol{K}_1(\neg p \wedge \neg \boldsymbol{K}_2(\neg p)), \epsilon, u, v] \\
=&\forall v\ E_1(v, u) \supset \mathcal{R}[\neg p \wedge \neg \boldsymbol{K}_2(\neg p), 1, v, u] \\
=&\forall v\ E_1(v, u) \supset (\neg P(v) \wedge \neg \forall u(E_{1\cdot 2}(u, v) \supset \mathcal{R}[\neg p, 1 \cdot 2, u, v])) \\
=&\forall v\ E_1(v, u) \supset (\neg P(v) \wedge \neg \forall u(E_{1\cdot 2}(u, v) \supset \neg P(u))
\end{aligned}$$

For $KD45_n$ system, we define function $\mathcal{R}_D$ which is identical to $\mathcal{R}$ except for items (4) and (5):

(4') If $end(\sigma) = i$, $\mathcal{R}_D[\boldsymbol{K}_i\alpha, \sigma, u, v] :=$
$(\forall u\ E_\sigma(u, v) \supset (\mathcal{R}_D[\alpha, \sigma, u, v])) \wedge \exists u\ E_\sigma(u, v)$

(5') If $end(\sigma) \neq i$, $\mathcal{R}_D[\boldsymbol{K}_i\alpha, \sigma, u, v] :=$
$(\forall v\ E_{\sigma\cdot i}(v, u) \supset (\mathcal{R}_D[\alpha, \sigma \cdot i, v, u])) \wedge \exists v\ E_{\sigma\cdot i}(v, u)$

For $S5_n$, we define $\mathcal{R}_T$ which also differs only on (4) and (5):

(4") If $end(\sigma) = i$, $\mathcal{R}_T[\boldsymbol{K}_i\alpha, \sigma, u, v] :=$
$(\forall u\ E_\sigma(u, v) \supset (\mathcal{R}_D[\alpha, \sigma, u, v])) \wedge E_\sigma(v, v)$

(5") If $end(\sigma) \neq i$, $\mathcal{R}_D[\boldsymbol{K}_i\alpha, \sigma, u, v] :=$
$(\forall v\ E_{\sigma\cdot i}(v, u) \supset (\mathcal{R}_D[\alpha, \sigma \cdot i, v, u])) \wedge E_{\sigma\cdot i}(u, u)$

Construction of Canonical FOL Model

To prove that $\mathcal{ML}_n$ can be embedded into FOL, we construct a canonical FOL model as follows:

Definition 5.18. We define $M^c = \langle \mathcal{D}^c, I \rangle$ as follows:

C1 Domain of discourse $\mathcal{D}^c$ is the set of $\mathcal{ML}_n$ models $\mathbb{M}$.

C2 For $P \in \mathsf{Ap}$ and $(w, \vec{e}) \in \mathcal{D}^c$, $(w, \vec{e}) \in P^{M^c}$ iff $P \in w$

C3 For $\omega, \omega' \in \mathcal{D}^c$, $\sigma \in \mathsf{Ag}^+$ s.t. $end(\sigma) = i$. $(\omega', \omega) \in E_\sigma^{M^c}$ iff
$\omega = (w, \vec{e})$ and $\omega' = s \cup_i e_i$ for some $s \in e_i$

Since C1-C3 specifies different aspects of the model, the existence of the canonical model is obvious. For each satisfiable formula in $\mathcal{ML}_n$, we show the satisfiability of the resulting formula.

Theorem 5.19. For any compatible model $(w, \vec{e})$ and sentence α,

1. $w, \vec{e} \models \alpha$ iff $M^c, \mu \models_{FO} \mathcal{R}[\alpha, \epsilon, u, v]$ for $\mu(u) = (w, \vec{e})$.

2. If $(w', \vec{e}')$ exists s.t. $(w, \vec{e}) = s \cup_i e_i'$ for some $s \in e_i$, then

$$w, \vec{e} \models \alpha \text{ iff } M^c, \mu \models_{FO} \mathcal{R}[\alpha, \sigma \cdot i, u, v]$$

Where $\mu(u) = (w, \vec{e})$ and $\mu(v) = (w', \vec{e}')$.

Proof. We prove this via induction on the formula structure. As for the base case, let $p \in \mathsf{Ap}$, $\sigma \in \mathsf{Ag}^*$

$$\begin{aligned}
w, \vec{e} \models p \Longleftrightarrow & p \in w && (\mathcal{ML}_n \text{ semantics}) \\
\Longleftrightarrow & \mu(u) \in P^{M^c} && (\text{C2}) \\
\Longleftrightarrow & M^c, \mu \models P(u) && (\text{FOL semantics}) \\
\Longleftrightarrow & M^c, \mu \models \mathcal{R}[p, \sigma, u, v] && (\text{Def. } \mathcal{R}[\cdot])
\end{aligned}$$

The induction on $\neg, \wedge$ is straightforward. For induction on the $\boldsymbol{K}_i$-operator where $end(\sigma) \neq i$, as hypothesis we assume that for any $(w, \vec{e})$, $s \in e_i$,

$$s \cup_i e_i \models \alpha \text{ iff } M^c, \mu' \models \mathcal{R}[\alpha, \sigma \cdot i, u', u] \tag{5.2}$$

Where $\mu'(u') = (s \cup_i e_i)$, $\mu'(u) = (w, \vec{e})$, then

$$
\begin{aligned}
& w, \vec{e} \models \boldsymbol{K}_i\alpha \\
\Longleftrightarrow & \text{for all } s' \in e_i, s' \cup_i e_i \models \alpha && (\mathcal{ML}_n \text{ semantics}) \\
\Longleftrightarrow & \text{for all } s' \in e_i \text{ and } \mu' \text{ s.t. } \mu'(u') = (s' \cup_i e_i), \\
& \mu'(u) = (w, \vec{e}), M^c, \mu' \models \mathcal{R}[\alpha, \sigma \cdot i, u', u] && (\text{ I.H.}) \\
\Longleftrightarrow & \text{for all } \mu' \text{ s.t. } \mu'(u) = (w, \vec{e}) \text{ and } M^c, \mu' \models E_{\sigma\cdot i}(u', u), \\
& \qquad M^c, \mu' \models \mathcal{R}[\alpha, \sigma \cdot i, u', u] && (\text{ cond. C3}) \\
\Longleftrightarrow & M^c, \mu \models \forall u' E_{\sigma\cdot i}(u', u) \supset \mathcal{R}[\alpha, \sigma \cdot i, u', u] && (\text{FOL semantics}) \\
\Longleftrightarrow & M^c, \mu \models \mathcal{R}[\boldsymbol{K}_i\alpha, \sigma, u, v] && (\text{Def. } \mathcal{R}[\cdot])
\end{aligned}
$$

Considering the case $\sigma = \epsilon$, then it is proved that for $\mu(u) = (w, \vec{e})$,

$$w, \vec{e} \models \boldsymbol{K}_i\alpha \text{ iff } M^c, \mu \models \mathcal{R}[\boldsymbol{K}_i\alpha, \epsilon, u, v], \tag{5.3}$$

which is part (1) of the theorem. To complete the induction on $\boldsymbol{K}_i$ for part (2), it remains to prove that for any $\sigma \in \mathsf{Ag}^*$,

$$w, \vec{e} \models \boldsymbol{K}_i\alpha \text{ iff } M^c, \mu \models_{FO} \mathcal{R}[\boldsymbol{K}_i\alpha, \sigma \cdot i, u, v] \tag{5.4}$$

$$
\begin{aligned}
& M^c, \mu \models \mathcal{R}[\boldsymbol{K}_i\alpha, \sigma \cdot i, u, v] \\
\Longleftrightarrow & M^c, \mu \models \forall u'.E_{\sigma\cdot i}(u', v) \supset \mathcal{R}[\alpha, \sigma \cdot i, u', v] && (\text{Def. } \mathcal{R}[\cdot]) \\
\Longleftrightarrow & M^c, \mu' \models E_{\sigma\cdot i}(u', v) \supset \mathcal{R}[\alpha, \sigma \cdot i, u', v] \text{ for all } \mu' \sim_{u'} \mu && (\text{FOL semantics}) \\
\Longleftrightarrow & \text{for all } \mu' \sim_{u'} \mu, M^c, \mu' \models E_{\sigma\cdot i}(u', v) \Rightarrow M^c, \mu' \models \mathcal{R}[\alpha, \sigma \cdot i, u', v] \\
\Longleftrightarrow & \text{for all } \mu' \sim_{u'} \mu, \text{ if } \mu'(u') = (s \cup_i e_i) \text{ for some } s \in e_i \\
& \qquad \text{then } M^c, \mu' \models \mathcal{R}[\alpha, \sigma \cdot i, u', v] && (\text{cond. C3, } e'_i = e_i) \\
\Longleftrightarrow & \text{for all } \mu' \sim_{u'} \mu, \text{ if } \mu'(u') = (s \cup_i e_i) \text{ for some } s \in e_i \\
& \qquad \text{then } s \cup_i e_i \models \alpha && (\text{I.H case 2}) \\
& \text{Since for all } s \in e_i, \mu' \text{ exists s.t. } \mu' \sim_{u'} \mu \text{ and } \mu'(u') = s \cup_i e_i \\
\Longleftrightarrow & (s \cup_i e_i) \models \alpha \text{ for all } s \in e_i \\
\Longleftrightarrow & w, \vec{e} \models \boldsymbol{K}_i\alpha && (\mathcal{ML}_n \text{ semantics})
\end{aligned}
$$

□

Preserving Satisfiability and Decidability

Thm. 5.19 implies that if the original formula is $\mathcal{ML}_n$-satisfiable, then the translation is FO-satisfiable. In particular, it is satisfied by the canonical model M^c. We are also interested in the other direction, namely, when the translation is satisfied in some model M (not necessarily the canonical one), whether a model for the modal formula exists. We construct a set of mappings from the domain of M to structures and models of $\mathcal{ML}_n$. It is similar to the proof of Lem. 5.10: Given a FO model M, element $\omega, \omega' \in \mathcal{D}^M$ and $k > 0$, we define structure mappings $\phi^{(\sigma)}_{M,k}$ and model mappings $\Phi^{(\sigma)}_{M,k}$.

- When $|\sigma| = k$, $\phi^{(\sigma\cdot i)}_{M,k}(\omega) = \varnothing$;
- When $|\sigma| < k$, $\phi^{(\sigma\cdot i)}_{M,k} = (w, e_1, \ldots, e_{i-1}, e_{i+1}, \ldots, e_n)$, where $w = \{p \,|\, \omega \in P^M\}$, $e_j = \{\phi^{(\sigma\cdot i\cdot j)}_{M,k}(\omega') \,|\, (\omega', \omega) \in E^M_{\sigma\cdot i\cdot j}\}$.

Given $\omega, \omega' \in \mathcal{D}^M$, we define $\Phi^{(\sigma)}_{M,k}(\omega, \omega') = (w, \vec{e})$ as follows

- $w = \{p \in \mathsf{Ap} \,|\, \omega \in P^M\}$
- For any $j \neq end(\sigma)$, $e_j = \{\phi^{(\sigma\cdot j)}_{M,k}(\omega'') \,|\, (\omega'', \omega) \in E^M_{\sigma\cdot j}\}$
- For $i = end(\sigma)$, $e_i = \{\phi^{(\sigma)}_{M,k}(\omega'') \,|\, (\omega'', \omega') \in E^M_{\sigma}\}$

For simplicity, we omit the subscripts M, k when the context is clear. When $\sigma = \epsilon$, the model is completely determined by ω. By definition of $\phi^{(\sigma)}$ and $\Phi^{(\sigma)}$, we have the following properties:

Lemma 5.20. $(\omega, \omega') \in E^M_{\sigma\cdot i}$ if and only if $\Phi^{(\sigma\cdot i)}(\omega, \omega') = s \cup_i e_i$ for some $s \in e_i$ and $e_i = \{\phi^{(\sigma\cdot i)}(\omega'') \,|\, (\omega'', \omega') \in E^M_{\sigma\cdot i}\}$

As we will show, given that (M, μ) satisfies the translating formula $\mathcal{R}[\alpha, \epsilon, u, v]$, the original formula will be satisfied by the $\mathcal{ML}_n$ model $(w, \vec{e}) = \Phi^{(\epsilon)}(\mu(u), \mu(v))$.

Lemma 5.21. Given $\alpha \in \mathcal{ML}_n$ and $\sigma \in \mathsf{Ag}^*$, Then for μ such that $\mu(u) = \omega$, $\mu(v) = \omega'$,

1. If $(w, \vec{e}) = \Phi^{(\epsilon)}(\omega, \omega')$, then

$$M, \mu \models \mathcal{R}[\alpha, \epsilon, u, v] \text{ iff } (w, \vec{e}) \models \alpha$$

2. If $(w, \vec{e}) = \Phi^{(\sigma \cdot i)}(\omega, \omega')$ and $(\omega, \omega') \in E^M_{\sigma \cdot i}$

$$M, \mu \models \mathcal{R}[\alpha, \sigma \cdot i, u, v] \text{ iff } (w, \vec{e}) \models \alpha$$

Proof. Since induction proofs with similar spirit are shown repeatedly, we only give the induction step for $\boldsymbol{K}$-operators.
For each $i \in \mathsf{Ag}$, when $i \neq end(\sigma)$,

$$\begin{aligned}
& M, \mu \models \mathcal{R}[\boldsymbol{K}_i\alpha, \sigma, u, v] \\
\Longleftrightarrow & M, \mu \models \forall v\ E_{\sigma \cdot i}(v, u) \supset \mathcal{R}[\alpha, \sigma \cdot i, v, u] && \text{(Def. } \mathcal{R}[\cdot]) \\
\Longleftrightarrow & \text{ for all } \mu' \sim_v \mu, \text{ if } (\mu'(v), \mu'(u)) \in E^M_{\sigma \cdot i} \text{ then} \\
& M, \mu' \models \mathcal{R}[\alpha, \sigma \cdot i, v, u] && \text{(FOL semantics)} \\
\Longleftrightarrow & \text{ for all } \mu' \sim_v \mu, \text{ if } (\mu'(v), \mu'(u)) \in E^M_{\sigma \cdot i} \text{ then} \\
& w', \vec{e}' \models \alpha \text{ for } (w', \vec{e}') = \Phi^{(\sigma \cdot i)}(\mu'(v), \mu'(u)) && \text{(I.H. case 2)} \\
\Longleftrightarrow & \text{ for all } s \in e_i, s \cup_i e_i \models \alpha && (\text{ Lem. 5.20}) \\
\Longleftrightarrow & e_i \models \boldsymbol{K}_i\alpha && (\mathcal{ML}_n \text{ semantics}) \\
\Longleftrightarrow & w, \vec{e} \models \boldsymbol{K}_i\alpha
\end{aligned}$$

Part (1) is covered when $\sigma = \epsilon$. Analogously we proof the induction for the case when $i = end(\sigma)$.

□

We introduce a constant c and define $\mathcal{T}[\alpha]$ as a sentence replacing every free appearance of u in $\mathcal{R}[\alpha, \epsilon, u, v]$ by c.

Theorem 5.22. α is $\mathcal{ML}_n$-satisfiable iff $\mathcal{T}[\alpha]$ is FO-satisfiable.

Proof. Suppose that α is $\mathcal{ML}_n$-satisfiable, by Thm. 5.19 M^c satisfies $\mathcal{R}[\alpha, \epsilon, u, v]$ with some μ. By Skolemization, we replace the free variable u with constant c and it results in $\mathcal{T}[\alpha]$ (v never occurs in $\mathcal{R}[\alpha, \epsilon, u, v]$ as a free variable so it can be ignored). For the other direction, suppose that $\mathcal{T}[\alpha]$ is FO-satisfiable, then $\mathcal{R}[\alpha, \epsilon, u, v]$ is satisfied in some (M, μ). By Lem. 5.21, α is $\mathcal{ML}_n$-satisfiable. □

With satisfiability-preservation, the properties of validity checking and theorem proving can also be derived:

Corollary 5.23. Given finite Σ and $\alpha \in \mathcal{ML}_n$,

- $\Sigma \models \alpha$ iff $\mathcal{T}[\Sigma] \models \mathcal{T}[\alpha]$;
- α is valid iff $\mathcal{T}[\alpha]$ is FO-valid.

We also show that the translation will convert any $\mathcal{ML}_n$ formula to a decidability-preserving formula in linear time and space.

Theorem 5.24. For any $\alpha \in \mathcal{ML}_n$, $\mathcal{T}[\alpha]$ can be computed in $O(|\alpha|)$, $\mathcal{T}[\alpha] \in GF^2$ and the size of $\mathcal{T}[\alpha]$ is linear in $|\alpha|$.

Proof. Let $\varphi = \mathcal{T}[\alpha]$, it is obvious that $\varphi \in FO^2$ since the translation only uses variables u, v. By induction on the structure of φ we prove $\varphi \in GF$. As for the base case, $\mathcal{R}[P, \sigma, u, v] = P(u)$, which is guarded. Negation or conjunction does not affect the guarded condition. Quantifiers only occur in translating a (sub-)formula of form $\boldsymbol{K}_i\alpha$. Apparently, the resulting formula has a guard E_σ containing all free variables in the conclusion. Linear complexity in terms of time and size is straightforward. By Def. 5.16, the number of translation steps is linear in $|\alpha|$, and for each step, only a constant number of symbols will be added to the resulting formulae. □

Since the translation results in GF^2, as the straightforward result, the satisfiability of the resulting formulae is decidable.

Corollary 5.25. Given $\alpha \in \mathcal{ML}_n$, the satisfiability of $\mathcal{T}[\alpha]$ can be decided in EXPTIME.

Note that the size of standard translation ST_u also grows linearly, our approach $\mathcal{T}$ is as succinct as ST_u. We use the following example to compare two translations.

Example 5.26. Let $\alpha := \boldsymbol{K}_1 p \wedge \neg \boldsymbol{K}_1 \boldsymbol{K}_1 p$. Apparently, α is not $K45_n$-satisfiable. The standard translation $ST_u[\alpha]$ is

$$\forall v(R_1(u,v) \supset P(v)) \wedge \neg \forall v(R_1(u,v) \supset \forall u(R_1(v,u) \supset P(v)))$$

The resulting formula is FO-satisfiable. To reason about the $K45_n$-properties via the standard translation, for each $i \in \mathsf{Ag}$ the following axioms need to be added:

$$\forall u \forall v \forall v'(R_i(u,v) \wedge R_i(v,v')) \supset R_i(u,v') \qquad \text{(Transitivity)}$$

$$\forall u \forall v \forall v'(R_i(u,v) \wedge R_i(u,v')) \supset R_i(v,v') \qquad \text{(Euclideanity)}$$

Apparently, these axioms are not in FO^2 since three variables are needed, and they are not in GF since a guard containing all free variables does not exist.[5] Considering the same formula α, our translation $\mathcal{R}[\alpha, \epsilon, u, v]$ is as follows:

$$\forall v(E_1(v,u) \supset P(v)) \wedge \neg \forall v(E_1(v,u) \supset \forall v(E_1(v,u) \supset P(v)))$$

Which preserves the $K45_n$-(un)satisfiability and no extra axioms are needed.

Let $\mathcal{T}_D[\alpha]$ be the sentence which replaces every free occurrence of u in $\mathcal{R}_D[\alpha, \epsilon, u, v]$ by c (and $\mathcal{T}_T[\alpha]$ is defined similarly). The translation for $KD45_n/S5_n$ systems is obtained.

Theorem 5.27. For any $\alpha \in \mathcal{ML}_n$,

- α is $KD45_n$-consistent iff $\mathcal{T}_D[\alpha]$ is FO-satisfiable.
- α is $S5_n$-consistent iff $\mathcal{T}_T[\alpha]$ is FO-satisfiable.

Theorem 5.28. For $* \in \{T, D\}$, $\alpha \in \mathcal{ML}_n$, $\mathcal{T}_*[\alpha]$ can be computed in $O(|\alpha|)$, $|\mathcal{T}_*[\alpha]| \in O(|\alpha|)$ and $\mathcal{T}_*[\alpha] \in GF^2$.

[5] It can be shown that these axioms are not even in the *loosely guarded fragment* [vB97]

5.1.5. First-order Translation

Our translation method can be extended to first-order modal logic (FOML). We lift $\mathcal{ML}_n$ to $\mathcal{KL}_n$, which can be considered as the multi-agent extension of the logic $\mathcal{KL}$ [LL01] and the fragment of $\mathcal{OL}_n$ [BL15b] without mentioning only-knowing or the "knowing at most" modalities.

The Logic $\mathcal{KL}_n$

Compared with ordinary first-order (modal) logic, $\mathcal{KL}_n$ contains a fixed, countably infinite domain of discourse $\mathcal{N}$, where each element refers to a unique *standard name*. There is a countably infinite supply of predicate and functional symbols of every arity. Formally, we have the following definition:

- By *terms* of the language we mean the minimal set which contains all variables, standard names, and expressions of form $f(t_1, \ldots, t_m)$ where f is a functional symbol and each t_j itself is a term.
- The *well-formed formulae* are inductively defined as follows:
  - For terms t and t', $t = t'$ is a formula,
  - For predicate symbol P and terms $t_1, \ldots, t_m$, $P(t_1, \ldots, t_m)$ is a formula,
  - If α, β are formulae, then $\neg\alpha, \alpha \wedge \beta, \forall x\ \alpha, \boldsymbol{K}_i\alpha$ are formulae.

We say a term is *ground* if it contains no variables. A ground term containing only a single function symbol is called *primitive*. A formula of form $P(t_1, \ldots, t_l)$ is called an *atomic formula* or simply an atom. Furthermore, an atom with no variables is called *ground* and a ground atom with no function symbols is called *primitive*.

Similar to $\mathcal{ML}_n$, the semantics of $\mathcal{KL}_n$ is also based on k-structures, but the worlds in $\mathcal{KL}_n$ are defined as assignments of primitive terms to the co-referring standard names and primitive atoms to $\{0, 1\}$. For any ground term t, the value of t at world w, written as $|t|_w$, is defined as follows:

- If t is a standard name, then $|t|_w = t$;
- For function f, $|f(t_1, \ldots, t_m)|_w = w[f(|t_1|_w, \ldots, |t_m|_w)]$.

The satisfaction relation is defined as follows:

- $(w, \vec{e}) \models P(t_1, \ldots, t_m)$ iff $w[P(|t_1|_w, \ldots, |t_m|_w)] = 1$;
- $(w, \vec{e}) \models (t_1 = t_2)$ iff $|t_1|_w$ is the same name as $|t_2|_w$;
- $(w, \vec{e}) \models \neg\alpha$ iff it is not the case that $(w, \vec{e}) \models \alpha$
- $(w, \vec{e}) \models \alpha \wedge \beta$ iff $(w, \vec{e}) \models \alpha$ and $(w, \vec{e}) \models \beta$
- $(w, \vec{e}) \models \forall x\ \alpha$ iff $(w, \vec{e}) \models \alpha^x_n$ for all $n \in \mathcal{N}$
- $(w, \vec{e}) \models \boldsymbol{K}_i\varphi$ iff for all $s \in e_i$, $s \cup_i e_i \models \varphi$

Here α^x_n means the formula obtained by substituting each occurrence of free x in α by n. Notions such as satisfiability, validity, or entailment are defined similar to $\mathcal{ML}_n$. The properties of $\mathcal{KL}_n$ and the comparison with the ordinary first-order logic are discussed in [BL15b, LL01]. It is shown that $\mathcal{KL}_n$ satisfies the $K45_n$ properties and first-order properties such as the *Barcan formula*:

- $\models \forall x \boldsymbol{K}_i\alpha \supset \boldsymbol{K}_i\forall x\alpha$
- $\models \exists x \boldsymbol{K}_i\alpha \supset \boldsymbol{K}_i\exists x\alpha$

Translation for $\mathcal{KL}_n$

We extend the translation $\mathcal{R}$ to deal with first-order modal formulae. Note that the translation implicitly results in a two-sorted first-order logic[BG07]: Quantifiers in the original formulae are of sort *object* and quantifiers introduced during translation are of sort *state*. It is commonplace that many-sorted logic is easily convertible to ordinary first-order logic [End01]. Here we present the version where one-sorted conversion has been embedded:

1. For atomic formula α, $\mathcal{R}[\alpha, \sigma, u, v] := \alpha \uparrow^u$, where:
   - For $n \in \mathcal{N}$, $n \uparrow^u := n$
   - For variable x, $x \uparrow^u := x$
   - $f(t_1, \ldots t_m) \uparrow^u := f(t_1 \uparrow^u, \ldots, t_m \uparrow^u, u)$
   - $P(t_1, \ldots t_m) \uparrow^u := P(t_1 \uparrow^u, \ldots, t_m \uparrow^u, u)$
   - $(t_1 = t_2) \uparrow^u := (t_1 \uparrow^u = t_2 \uparrow^u)$
2. $\mathcal{R}[\neg\alpha, \sigma, u, v] := \neg\mathcal{R}[\alpha, \sigma, u, v]$
3. For $\odot \in \{\wedge, \vee, \supset, \equiv\}$,
   $\mathcal{R}[\alpha_1 \odot \alpha_2, \sigma, u, v] := \mathcal{R}[\alpha_1, \sigma, u, v] \odot \mathcal{R}[\alpha_2, \sigma, u, v]$
4. If $end(\sigma) = i$, $\mathcal{R}[\boldsymbol{K}_i\alpha, \sigma, u, v] :=$
   $\forall u\ (E_\sigma(u, v) \wedge S(u)) \supset (\mathcal{R}[\alpha, \sigma, u, v])$
5. If $end(\sigma) \neq i$, $\mathcal{R}[\boldsymbol{K}_i\alpha, \sigma, u, v] :=$
   $\forall v\ (E_{\sigma \cdot i}(v, u) \wedge S(v)) \supset (\mathcal{R}[\alpha, \sigma \cdot i, v, u])$
6. $\mathcal{R}[\forall x\ \alpha, \sigma, u, v] := \forall x\ O(x) \supset \mathcal{R}[\alpha, \sigma, u, v]$
7. $\mathcal{R}[\exists x\ \alpha, \sigma, u, v] := \exists x\ O(x) \wedge \mathcal{R}[\alpha, \sigma, u, v]$

Predicates O and S denote the membership of sort object and state. Let $\mathcal{T}[\alpha]$ be to replace every free appearance of u in $\mathcal{R}[\alpha, \epsilon, c, v]$.

Theorem 5.29. If α is $\mathcal{KL}_n$-satisfiable then $\mathcal{T}[\alpha]$ is FO-satisfiable.

We leave the proof to the appendix. Similar to Thm. 5.19, we prove it by constructing a canonical model. Different from the propositional case, the other direction does not hold naively. A counter-example can be easily constructed:

Example 5.30. Let $\alpha := \forall x\ (x = n_1 \vee x = n_2)$, then $\mathcal{T}[\alpha]$ is satisfiable but α is not $\mathcal{KL}_n$-satisfiable.

Note that the difference has already been discussed in [LL01]. It is mainly because $\mathcal{KL}_n$ interprets equality in the usual sense, and requires a countably infinite domain of discourse where each element has a unique name. To achieve satisfiability-preservation, it requires additional assumptions Σ_{eq} about equality:[6]

- $(x = x)$
- $(x = y) \supset (y = x)$
- $((x = y) \wedge (y = z)) \supset (x = z)$
- for any function symbol f in $\mathcal{T}[\alpha]$:
  $((x_1 = y_1) \wedge \ldots \wedge (x_m = y_m)) \supset f(x_1, \ldots, x_m) = f(y_1, \ldots, y_m)$
- for any predicate symbol P in $\mathcal{T}[\alpha]$:
  $((x_1 = y_1) \wedge \ldots \wedge (x_m = y_m)) \supset P(x_1, \ldots, x_m) \equiv P(y_1, \ldots, y_m)$

$\Sigma_{una} = \{(n \neq n') \mid n, n' \text{ are distinct standard names in } \mathcal{N}\}$.
Σ_{sort} includes the following formulae:

- $\exists x\ O(x)$
- $S(c)$
- For each function symbol f in $\mathcal{T}[\alpha]$:
  $(O(x_1) \wedge \ldots \wedge O(x_m) \wedge S(u)) \supset O(f(x_1, \ldots x_m, u))$
- $O(n)$ for all $n \in \mathcal{N}$

Intuitively, Σ_{eq} means that the symbol $=$ is interpreted in its usual sense. Σ_{una} requires the domain of discourse to be infinite. Σ_{sort} preserves the satisfiability in the conversion from many-sorted to one-sorted (See [LL01, End01] for details). Let $\Sigma = \Sigma_{eq} \cup \Sigma_{una} \cup \Sigma_{sort}$, similar with Lem. 5.21 we have

Theorem 5.31. Given $\alpha \in \mathcal{KL}_n$, if $\mathcal{T}[\alpha] \cup \Sigma$ is FO-satisfiable, then α is $\mathcal{KL}_n$-satisfiable.

[6] Here all free variables are implicitly universally quantified.

A proof is left to the appendix. By Thm. 5.29 and Thm. 5.31, we derive the following result:

Theorem 5.32. α is $\mathcal{KL}_n$-satisfiable iff $\mathcal{T}[\alpha] \cup \Sigma$ is FO-satisfiable.

As shown in the following corollary, the infinite set of axioms Σ is needed only when equality or standard names occur in the formula. Let Σ'_{sort} be the subset of Σ_{sort} without $O(n)$ for any n:

Corollary 5.33. For $\alpha \in \mathcal{KL}_n$ without standard names or equality, α is $\mathcal{KL}_n$-satisfiable iff $\mathcal{T}[\alpha] \cup \Sigma'_{sort}$ is FO-satisfiable.

Since FOL is a fragment of $\mathcal{KL}_n$, it is clear that the decision problems of $\mathcal{KL}_n$ are undecidable. Thus it is valuable to investigate decidable first-order modal fragments. However, as shown in previous work including [BG07], the full modal extensions of many decidable first-order fragments, including the monadic fragment (with only unary predicates), GF and FO^2, cannot preserve decidability. With an analysis of the translation, we study the decidability of a fragment, which we call the *modal bounded guarded fragment* ($MBGF$).

Definition 5.34 (modal bounded guarded fragment)**.** A formula $\alpha \in \mathcal{KL}_n$ is in $MBGF$, if it contains no standard names, equals, or functional symbols,[7] and one of the following holds:

- α is an atom $P(\vec{t})$;
- α is a Boolean combination of formulae in $MBGF$;
- α is of form $\boldsymbol{K}_i\varphi$, where φ is a bounded $MBGF$;
- α is of form $\forall\vec{x}.G(\vec{x},\vec{y}) \supset \varphi$, where φ is objective and is in $MBGF$, $G(\vec{x},\vec{y})$ is an atom and $Free(\varphi) \subseteq \{\vec{x},\vec{y}\}$, i.e. all free variables in φ are in $\vec{x}$ or $\vec{y}$.

[7] Constants, which can be understood as nullary functions, can be accepted though.

Intuitively, a formula is in $MBGF$ if all the objective sub-formulae are in GF, and all subjective sub-formulae are bounded. For example, a formula of the form $\boldsymbol{K}_1(\alpha \wedge (\boldsymbol{K}_2\beta \supset \boldsymbol{K}_1\gamma))$ is in $MBGF$, provided that α, β, γ are all bounded formulae in GF. However, $\boldsymbol{K}_1\forall x(G(x) \supset \boldsymbol{K}_1F(x))$ is not in $MBGF$.

We consider the loosely guarded fragment (LGF), where the notion of guard is relaxed. In LGF quantifiers can be guarded by a conjunction of atomic formulae, provided that each quantified variable co-exists with each free or bound variable in some atom. LGF has very similar properties as GF. It has been proved that the satisfiability problem for both GF and LGF is 2EXPTIME-complete [Grä99].

Theorem 5.35. For any $\alpha \in MBGF$, $\mathcal{R}[\alpha, \sigma, u, v] \in LGF$.

Proof. When $\alpha := P(\vec{t})$, $\mathcal{R}[\alpha, \sigma, u, v] = P(\vec{t}\uparrow^u, u)$. Thus the theorem holds for all atomic formulae. The induction on $\vee, \wedge, \neg$ is trivial. Suppose that α is of form $\boldsymbol{K}_i\varphi$ where φ is a bounded $MBGF$, i.e. φ has no free variables. When $end(\sigma) = i$,

$$\mathcal{R}[\alpha, \sigma, u, v] = \forall u\ (E_\sigma(u, v) \wedge S(u)) \supset \mathcal{R}[\varphi, \sigma, u, v]$$

By the induction hypothesis, $\mathcal{R}[\varphi, \sigma, u, v] \in LGF$ and has no free variables other than u, v. Thus $\mathcal{R}[\alpha, \sigma, u, v] \in LGF$. The proof is similar when $end(\sigma) \neq i$. Suppose that α is of form $\forall\vec{x}.G(\vec{x}, \vec{y}) \supset \varphi$, then

$$\begin{aligned}&\mathcal{R}[\forall\vec{x}.G(\vec{x}, \vec{y}) \supset \varphi, \sigma, u, v]\\ =&\forall\vec{x}(O(x_1) \wedge \ldots \wedge O(x_m) \wedge G(\vec{x}, \vec{y}, u)) \supset \mathcal{R}[\varphi, \sigma, u, v]\end{aligned}$$

By induction hypothesis $\mathcal{R}[\varphi, \sigma, u, v] \in LGF$ and $\mathcal{R}[\varphi, \sigma, u, v]$ has free variables among $\{\vec{x}, \vec{y}, u\}$. Since each pair in $\{\vec{x}, \vec{y}\}$ co-occurs in $G(\vec{x}, \vec{y}, u)$, we have $\mathcal{R}[\alpha, \sigma, u, v] \in LGF$. □

Corollary 5.36. Given $\alpha \in MBGF$, the satisfiability of α can be decided in 2EXPTIME.

Proof. By Cor. 5.33, α is satisfiable iff $\mathcal{T}[\alpha] \cup \Sigma'_{sort}$ is FO-satisfiable. Thm. 5.35 shows that $\mathcal{T}[\alpha] \in LGF$. Since $\alpha \in MBGF$, α contains at most nullary functions and hence $\mathcal{T}[\alpha]$ contains functional terms with at most one variable. Thus $\Sigma'_{sort} \in LGF$, and $\mathcal{T}[\alpha] \cup \Sigma'_{sort}$ can be decided in 2EXPTIME. □

5.2. Planning with Higher-order Beliefs

Multi-agent epistemic planning (MEP) refers to the task of generating action sequences to achieve goals formulated not only in terms of the world state but also the mental states of agents. Unlike classical planning, epistemic planning explicitly characterizes and tracks how knowledge and beliefs evolve through actions. This capability enables planning in nondeterministic, partially observable, and multi-agent domains, where agents may reason about others' knowledge and beliefs. MEP plays important roles in applications such as games, human-agent interaction, and cognitive science [Pie03, DGMR+16, BBvDM17].

Most existing approaches for MEP are based on Dynamic Epistemic Logic (DEL). In the DEL approach, both epistemic states and actions (events) are represented using pointed Kripke models, and state transitions with respect to actions are computed through the *product update*, which integrates the epistemic model with the event model [BA11]. The DEL approach offers several advantages for modeling MEP: it provides high expressivity for representing epistemic concepts such as nested beliefs and common knowledge, and it allows capturing complex epistemic dynamics arising from diverse action types and effects. In addition, the product update offers a uniform mathematical method for computing how agents' knowledge evolves in response to actions.

However, despite its expressiveness, the DEL approach faces several limitations. As a semantic approach, it requires a complete and explicit construction of both the initial model and the event model, which often demands expertise in modal logic. In complex domains, the automatic generation of these models is challenging. Moreover, it has been shown that MEP in DEL is undecidable, mainly due to

the potential for unbounded growth of nested knowledge [BA11]. While decidable fragments exist, such as planning with propositional actions or pure epistemic actions [LPW11, YWL13, BJS15], strong restrictions are imposed that limit the expressivity.

To address the drawbacks of the semantic-based DEL approach, alternative syntactic approaches have also been explored. [Auc11, Auc12] investigates epistemic planning using DEL-sequent calculus and projection. Unlike the DEL approach, these works do not explicitly consider action sequences but instead produce a formal specification that the transition must fulfill. *Situation Calculus* [MH81, Rei01], which is a first-order logical framework originally designed for physical actions, is used for relevant tasks such as belief change or progression [FL13, FLW15]. However, to express knowledge and beliefs in the original design of the situation calculus requires second-order quantification, which makes it difficult to design a general planner but still preserve decidability. Via considering a modal variant of the epistemic situation calculus, [BL14a] presents a regression framework for multi-agent scenarios, which reduces queries after actions to the initial state so that tools for theorem proving can be applied. However, their approach does not handle MEP directly, and also suffers from limited expressivity, as it can only represent actions with objective pre- and postconditions, and cannot express announcements or nondeterministic actions.

To further extend the prior works and overcome the restriction on action theories, we investigate the cognitive-science concept *Theory of Mind* (ToM) - the cognitive mechanism by which one predicts other people's actions by inferring their perceptions, beliefs, and desires [HSC22]. Higher-order ToM typically corresponds to more sophisticated cognitive and reasoning abilities, involving recursive mental-state attribution [Pel23]. Inspired by the hierarchical structure of ToM, we propose a formalism of collective action theories, where each agent has an individual action theory that amounts to ToM of a certain order, and beliefs about other agents' theories, which amount to lower-order ToMs. By preserving *bounded-order*

ToM (BoToM) throughout the regression, our framework ensures the decidability of MEP represented within this formalism.

The rest of this section is organized as follows. We begin by formalizing the logic for multi-agent epistemic reasoning in dynamic systems. Next, Section 5.2.2 introduces our hierarchical formalism for collective theories. We then discuss the various action types that can be represented within our framework. Subsequently, we present the regression approach in Section 5.2.4. Section 5.2.5 provides examples to demonstrate the expressiveness of our framework. In Section 5.2.6, we define MEP problems via logical entailment and prove their decidability. We propose an algorithm which solves MEPs and returns optimal solutions.

5.2.1. The Logic $\mathcal{MS}$

We consider a propositional variant of the logic $\mathcal{ES}_n$[BL14a]: Given finite sets of agents Ag, actions Ac and atomic propositions Ap, the logic $\mathcal{MS}(\mathsf{Ag}, \mathsf{Ac}, \mathsf{Ap})$ includes the following formulae:[8]

$$\varphi ::= p \mid \neg\varphi \mid \varphi \wedge \psi \mid \boldsymbol{K}_i\varphi \mid [a]\varphi \mid \Box\varphi, \tag{5.5}$$

where $p \in \mathsf{Ap}$. Other connectives like $\vee, \supset, \equiv$ are used as the usual abbreviations. $\boldsymbol{K}_i\varphi$ means "agent i knows/believes φ", $[a]\varphi$ is read as "φ holds after action a" and $\Box\varphi$ means "φ holds after any sequence of actions". For $G \subseteq \mathsf{Ag}$, we use $\boldsymbol{E}_G\varphi$ as the abbreviation of $\wedge_{i\in G}\boldsymbol{K}_i\varphi$ and $\boldsymbol{E}_G^k\varphi$ means $\boldsymbol{E}_G(\cdots(\boldsymbol{E}_G\varphi))$ with k nested $\boldsymbol{E}_G$. When $G = \mathsf{Ag}$, we write $\boldsymbol{E}\varphi$ and $\boldsymbol{E}^k\varphi$. Note that $\mathcal{ES}_n$ is a first-order logic with a countably infinite set of objects and actions. We restrict our language to a propositional fragment for decidable MEP. A first-order extension is certainly possible, but we leave it for future work. In addition, we include two "predicates":

- $Poss(a)$: a is executable (in an agent's mind);
- $I(a, a')$: a' is indistinguishable from a (for an agent).

[8] When the context is clear, we simply write $\mathcal{MS}$.

Essentially, these predicates are interpreted as propositions in Ap, e.g., $Poss(move)$ as *poss_move*. Let Z denote the set of finite action sequences, i.e. $Z = \mathsf{Ac}^*$. We use $\langle\rangle$ to denote the empty one. A formula is *bounded* if it does not mention $\Box$, and it is *static* if it does not mention $\Box$ or any $[a]$. We say a formula is *objective* if it mentions no $\boldsymbol{K}$ operators and a formula is *subjective* if every proposition occurs in the scope of a $\boldsymbol{K}$ operator. A formula is *i-objective* if every occurrence of $\boldsymbol{K}_i$ is in the scope of a $\boldsymbol{K}_j$ where $j \neq i$, and a formula is *i-subjective* if every occurrence of propositions or $\boldsymbol{K}_j$ for $j \neq i$ is in the scope of a $\boldsymbol{K}_i$. In addition, we use $\top$ and $\bot$ to represent tautologies and falsities. For $z = a_1 \cdots a_m$, we write $[z]\varphi$ instead of $[a_1] \cdots [a_m]\varphi$. Executability of sequence $Exec(z)$ is recursively defined as

- $Exec(\langle\rangle) := \top$;
- For $z = a \cdot \sigma$, $Exec(z) := Poss(a) \wedge [a]Exec(\sigma)$.

To evaluate the depth of a formula, we follow [FWKL25] and define:

Definition 5.37 (i-depth)**.** For $i \in \mathsf{Ag}$, the i-depth of a formula α, written as $dep[\alpha, i]$, is recursively defined as

- $dep[p, i] = 0$ for $p \in \mathsf{Ap}$;
- $dep[\neg\alpha, i] = dep[\alpha, i]$;
- $dep[\alpha \wedge \beta, i] = max(dep[\alpha, i], dep[\beta, i])$;
- $dep[[a]\alpha, i] = dep[\alpha, i]$ for $a \in \mathsf{Ac}$;
- $dep[\Box\alpha, i] = dep[\alpha, i]$;
- $dep[\boldsymbol{K}_i\alpha, i] =$
  $\max(\max\{dep[\alpha, j] \mid j \neq i\} + 1, dep[\alpha, i])$;
- $dep[\boldsymbol{K}_j\alpha, i] = 0$ for $j \neq i$.

Intuitively, $dep[\alpha, i]$ measures the degree of nesting of agent i's mental state. For example, a formula $p \wedge \boldsymbol{K}_A\neg\boldsymbol{K}_B p$ has A-depth 2 and B-depth 0, as it involves

what A believes that B believes, but not the actual beliefs of B. The depth of a formula α is defined as the maximal i-depth, i.e. $dep[\alpha] = \max_{i \in \mathsf{Ag}} dep[\alpha, i]$.

The semantics of the logic involves several terms: *world*, *structure*, and *epistemic state*. Similar notions are in [BL14a, BL15b, FWKL25]. A world is defined as a mapping $w \colon Z \to 2^{\mathsf{Ap}}$. We use W to denote the set of all worlds. Let $|\mathsf{Ag}| = n$, for $k \geq 0$, we inductively define the epistemic states and structures as follows. Note that the construction is essentially the same as in Section 5.1.1. The two only differ in the definition of world:

- $\mathcal{E}^0 = \{\{\varnothing\}\}$;
- $\mathcal{S}^{k+1} = \{(w, e_1, \dots e_{n-1}) \mid w \in W,\ e_j \in \mathcal{E}^k \text{ for all } j\}$;
- $\mathcal{E}^{k+1} = \wp(\mathcal{S}^{k+1})$, i.e. $\mathcal{E}^{k+1}$ is the power set of $\mathcal{S}^{k+1}$.

We call $s^k \in \mathcal{S}^k$ a k-structure and $e^k \in \mathcal{E}^k$ a k-state. We call k the depth of s^k and e^k. Intuitively, a structure consists of what an agent considers possible about the world and about other agents' beliefs. When the context is clear, we omit the superscript k. By a model, we mean a tuple $(w, e_1, \dots e_n)$ (also written as $(w, \vec{e})$), where w is a world and e_i is the state of agent i. Note the difference between structures and models: a structure consists of $n-1$ states of the same depth, while a model consists of n states whose depths can be different. For state $\hat{e}$ and structure $s = (w, e_1, \dots e_{n-1})$, we use $s \cup_i \hat{e}$ to denote the model obtained by inserting $\hat{e}$ at the i-th position, i.e. $(w, e_1, \dots e_{i-1}, \hat{e}, e_i, \dots e_{n-1})$.
The satisfaction relation is defined as follows:

- $(w, \vec{e}) \models p$ iff $p \in w[\langle\rangle]$;
- $(w, \vec{e}) \models \neg\alpha$ iff it is not the case that $(w, \vec{e}) \models \alpha$;
- $(w, \vec{e}) \models \alpha \wedge \beta$ iff $(w, \vec{e}) \models \alpha$ and $(w, \vec{e}) \models \beta$;
- $(w, \vec{e}) \models \boldsymbol{K}_i\alpha$ iff $s \cup_i e_i \models \alpha$ for all $s \in e_i$.

We introduce the following notions. Essentially, they extend the semantics of $Poss$ and I to sequences:

Definition 5.38. We define $\Pi_w \subseteq Z$ and $\sim_w \subseteq Z \times Z$ as:

- $\langle\rangle \in \Pi_w$. For $z = \sigma \cdot a$, $z \in \Pi_w$ iff $\sigma \in \Pi_w$ and $Poss(a) \in w[\sigma]$;
- $\langle\rangle \sim_w z$ iff $z = \langle\rangle$. For $z = \sigma \cdot a$ and $z' = \sigma' \cdot a'$, $z \sim_w z'$ iff $\sigma \sim_w \sigma'$ and $I(a, a') \in w[\sigma']$.

To characterize the situation after actions, we extend the world/state progression [Cla13, LF23] to multi-agent cases:

Definition 5.39. For $w \in W$ and $z \in Z$, the progression of w wrt z, written as w_z, is defined as the world w' such that for all $z' \in Z$, $w[z \cdot z'] = w'[z']$. For epistemic state $e \in \mathcal{E}^k$, the progression of e wrt z, written as e_z, is defined as follows:

- When $k = 1$, i.e. e is a state of depth 1,

$$e_z = \{(w_\sigma, \{\varnothing\}, \ldots \{\varnothing\}) \mid (w, \{\varnothing\}, \ldots \{\varnothing\}) \in e, \sigma \in \Pi_w, z \sim_w \sigma\};$$

- When $k > 1$, i.e. e is a state deeper than 1,

$$e_z = \{(w_\sigma, e'_{1,\sigma}, \ldots e'_{n-1,\sigma}) \mid (w, e'_1, \ldots e'_{n-1}) \in e, \sigma \in \Pi_w \text{ and } z \sim_w \sigma\},$$

where $e'_{j,\sigma}$ is the progression of e'_j wrt σ.

Now we can consider formulae with action modalities:

- $(w, \vec{e}) \models [a]\varphi$ iff $(w_a, \vec{e}_a) \models \varphi$
- $(w, \vec{e}) \models \Box\alpha$ iff $(w_z, \vec{e}_z) \models \alpha$ for all $z \in Z$

We say a model and a formula α is *compatible* if, for all $i \in \mathsf{Ag}$, the i-depth of α is not deeper than the epistemic state of agent i. Given a finite set $\Sigma \subseteq \mathcal{MS}$,[9]

[9] Σ can also be an infinite set, provided that the depth of formulae is bounded, i.e. it exists $k \in \mathbb{N}$ s.t. $dep[\varphi] \leq k$ for all $\varphi \in \Sigma$.

$\alpha \in \mathcal{MS}$, we say Σ entails α (written as $\Sigma \models \alpha$) iff for all model $(w, \vec{e})$ compatible with each formula in $\Sigma \cup \{\alpha\}$, if $(w, \vec{e}) \models \varphi$ for all $\varphi \in \Sigma$, then $(w, \vec{e}) \models \alpha$. We say α is *valid* (written as $\models \alpha$) iff $\{\} \models \alpha$. We say a formula α is *satisfiable* if there is a compatible model $(w, \vec{e})$ s.t. $(w, \vec{e}) \models \alpha$. When α is objective, we write $w \models \alpha$ instead of $(w, \vec{e}) \models \alpha$. When α is i-subjective, we write $e_i \models \alpha$. Note that $\mathcal{MS}$ cannot handle unbounded depth of formulae. When Σ is an infinite set of formulae with unbounded depth, i.e. for all k it exists $\psi \in \Sigma$ s.t. $dep[\psi] > k$, then $\Sigma \models \alpha$ is not well-defined.

The logic satisfies the $K45_n$ properties after any actions.

Theorem 5.40. For any $\alpha, \beta \in \mathcal{MS}$ and $i \in \mathsf{Ag}$:

- (Prop) $\models \Box \boldsymbol{K}_i \alpha$ if α is a (propositional) tautology;
- (Dist)$\models \Box \boldsymbol{K}_i \alpha \wedge \boldsymbol{K}_i(\alpha \supset \beta) \supset \boldsymbol{K}_i \beta$;[10]
- (4)$\models \Box \boldsymbol{K}_i \alpha \supset \boldsymbol{K}_i \boldsymbol{K}_i \alpha$;
- (5)$\models \Box \neg \boldsymbol{K}_i \alpha \supset \boldsymbol{K}_i \neg \boldsymbol{K}_i \alpha$.

The static entailment amounts to $K45_n$ theorem proving.

Theorem 5.41. Given static Σ and α, $\Sigma \models \alpha$ iff α is derivable from Σ under the $K45_n$ axiom system.

It is proven by constructing a bisimulation between $\mathcal{MS}$ models and $K45_n$ Kripke models. We refer to [FWKL25] for the proof.

With the syntax and semantics of the logic $\mathcal{MS}$, we now turn to multi-agent action theories, which specify the transition of not only the world but also agents' mental states.

[10] Throughout the paper, we assume that the modality $\Box$ always has the lowest precedence among all operators.

5.2.2. Multi-Agent Action Theories

The proposed action theories are inspired by the BHL framework [BHL99a]. As in BHL, a theory includes *precondition axioms*, *successor-state axioms* (SSAs) which describe the effect of actions, and *observation-indistinguishability* (OI) axioms, which distinguish the action actually being executed from those likely being executed. In addition, we include axioms on other agents' knowledge of actions.

Definition 5.42 (Individual Action Theory)**.** For $i \in \mathsf{Ag}$ and $k \in \mathbb{N}$, We define the *k-order individual action theory (IAT) of agent i*, denoted as $\Sigma^{i,k}$, as the following axioms:

- $\Sigma^{i,k}_{pre}$: For each action $a \in \mathsf{Ac}$,

$$\Box Poss(a) \equiv \pi_a \tag{5.6}$$

  where π_a is a static formula with a depth at most k.

- $\Sigma^{i,k}_{ssa}$: For each $a \in \mathsf{Ac}$ and $p \in \mathsf{Ap}$,[11]

$$\Box[a]p \equiv \gamma_{p,a} \tag{5.7}$$

  where $\gamma_{p,a}$ is a static formula with a depth at most k.

- $\Sigma^{i,k}_{oi}$: For each $a, a' \in \mathsf{Ac}$,

$$\Box I(a, a') \equiv \phi_{a,a'} \tag{5.8}$$

  where $\phi_{a,a'}$ is a static formula with a depth at most k.

- $\Sigma^{i,k}_{iat}$: For each $j \neq i$, j's knowledge of actions $\boldsymbol{K}_j\Sigma^{j,k-1}$ is included. When $k = 0$, $\Sigma^{j,k-1} := \top$. When $k > 0$, $\Sigma^{j,k-1}$ is an IAT of order $k-1$.

Note that when $k > 1$, $\Sigma^{j,k-1}$ also includes other agents' knowledge about actions. The nesting continues until $k = 0$. Hence, a hierarchical structure is formed. For each $i \in \mathsf{Ag}$ and theory $\Sigma^{i,k}$, we require that π_a, $\gamma_{p,a}$ and $\phi_{a,a'}$ do not mention the predicates $Poss$ and I. Based on the IAT of each agent, we define the *collective action theory* (CAT):

[11] Propositions about $Poss$ or I, e.g. *poss_move*, are excluded.

Definition 5.43 (Collective Action Theory)**.** For $k \in \mathbb{N}$, a k-order CAT, written as Γ, is defined as

$$\Gamma := \Gamma_{pre} \cup \Gamma_{ssa} \cup \Gamma_{oi} \cup \bigcup_{i \in \mathsf{Ag}} \boldsymbol{K}_i \Sigma^{i,k_i} \tag{5.9}$$

where Γ_{pre}, Γ_{ssa} and Γ_{oi} are respectively of the form (5.6-5.8) and the corresponding π_a, $\gamma_{p,a}$ and $\phi_{a,a'}$ are static formulae of depth at most k. Σ^{i,k_i} is the IAT of agent i and $k_i < k$. We use $\mathsf{Ord}(\Gamma, i) := dep[\Gamma, i]$ to denote the order of agent i's mind, and $\mathsf{Ord}(\Gamma)$ denotes the maximal order of all agents. By a Γ-model, we mean a model $(w, \vec{e})$ compatible with Γ, and $(w, \vec{e}) \models \Gamma$.

Essentially, a k-order IAT corresponds to the mind of an agent with k-order ToM. There are two reasons for having such a hierarchy: The first is to distinguish an agent's IAT from what other agents believe to be his or her IAT, and the second is for the hierarchical nature of ToM: A child with 1-order ToM could accurately reason about others' beliefs about the world state, and can distinguish what is believed by others from what is true. When two children with 1-order ToM share their rationale behind actions, can they accurately model the knowledge about the mental state transition of the other? The answer is no. A k-order agent cannot reason about the knowledge of another k-order agent, as representing others' perspectives consumes one level of nesting.

To accurately reason about k-order action theories, it requires a hierarchy of order $2k$. Considering a single theory per agent of the form $\Delta^{i,k} := \Sigma^{i,k}_{pre} \cup \Sigma^{i,k}_{ssa} \cup \Sigma^{i,k}_{oi}$, which is the union of the axioms (5.6-5.8), then a collective action theory, where the theories of all agents are "commonly known" up to depth k, is represented as:

$$\Delta := \boldsymbol{E}^k\big(\bigcup_{i \in \mathsf{Ag}} \boldsymbol{K}_i \Delta^{i,k}\big) \tag{5.10}$$

Then Δ is equivalent to a $2k$-order CAT.

Definition 5.44. Given theory Γ which includes $\boldsymbol{K}_i\Sigma^{i,k}$ and $\Sigma^{i,k}$ of the form $\Sigma^{i,k}_{pre} \cup \Sigma^{i,k}_{ssa} \cup \Sigma^{i,k}_{oi} \cup \bigcup_{j \neq i} \boldsymbol{K}_j \Sigma^{j,k-1}$, the **slice of** Γ **wrt agent** i, written as Γ^i, is

a CAT defined as

$$\Gamma^i := \Sigma^{i,k}_{pre} \cup \Sigma^{i,k}_{ssa} \cup \Sigma^{i,k}_{oi} \cup \boldsymbol{K}_i\Sigma^{i,k} \cup \bigcup_{j\neq i} \boldsymbol{K}_j\Sigma^{j,k-1}$$

i.e. the precondition axioms, SSAs and OI axioms of Γ^i are the corresponding parts in $\Sigma^{i,k}$.

Essentially, Γ^i is a CAT reconstructed via agent i's IAT in Γ. By induction we prove that

Lemma 5.45. For any CAT Γ and $i, j \in \mathsf{Ag}$, $i \neq j$:

- $\mathsf{Ord}(\Gamma^i, i) = \mathsf{Ord}(\Gamma, i)$;
- $\mathsf{Ord}(\Gamma^i, j) < \mathsf{Ord}(\Gamma, i)$.

Now we consider the existence of Γ-models: For any model $(w, \vec{e})$, since Γ only regulates the state transition but not the initial state, one can construct a Γ-model wrt $(w, \vec{e})$ which satisfies the same static formulae:

Theorem 5.46. Given theory Γ and model $(w, \vec{e})$ compatible with Γ, there exists a Γ-model $(w_\Gamma, e_{1,\Gamma}, \ldots, e_{n,\Gamma})$ (also written as $(w_\Gamma, \vec{e}_\Gamma)$) of the same depth such that for each compatible static formula φ:

$$(w, \vec{e}) \models \varphi \text{ iff } (w_\Gamma, \vec{e}_\Gamma) \models \varphi \tag{5.11}$$

Proof. Given $(w, \vec{e})$, we recursively construct $(w_\Gamma, \vec{e}_\Gamma)$ such that for all $p \in \mathsf{Ap}$, $z \in Z$ and $a, a' \in \mathsf{Ac}$,

- $w_\Gamma[\langle\rangle] = w[\langle\rangle]$;
- $p \in w_\Gamma[z \cdot a]$ iff $(w_\Gamma, \vec{e}_\Gamma) \models [z]\gamma^\Gamma_{p,a}$;
- $Poss(a) \in w_\Gamma[z]$ iff $(w_\Gamma, \vec{e}_\Gamma) \models [z]\pi^\Gamma_a$;
- $I(a, a') \in w_\Gamma[z]$ iff $(w_\Gamma, \vec{e}_\Gamma) \models [z]\phi^\Gamma_{a,a'}$;

- For any $i \in \mathsf{Ag}$, $e_{i,\Gamma} = \{s_{\Gamma^i} \mid s \in e_i\}$, where
  For $s = (w, e_1, \dots e_{i-1}, e_{i+1}, \dots e_n)$,
  $s_{\Gamma^i} = (w_{\Gamma^i}, e_{1,\Gamma^i}, \dots e_{i-1,\Gamma^i}, e_{i+1,\Gamma^i}, \dots e_{n,\Gamma^i})$.

Here $\gamma^{\Gamma}_{p,a}, \pi^{\Gamma}_{a}$ and $\phi^{\Gamma}_{a,a'}$ are respectively the right-hand side of $\Gamma_{pre}, \Gamma_{ssa}$ and Γ_{oi}. For all e_j and w, we construct e_{j,Γ^i} and w_{Γ^i} in a similar way but wrt Γ^i instead of Γ. We prove Eq.5.11 via a double-induction: The outer induction is on the order of Γ and the inner induction is on the construction of formula φ.

Base case 1: For a zero-order CAT, it reduces to the single-agent cases with objective φ. We refer to [LL11] for the proof details.

I.H.1: Suppose that Eq. (5.11) holds for all Γ' which satisfies $\mathsf{Ord}(\Gamma') < k$, formula φ, and $(w, \vec{e})$ compatible with Γ'.

Now we fix the theory Γ and induce on φ:

Base case 2: For any Γ with $\mathsf{Ord}(\Gamma) = k$, It is trivial that $(w, \vec{e}) \models p$ iff $(w_\Gamma, \vec{e}_\Gamma) \models p$.

I.H.2: Suppose that $(w, \vec{e}) \models \alpha$ iff $(w_{\Gamma'}, \vec{e}_{\Gamma'}) \models \alpha$ for all Γ' with $\mathsf{Ord}(\Gamma') = k$ and $(w, \vec{e})$ compatible with Γ'. Now we consider a fixed Γ. Induction on Boolean operators is trivial. For induction on the modality $\boldsymbol{K}_i$,

$$
\begin{aligned}
& (w, \vec{e}) \models \boldsymbol{K}_i\alpha \\
\Leftrightarrow\ & \text{f.a. } s \in e_i, s \cup_i e_i \models \alpha && \text{(Semantics)} \\
\Leftrightarrow\ & \text{f.a. } s \in e_i, s_{\Gamma^i} \cup_i e_{i,\Gamma^i} \models \alpha && (\#) \\
\Leftrightarrow\ & \text{f.a. } s' \in e_{i,\Gamma}, s' \cup_i e_{i,\Gamma} \models \alpha && (e_{i,\Gamma} = e_{i,\Gamma^i}) \\
\Leftrightarrow & e_{i,\Gamma} \models \boldsymbol{K}_i\alpha \Leftrightarrow (w_\Gamma, \vec{e}_\Gamma) \models \boldsymbol{K}_i\alpha
\end{aligned}
$$

There are two cases for $(\#)$: When $\mathsf{Ord}(\Gamma^i) < k$, then I.H.1 is applied. Otherwise, we apply I.H. 2 as $\mathsf{Ord}(\Gamma^i) = k$.

To prove that $(w_\Gamma, \vec{e}_\Gamma)$ is indeed a Γ-model, we check each entry of Γ. For the precondition, SSAs and OI axioms, the proof applies simply the semantics and

the definition of $(w_\Gamma, \vec{e}_\Gamma)$. For the IAT of each agent i:

$$\begin{aligned}
&(w_\Gamma, \vec{e}_\Gamma) \models \boldsymbol{K}_i \Sigma^{i,k_i} \\
\Leftrightarrow &s' \cup_i e_{i,\Gamma} \models \Sigma^{i,k_i} \text{ f.a. } s' \in e_{i,\Gamma} \\
\Leftrightarrow &s' \cup_i e_{i,\Gamma} \models \Sigma^{i,k_i} \wedge \boldsymbol{K}_i \Sigma^{i,k_i} \text{ f.a. } s' \in e_{i,\Gamma} \\
\Leftrightarrow &s' \cup_i e_{i,\Gamma} \models \Gamma^i \text{ f.a. } s' \in e_{i,\Gamma} \\
\Leftrightarrow &s_{\Gamma^i} \cup_i e_{i,\Gamma^i} \models \Gamma^i \text{ f.a. } s \in e_i
\end{aligned}$$

It is equivalent to prove that for each $s \in e_i$, $s_{\Gamma^i} \cup_i e_{i,\Gamma^i}$ is a Γ^i-model, which can be proved via induction. □

We define a knowledge base with bounded-order ToM (BoToM-KB) as follows:

Definition 5.47 (BoToM-KB)**.** A finite set of formulae $\Sigma_0 \cup \Gamma$ forms a BoToM-KB, if there is $k \in \mathbb{N}$ such that Σ_0 is a static formula of depth at most k, and Γ is a k-order CAT.

Σ_0 can be seen as the initial state in the DEL approach, and to some extent, Γ corresponds to the DEL state-transition system. However, to represent the ToM hierarchy in DEL, the event models and product update need to be modified such that the depths of the Kripke states are considered. So far, no one has investigated such modification.

5.2.3. Action Types

Three types of action are usually considered in multi-agent domains: *ontic actions*, *sensing*, and *announcements*. We show how these actions can be modeled in our framework.

Ontic Action

Ontic actions have actual effects on the world state. Given a theory Γ, an action a is ontic if there is a $p \in \mathsf{Ap}$ such that Γ does not entail $\Box\gamma_{p,a} \equiv p$. Our framework

can represent ontic action with uncertain or unexpected effects. Uncertain effects, either caused by the partial observability of the agent or the non-deterministic nature of the world, can be modeled in the sense that $\phi_{a,a'}$ of a fixed a holds for more than one a'. Unexpected effects are modeled in the sense that the action which actually happens is considered impossible, i.e. $\phi_{a,a}$ does not hold.

Sensing and Announcement

In contrast to ontic actions, sensing and announcements do not change the actual world but only what agents believe. Thus, these actions only have trivial SSAs with $\gamma_{p,a} \equiv p$. Let see_h be sensing the tossing result $\underline{H}ea\underline{d}$, then a precise sensing can be represented by

$$\Box Poss(see_h) \equiv Hd; \quad \Box I(see_h, see_h) \equiv \top;$$
$$\Box I(see_h, a') \equiv \bot \text{ for } a' \neq see_h$$

Let Σ^i include these axioms, then $\boldsymbol{K}_i\Sigma^i \models [see_h]\boldsymbol{K}_i Hd$.
Our framework supports announcement to any group of agents. Let $G!Hd$ denote the announcement to $G \subseteq \mathsf{Ag}$ that the coin shows head. For each $i \in G$, let $\Sigma^{i,k}$ include

$$\Box Poss(G!Hd) \equiv Hd; \quad \Box I(G!Hd, G!Hd) \equiv \top;$$
$$\Box I(G!Hd, a') \equiv \bot \text{ for } a' \neq G!Hd.$$

Let Γ include $\boldsymbol{K}_i\Sigma^{i,k}$ for each $i \in G$, then

$$\Gamma \models [G!Hd]\boldsymbol{E}^l_G Hd$$

for any l up to the order of theory Γ. Other forms of actions, such as noisy sensing or deceptive announcements, will be demonstrated by examples in a later section.

In fact, the notion of "actions" here is similar to "events" in the DEL approach, as the executor of an action is not specified. While complex constructs such as action profiles (multiple agents simultaneously choose actions) can naturally be

expressed within our framework, we focus on the general formalism and do not elaborate on these constructs.

5.2.4. Regression

For the action theories mentioned above, we introduce the regression method, which transforms epistemic queries on a future situation to those on the current situation.

Definition 5.48 (Regressable)**.** We call a formula φ *regressable wrt* Γ iff φ is bounded, and $dep[\varphi, i] \leq \mathsf{Ord}(\Gamma, i)$ for all i.

Definition 5.49 (Regression function)**.** For $z \in Z$ and k-order collective theory Γ, we recursively define $\mathcal{R}[\Gamma, z, \alpha]$:

1. $\mathcal{R}[\Gamma, \langle\rangle, p] := p$ for p not mentioning $Poss$ or I;
2. $\mathcal{R}[\Gamma, \sigma \cdot a, p] := \mathcal{R}[\Gamma, \sigma, \gamma^{\Gamma}_{p,a}]$;
3. $\mathcal{R}[\Gamma, \sigma, Poss(a)] := \mathcal{R}[\Gamma, \sigma, \pi^{\Gamma}_{a}]$;
4. $\mathcal{R}[\Gamma, \sigma, I(a, a')] := \mathcal{R}[\Gamma, \sigma, \phi^{\Gamma}_{a,a'}]$;
5. $\mathcal{R}[\Gamma, z, \alpha \wedge \beta] := \mathcal{R}[\Gamma, z, \alpha] \wedge \mathcal{R}[\Gamma, z, \beta]$;
6. $\mathcal{R}[\Gamma, z, \neg\alpha] := \neg\mathcal{R}[\Gamma, z, \alpha]$;
7. $\mathcal{R}[\Gamma, z, [a]\alpha] := \mathcal{R}[\Gamma, z \cdot a, \alpha]$;
8. $\mathcal{R}[\Gamma, \langle\rangle, \boldsymbol{K}_i\varphi] := \boldsymbol{K}_i\big(\mathcal{R}[\Gamma^i, \langle\rangle, \varphi]\big)$;
9. $\mathcal{R}[\Gamma, \sigma \cdot a, \boldsymbol{K}_i\varphi] := \mathcal{R}[\Gamma, \sigma, \boldsymbol{K}_i \bigwedge_{a'} \big((Poss(a') \wedge I(a, a')) \supset [a']\varphi\big)]$

The regression always terminates and returns a static formula of size $O(|\varphi| \cdot |\Gamma| \cdot |\mathsf{Ac}|^{k \cdot m})$, where k is the order of Γ and m is the action length of $[z]\varphi$. When the second argument is an empty sequence, we write $\mathcal{R}[\Gamma, \alpha]$ instead of $\mathcal{R}[\Gamma, \langle\rangle, \alpha]$.

The main motivation for introducing hierarchical structures and descending orders of action theories is to limit the depth of regression results. Via induction, we prove that

Theorem 5.50. For any theory Γ, φ regressable wrt Γ, for all z and $i \in \mathsf{Ag}$,

$$dep[\mathcal{R}[\Gamma, z, \varphi], i] \leq \mathsf{Ord}(\Gamma, i) \tag{5.12}$$

Namely, the regression never results in a formula deeper than Γ. The theorem makes sense intuitively, but the proof is complicated: A complete proof requires a triple-induction, one layer on the order of Γ, one on the length of sequence z and one on the construction of φ.

Proof sketch: W.l.o.g. we consider static φ, as the non-static cases can be converted via the 7-th regression rule. For the basis where $z = \langle\rangle$, the only non-trivial cases are $\varphi = Poss(a)$, $\varphi = I(a, a')$, and $\varphi = \boldsymbol{K}_i\alpha$.

By the corresponding regression rules, $\mathcal{R}[\Gamma, \sigma, Poss(a)] := \mathcal{R}[\Gamma, \sigma, \pi_a^\Gamma]$, where π_a^Γ is a static function of depth at most $\mathsf{Ord}(\Gamma, i)$ and does not mention $Poss$ or I, by induction on the formula construction, $dep[\pi_a^\Gamma, i] \leq \mathsf{Ord}(\Gamma, i)$. Cases for $I(a, a')$ can be proved similarly. For $\boldsymbol{K}_i\alpha$, Γ^i is an IAT where $\mathsf{Ord}(\Gamma^i, i) = \mathsf{Ord}(\Gamma, i)$ and $\mathsf{Ord}(\Gamma^i, j) = \mathsf{Ord}(\Gamma, i) - 1$ for all $j \neq i$. It is reduced to the cases where a theory has lower orders.

When $z \neq \langle\rangle$, suppose that $z = \sigma \cdot a$,

- when $\varphi = p$, then $\mathcal{R}[\Gamma, z, \varphi] = \mathcal{R}[\Gamma, \sigma, \gamma_{p,a}^\Gamma]$, which is reduced to cases with a shorter sequence.
- when $\varphi = \boldsymbol{K}_i\alpha$, analogously, we apply the last regression rule and reduce the case to regression with sequence σ. The rest can be proved similarly with the cases when z is empty.

A complete proof is left to the appendix A.1.2.

The regression is correct in the sense that any Γ-model satisfies α after sequence z if and only if the model satisfies the initial condition $\mathcal{R}[\Gamma, z, \alpha]$:

Lemma 5.51. Given action theory Γ, for all Γ-model $(w, \vec{e})$, $z \in Z$ and static formula α regressable wrt Γ,

$$(w, \vec{e}) \models [z]\alpha \text{ iff } (w, \vec{e}) \models \mathcal{R}[\Gamma, z, \alpha] \tag{5.13}$$

We proceed with a triple-induction: The outer-induction is on the order of the theory Γ, the mid-induction is on the length of the action sequence z, and the inner-induction is on the construction of formula α. We leave the complete proof to the appendix A.1.2.

The lemma can be easily extended to regression on arbitrary bounded formulae:

Theorem 5.52. Given action theory Γ, for all Γ-model $(w, \vec{e})$, $z \in Z$ and bounded formula α regressable wrt Γ,

$$(w, \vec{e}) \models [z]\alpha \text{ iff } (w, \vec{e}) \models \mathcal{R}[\Gamma, z, \alpha] \tag{5.14}$$

Now we are ready for the regression theorem:

Theorem 5.53 (Regression)**.** Given BoToM-KB $\Sigma_0 \cup \Gamma$, for all $z \in Z$, and formula α regressable wrt Γ,

$$\Sigma_0 \cup \Gamma \models [z]\alpha \quad \text{iff} \quad \Sigma_0 \models \mathcal{R}[\Gamma, z, \alpha] \tag{5.15}$$

Proof. It suffices to prove the equivalence of the statements:

$$\text{F.a. } (w, \vec{e}) \models \Gamma, \text{ if } (w, \vec{e}) \models \Sigma_0, \text{ then } (w, \vec{e}) \models [z]\alpha \tag{5.16}$$

$$\text{F.a. } (w, \vec{e}), \text{ if } (w, \vec{e}) \models \Sigma_0, \text{ then } (w, \vec{e}) \models \mathcal{R}[\Gamma, z, \alpha] \tag{5.17}$$

(5.16)⇒(5.17): For any model $(w, \vec{e})$:

$$\begin{aligned}
(w, \vec{e}) \models \Sigma_0 &\Rightarrow (w_\Gamma, \vec{e}_\Gamma) \models \Sigma_0 && \text{(Thm.5.46)} \\
&\Rightarrow (w_\Gamma, \vec{e}_\Gamma) \models [z]\alpha && \text{(Eq. 5.16)} \\
&\Rightarrow (w_\Gamma, \vec{e}_\Gamma) \models \mathcal{R}[\Gamma, z, \alpha] && \text{(Thm.5.52)} \\
&\Rightarrow (w, \vec{e}) \models \mathcal{R}[\Gamma, z, \alpha] && \text{(Thm.5.46)}
\end{aligned}$$

(5.16)⇐(5.17): For any Γ-model $(w_\Gamma, \vec{e}_\Gamma)$,

$$\begin{aligned}(w_\Gamma, \vec{e}_\Gamma) \models \Sigma_0 &\Rightarrow (w_\Gamma, \vec{e}_\Gamma) \models \mathcal{R}[\Gamma, z, \alpha] && \text{(Eq. 5.17)}\\ &\Rightarrow (w_\Gamma, \vec{e}_\Gamma) \models [z]\alpha && \text{(Thm.5.52)}\end{aligned}$$

□

By Thm.5.52 we proved the following result, which will later be used in the planning task.

Corollary 5.54. For $z \in Z$, and any φ, ψ regressable wrt. Γ,

- if $\models \varphi \equiv \psi$, then $\models \mathcal{R}[\Gamma, z, \varphi] \equiv \mathcal{R}[\Gamma, z, \psi]$
- if $\models \mathcal{R}[\Gamma, \varphi] \equiv \mathcal{R}[\Gamma, \psi]$, then $\models \mathcal{R}[\Gamma, z, \varphi] \equiv \mathcal{R}[\Gamma, z, \psi]$

5.2.5. Examples

We formalize action theories for two examples: Sally-Anne Test [WP83] and Die Rolling. Only the key entries of the theories relevant to regression are present. In particular, we consider the dynamics of four actions:

- *move*: To move the marble. We consider the unexpected effect of the action, or as described in [BGPS22], an agent is oblivious of the actions' occurrence.
- *roll*: To roll a die. A non-deterministic action.
- *show*2: An agent saw that the die showed two. We model noisy sensing, namely, a distracted agent might misread the number.
- !*bob*_3: Bob announces that the die shows three (even if it does not). An example of deceptive announcements.

Example 5.55 (Sally-Anne Test)**.** Initially, Sally's marble is in a basket (and everyone knows that), then Anne (secretly) moves the marble into a box. In a test, the child will be asked "Will Sally look in the box for her marble?"

The test is a classic psychological experiment used to assess the order of ToM: If the participant believes that Sally will not look in the box, it indicates that the child has developed first-order ToM and can understand another person's beliefs, even if they are false. Otherwise, the child has at most zero-order ToM. We show how different orders of ToM can be characterized:

- $\mathsf{Ag} = \{S, C_0, C_1\}$, which respectively stands for
  Sally, Child with 0-order ToM and with 1-order ToM.
- $\mathsf{Ap} = \{InBt, InBx\}$
  The marble is in the Basket / in the Box.
- $\mathsf{Ac} = \{move, void, look_bx\}$, *void* for nothing happens, *look_bx* for "look in the box for the marble"

We consider a collective theory Γ which includes

- $\boldsymbol{K}_{C_0}\Sigma^{C_0}$, i.e. a 0-order IAT for child C_0, and
- $\boldsymbol{K}_{C_1}\Sigma^{C_1}$, a 1-order IAT for child C_1.

Σ^{C_0} includes

- $\Box Poss(look_bx) \equiv InBx$,
- $\Box[move]InBx \equiv \top$, $\Box[move]InBt \equiv \bot$, and
- $\Box I(a, a) \equiv \top$, $\Box I(a, a') \equiv \bot$ for all $a \neq a'$.

Σ^{C_1} differs from Σ^{C_0} by

- $\Box Poss(look_bx) \equiv \boldsymbol{K}_S InBx$, i.e. Sally will look in the box only when she believes that the marble is in the box.
- Σ^{C_1} includes $\boldsymbol{K}_S\Sigma^{S,0}$ where $\Sigma^{S,0}_{oi}$ includes
  $\Box I(move, void) \equiv \top$, $\Box I(move, move) \equiv \bot$. Namely, Sally is oblivious to the move.

Let Σ_0 be the initial state. We encode the children's answers to the test as their knowledge about the possibility of action $look_bx$. By regression, we prove that

$$\begin{aligned}
&\Sigma_0 \cup \Gamma \models \boldsymbol{K}_{C_0}([move]Poss(look_bx)) \\
\Leftrightarrow &\Sigma_0 \models \mathcal{R}[\Gamma, \boldsymbol{K}_{C_0}([move]Poss(look_bx))] \\
\Leftrightarrow &\Sigma_0 \models \boldsymbol{K}_{C_0}(\mathcal{R}[\Gamma^{C_0}, move, Poss(look_bx)]) \\
\Leftrightarrow &\Sigma_0 \models \boldsymbol{K}_{C_0}(\mathcal{R}[\Gamma^{C_0}, move, \top]) \Leftrightarrow \Sigma_0 \models \top
\end{aligned}$$

Namely, in any initial case, C_0 believes that Sally will look in the box for the marble. In contrast, child C_1 does not believe that Sally will look in the box. By regression,

$$\begin{aligned}
&\Sigma_0 \cup \Gamma \models \neg\boldsymbol{K}_{C_1}([move]Poss(look_bx)) \\
\Leftrightarrow &\Sigma_0 \models \neg\boldsymbol{K}_{C_1}(\mathcal{R}[\Gamma^{C_1}, move, \boldsymbol{K}_S InBx]) \\
\Leftrightarrow &\Sigma_0 \models \neg\boldsymbol{K}_{C_1}\boldsymbol{K}_S(\mathcal{R}[\Gamma^{S}, void, InBx]) \\
\Leftrightarrow &\Sigma_0 \models \neg\boldsymbol{K}_{C_1}\boldsymbol{K}_S(InBx)
\end{aligned}$$

According to example, the child knows that Sally knows that initially, the marble is in the basket instead of the box.

Example 5.56 (Die Rolling)**.** Alice and Bob roll a die. The result can be shown to them, but when Alice is distracted, she may misread the number. In addition, Bob can announce the number, even if it is wrong, and Alice will believe it anyhow.

- $\mathsf{Ag} = \{A, B\}$, i.e. two agents Alice and Bob.
- $\mathsf{Ac} = \{roll, roll2, show2, !bob_3, \ldots\}$
  $roll$ is a nominal action and $roll1$ to $roll6$ are factual actions with the resulting number. $show2$ and $!bob_3$ are, respectively, the noisy sensing and announcement mentioned at the beginning of this section.
- $\mathsf{Ap} = \{N1, \ldots N6, Dt\}$. Dt means Alice is distracted, Nx stands for the number that the die shows.

We consider the collective theory Γ which includes $\boldsymbol{K}_A\Sigma^A$ and $\boldsymbol{K}_B\Sigma^B$. Alice's individual theory Σ^A consists of

- $\Box Poss(show2) \equiv (Dt \supset (N2 \vee N3)) \wedge (\neg Dt \supset N2)$
  Alice may misread 3 as 2 when she is distracted.
- $\Box Poss(!bob_3) \equiv N3 \wedge \neg N1 \wedge \cdots \wedge \neg N6$
  Alice always believes the truth of Bob's announcement.
- $\Box[roll2]N2 \equiv \top$ and $\Box[roll2]Nx \equiv \bot$ for $x \neq 2$.
- $\Box I(roll, rollx) \equiv \top$ for $x \in \{1, \ldots 6\}$

Bob's theory Σ^B differs from Alice's by

- $\Box Poss(show2) \equiv N2$, i.e. Bob never misreads.
- $\Box Poss(!bob_3) \equiv \top$, i.e. Bob can always announce that the number is 3, as long as he wants to do so.

After rolling the die and showing the result 2, to achieve a state where Alice does not know that the die shows 2, by regression we prove that the initial situation must fulfill

$$\Sigma_0 \cup \Gamma \models [roll][show2]\neg \boldsymbol{K}_A N2 \Leftrightarrow \Sigma_0 \models \neg \boldsymbol{K}_A \neg Dt$$

Namely, Alice cannot confirm that she stays focused. Lastly, we show how Bob can make a false announcement and manipulate Alice's beliefs:

$$\Sigma_0 \cup \Gamma \models [!bob_3](\boldsymbol{K}_A N3) \Leftrightarrow \Sigma_0 \models \top$$

Namely, Alices will always believe Bob's announcement.

5.2.6. Multi-agent Epistemic Planning

Multi-agent epistemic planning amounts to finding an action sequence such that, after taking the actions, a given goal is fulfilled. In the DEL approach, MEP is known to be robustly undecidable mainly because agents can reason with an

arbitrary nesting of beliefs [AB13]. With the bounded-order hierarchy of the action theories, we restrict the degree of nesting and thus preserve decidability.

Definition 5.57 (BoToM-MEP)**.** Given a BoToM-KB $\Sigma_0 \cup \Gamma$ and a goal α regressable wrt Γ, the BoToM-MEP problem is to decide whether a finite sequence z exists such that

$$\Sigma_0 \cup \Gamma \models Exec(z) \wedge [z]\alpha \tag{5.18}$$

A sequence z which fulfills (5.18) is called a *solution* of the MEP wrt Σ_0, Γ and α. Note that the DEL approach defines MEP based on the satisfaction of a single model and therefore requires complete specification of the initial state; and our definition is based on logical entailment, where multiple, even infinitely many models are involved. Thus, we accept incomplete specification.

Theorem 5.58. For $\sigma_1, \sigma_2 \in Z$ s.t. $\mathcal{R}[\Gamma, Exec(\sigma_1) \wedge [\sigma_1]\alpha]$ and $\mathcal{R}[\Gamma, Exec(\sigma_2) \wedge [\sigma_2]\alpha]$ are logically equivalent, if $z \cdot \sigma_1$ is a solution of the problem, then $z \cdot \sigma_2$ is also a solution.

Proof. Suppose that $z \cdot \sigma_1$ is a solution,

$$\begin{aligned}
&\Sigma_0 \cup \Gamma \models Exec(z \cdot \sigma_1) \wedge [z \cdot \sigma_1]\alpha \\
\Longleftrightarrow &\Sigma_0 \cup \Gamma \models Exec(z) \wedge [z]Exec(\sigma_1) \wedge [z \cdot \sigma_1]\alpha \\
\Longleftrightarrow &\Sigma_0 \cup \Gamma \models \mathcal{R}[\Gamma, Exec(z)] \\
&\qquad \wedge \mathcal{R}[\Gamma, z, Exec(\sigma_1) \wedge [\sigma_1]\alpha] && \text{(Thm. 5.53)} \\
\Longleftrightarrow &\Sigma_0 \cup \Gamma \models \mathcal{R}[\Gamma, Exec(z)] \\
&\qquad \wedge \mathcal{R}[\Gamma, z, Exec(\sigma_2) \wedge [\sigma_2]\alpha] && \text{(Cor. 5.54)} \\
\Longleftrightarrow &\Sigma_0 \cup \Gamma \models \mathcal{R}[\Gamma, Exec(z) \wedge [z]Exec(\sigma_2) \wedge [z \cdot \sigma_2]\alpha] \\
\Longleftrightarrow &\Sigma_0 \cup \Gamma \models Exec(z \cdot \sigma_2) \wedge [z \cdot \sigma_2]\alpha
\end{aligned}$$

□

Proposition 5.59. For $z = a \cdot \sigma$, $\Sigma_0 \models \mathcal{R}[\Gamma, Exec(z) \wedge [z]\alpha]$ if and only if $\Sigma_0 \models Poss(a) \wedge \mathcal{R}[\Gamma, a, \mathcal{R}[Exec(\sigma) \wedge [\sigma]\alpha]]$

With the above theorem and proposition, we are ready to prove the most important result of this work:

Theorem 5.60. BoToM-MEP is decidable.

Proof. Given a k-order theory Γ and α regressable wrt Γ. By Thm.5.53, to decide Eq.5.18 amounts to decide

$$\Sigma_0 \models \mathcal{R}[\Gamma, Exec(z) \wedge [z]\alpha] \tag{5.19}$$

Since both Σ_0 and the regression are static, it is reduced to $K45_n$ reasoning. By Thm.5.50, the regression function always returns a formula with depth bounded by k. As pointed out in [Hal95], for a fixed k and finite Ap, there are only finitely many inequivalent static formulae. Let the number of such formulae be $\mathcal{N}$. By Thm.5.58 and the Pigeonhole Principle, if (5.18) has a solution, there must be a solution z of length at most $\mathcal{N}$. Since Ac is also finite, by enumerating all action sequences up to length $\mathcal{N}$, one can decide if a solution exists. □

In fact, it is not needed to enumerate all sequences or to compute the exact number of $\mathcal{N}$. It can be solved more efficiently by Algorithm 1: A queue of action sequence candidates Cand is maintained and initialized with the empty one. A dictionary State, which maps action sequences to $\mathcal{MS}$ formulae, is maintained and initialized as $\{\langle\rangle : \alpha\}$. In each iteration of the while-loop, the algorithm pops out the first sequence z in Cand, generates a longer sequence $a \cdot z$ for each $a \in \mathsf{Ac}$, and tests whether Σ_0 entails the regression.[12] Line 13 tests if the regression is equivalent to a formula in State.[13] If the regression wrt $a \cdot z$ is equivalent to one wrt an explored sequence z', by Thm.5.58 the shortest solution expanded from $a \cdot z$ will not be shorter as those expanded from z'. Hence, we can prune $a \cdot z$. Since the size of State will be at most $\mathcal{N}$, either a solution is found, or eventually every

[12] By Prop.5.59 it suffices to perform a single step of regression.

[13] Though State could be huge, equivalence check can be efficient with proper knowledge compilation methods [Hal95].

regression can be found in State, then every sequence in Cand will be pruned and it terminates when the queue is empty.

Theorem 5.61. Algorithm 1 terminates and returns FALSE if the BoToM-MEP has no solution. When a solution exists, the algorithm returns a solution z such that for any other solution z', $|z| \leq |z'|$.

Algorithm 1 Breadth-First BoToM-MEP Search

function botommep(Σ_0, Γ, α):
  Cand $\leftarrow [\langle\rangle]$
  State$[\langle\rangle] \leftarrow \alpha$
  if $\Sigma_0 \models \mathcal{R}[\Gamma, \alpha]$ **then**
    return $\langle\rangle$
  end if
  while Cand not empty **do**
    $z \leftarrow$ POP(Cand)
    $\varphi \leftarrow$ State$[z]$
    for $a \in$ Ac **do**
      if $\Sigma_0 \models \mathcal{R}[\Gamma, Poss(a) \wedge [a]\varphi]$ **then**
        return $a \cdot z$
      else if $\mathcal{R}[\Gamma, Exec(a) \wedge [a]\varphi] \notin$ State **then**
        State$[a \cdot z] \leftarrow \mathcal{R}[\Gamma, Poss(a) \wedge [a]\varphi]$
        PUSH(Cand, $a \cdot z$)
      end if
    end for
  end while
  return False
end function

6

Joining Higher-order and Probability

In multi-agent systems, where agents constantly interact or coordinate, reasoning about knowledge and belief is of interest for many applications. By gathering detailed information, not only from the environment but also from other agents' mental states, an agent can determine when communication with other agents is necessary, enhance collaboration on tasks, and optimize its strategy against other agents in scenarios where agents operate concurrently.

Among many symbolic approaches that represent an agent's knowledge and belief, the notion of only-knowing [Lev90, LL01] is particularly valuable,

Numerous researches on single-agent only-knowing emerged, which have already been discussed in the previous chapters. Naturally, extending these works into the multi-agent scenario would result in an expressive account to represent and reason about the mental states of agents, and about how their minds change as a result of actions. However, research into multi-agent only-knowing has faced more obstacles than expected, and extensions in terms of probabilistic belief or belief after actions are rarely considered yet. Most studies are confined to characterizing only-knowing of categorical knowledge. Halpern and Lakemeyer attempted to handle the extension independently [Hal93, Lak93]. However, these accounts make use of arbitrary Kripke structures and lose the simplicity of Levesque's semantics. Furthermore, each account has some undesirable properties respectively. These properties are avoided in a joint work from Halpern and Lakemeyer [HL01], but it

forces us to have the semantic notion of validity directly in the language. For this reason, that proposal is not natural, and it is matched with a proof theory that has a set of new axioms to deal with these new notions. Belle and Lakemeyer revisited the construction and proposed an account $\mathcal{ONL}_n$ with a natural possible-world semantics[BL10].

Extensions in terms of probabilistic belief are rarely considered. The major difficulty is as follows: If we faithfully follow Levesque's principle of only-knowing, that the beliefs of an agent are precisely those following from its knowledge base, then in the multi-agent cases, only-knowing a sentence involves the belief and non-belief of sentences with an arbitrary depth of nesting beliefs. In a single-agent scenario, e.g. $\mathcal{OL}$, every formula specifies the agent's belief about the fact, or about its own belief, which we say to be of depth 1. In a multi-agent scenario, however, there are sentences describing an agent's belief about another agent's belief, which are of depth 2 at least. A sentence describing agent 1's belief about agent 2's belief about agent 3's belief is at least of depth 3. Such a belief nesting can be arbitrarily deep. If the depth is not restricted, constructing a structure which describes only-knowing up to all depths will be semantically difficult and will easily lead to a circular definition.

The study of belief with arbitrary depths of nesting is indeed meaningful, in particular for reasoning about common knowledge. However, we argue that it suffices to consider finite depths in many applications, where common knowledge or beliefs on an infinite set of sentences are not considered. We adapt the idea from Aucher and Belle [AB15] and use modalities of the form $\boldsymbol{O}_i^{(k)}$ to describe agent i's only-believing up to depth k. We focus on extending the notion of only-knowing to multi-agent, probabilistic cases and spare us from the troubles handling infinite nesting of beliefs.

The rest of the chapter is organized as follows. We begin with introducing the logic of multi-agent only-believing $\mathcal{OBL}_m$. Then we analyze the properties of the logic, in particular those exclusive for only-believing. Relation to other logical

accounts will also be discussed. Finally, we briefly conclude the chapter.

6.1. A Logic for Multi-agent Only-believing

We introduce the logic $\mathcal{OBL}_m$. It can be considered as a multi-agent extension of $\mathcal{OBL}$ [BLL16] and a first-order, probabilistic extension of $\mathcal{ML}$ in Section 5.1.

6.1.1. Syntax

The logic $\mathcal{OBL}_m$ is a first-order modal logic with equality. Let $\mathsf{Ag} = \{1, \ldots, n\}$ denote a set of n agents. The vocabulary consists of (FO) *variables* and *predicate symbols*. For simplicity, function symbols are excluded. The language includes a countable set of *standard names* $\mathcal{N}$. This can be viewed as having fixed infinite domain closure axioms with the *unique name assumption*, which further allows FO quantification to be understood substitutionally. The set of computable numbers $\mathbb{C}$ is included as a sub-sort of objects. We call a predicate other than $=$, applied to first-order variables or standard names, an *atomic* formula. An atomic formula without variables is called a *ground atom*. We use $\mathcal{P}$ to denote the set of ground atoms in $\mathcal{OBL}_m$.

Standard FO connectives $\wedge, \neg, \forall$ and epistemic modalities $\boldsymbol{B}_1, \ldots, \boldsymbol{B}_n$ are used to construct formulae. For each agent index i, $\boldsymbol{B}_i(\alpha\colon r)$ is read as "agent i believes α with degree r". To illustrate, a sentence $\boldsymbol{B}_1(\mathit{Rain}\colon 0.8)$ means agent 1 believes that there is an 80% chance of raining. In particular, we write $\boldsymbol{K}_i\alpha$ as an abbreviation of $\boldsymbol{B}_i(\alpha\colon 1)$ and read it as "agent i knows α".

To extend the notion of "only-believing" to multi-agent systems, we start from the single-agent cases and inductively extend the language to represent only-believing of any depth. We use modality $\boldsymbol{O}_i^{(1)}$ to specify agent i's only-believing up to depth 1, which captures the beliefs and non-beliefs about facts and the agent's own beliefs. An example of $\boldsymbol{O}_i^{(1)}$-formula is $\boldsymbol{O}_i^{(1)}(\exists x.[P(x) \wedge \neg \boldsymbol{K}_i P(x)])$, which means "All agent i believes up to depth 1 is the existence of an unknown object

x s.t. $P(x)$ holds." Now we go one step deeper: We use modality $\boldsymbol{O}_i^{(2)}$ to specify agent i's only-believing up to depth 2, which stands for the beliefs and non-beliefs about fact, about the agent's own beliefs and other agents' beliefs up to depth 1. An example is

$$\boldsymbol{O}_i^{(2)}(q \wedge \exists x.[\boldsymbol{K}_j(P(x)) \wedge \neg \boldsymbol{K}_i \boldsymbol{K}_j(P(x))])$$

with $i \neq j$. The difference between $\boldsymbol{O}_i^{(2)}$ and $\boldsymbol{O}_i^{(1)}$ is: While the former specifies agent i's beliefs about other agents' beliefs, the latter does not. For instance, $\boldsymbol{O}_i^{(2)}(p\colon 1)$ implies that agent i knows nothing about what other agents know or believe, e.g. $\boldsymbol{O}_i^{(2)}(p\colon 1)$ entails $\neg \boldsymbol{K}_i \boldsymbol{K}_j q$ for any proposition q, but $\boldsymbol{O}_i^{(1)}(p\colon 1)$ entails neither $\boldsymbol{K}_i \boldsymbol{K}_j q$ nor $\neg \boldsymbol{K}_i \boldsymbol{K}_j q$.

We include modalities $\boldsymbol{O}_1^{(k)}, \ldots \boldsymbol{O}_n^{(k)}$ with natural number $k > 0$. For each k, $\boldsymbol{O}_i^{(k+1)}$ specifies agent i's only-believing about other agents' beliefs up to depth k, and formula $\boldsymbol{O}_i^{(k)}(\alpha\colon r)$ is read as "All agent i believes up to depth k is α with degree r". We write $\boldsymbol{O}_i^{(k)}\alpha$ to mean $\boldsymbol{O}_i^{(k)}(\alpha\colon 1)$.

Definition 6.1 (i-depth)**.** For $i \in \mathsf{Ag}$, the i-depth of a formula α, written as $dep[\alpha, i]$, is inductively defined as

- $dep[\alpha, i] = 0$ for atomic formula α, $dep[t_1 = t_2, i] = 0$
- $dep[\neg\alpha, i] = dep[\alpha, i]$
- $dep[\alpha \wedge \beta, i] = max(dep[\alpha, i], dep[\beta, i])$
- $dep[\forall x.\alpha, i] = dep[\alpha, i]$
- $dep[\boldsymbol{B}_i(\alpha\colon r), i] = \max\{\max\{dep[\alpha, j] \mid j \neq i\} + 1, dep[\alpha, i]\}$
- $dep[\boldsymbol{B}_j(\alpha\colon r), i] = 0$ for $j \neq i$
- $dep[\boldsymbol{O}_i^{(k)}(\alpha\colon r), i] = k$
- $dep[\boldsymbol{O}_j^{(k)}(\alpha\colon r), i] = 0$ for $j \neq i$

For a $\boldsymbol{B}$-formula, the depth increases only when a nesting of different agents' beliefs occurs. For an $\boldsymbol{O}$-formula, the depth is given by the superscript k.

Example 6.2. Let α denote the formula $(\exists r.r > 0.5 \wedge \boldsymbol{B}_2(p\colon r))$. Consider the formula $\boldsymbol{K}_1\alpha$, namely, agent 1 knows that p is believed by agent 2 with a degree greater than 0.5. Then

$$\begin{aligned} dep[\boldsymbol{K}_1\alpha, 1] &= \max\{dep[\alpha, 1], 1 + dep[\alpha, 2]\} \\ &= \max\{0, 2\} = 2 \\ dep[\boldsymbol{K}_1\alpha, 2] &= 0 \end{aligned}$$

i.e. $\boldsymbol{K}_1\alpha$ has 1-depth 2 and 2-depth 0.

The set of well-formed formulae (wffs) is the smallest set including:

- any atomic formulae,
- $t = t'$ where t and t' are variables or standard names,
- if α, β are formulae, then $\alpha \wedge \beta$, $\neg\alpha$, and $\forall x.\alpha$ are formulae.
- if α is a formula, then for any $i \in \mathsf{Ag}$ and number r, $\boldsymbol{B}_i(\alpha\colon r)$ is a formula.
- if $\boldsymbol{B}_i(\alpha\colon r)$ has i-depth k, then for $l \geq k$, $\boldsymbol{O}_i^{(l)}(\alpha\colon r)$ is a formula.

An example of illegal formulae is $\boldsymbol{O}_1^{(1)}(p \wedge \boldsymbol{K}_2 q\colon 0.8)$, where the sentence being believed is of 1-depth 2. Only-believing this sentence requires at least $k = 2$. For the rest of the paper, we only consider wffs and proper superscript k.

Similar as usual and as in Chpt. 4, a formula with no free variables is called a *sentence*. A formula not mentioning modalities is said to be *objective*. A formula, where all predicate symbols appear within the scope of a modal operator is called *subjective*. Given $i \in \mathsf{Ag}$, a formula is called *i-objective* if modal operators $\boldsymbol{B}_i$ and $\boldsymbol{O}_i^{(k)}$ only appear within the scope of $\boldsymbol{B}_j$ or $\boldsymbol{O}_j^{(k')}$ where $j \neq i$. For instance, $p \wedge \boldsymbol{O}_2^{(3)}(q \wedge \boldsymbol{K}_1 q)$ is 1-objective since $\boldsymbol{K}_1$ only appears in the scope of $\boldsymbol{O}_2^{(3)}$, namely, knowing p is agent 2's conjecture of agent 1's mental state and not necessarily to

be the actual belief of agent 1. Inversely, a formula is called *i-subjective* if any predicates or modal operators $\boldsymbol{B}_j$ and $\boldsymbol{O}_j^{(k)}$ s.t. $j \neq i$ only appear within the scope of operators $\boldsymbol{B}_i$ or $\boldsymbol{O}_i^{(k')}$, e.g. $\neg \boldsymbol{K}_1 p \wedge \boldsymbol{B}_1(p \wedge \boldsymbol{K}_2 q\colon 0.5)$. *Basic* formulae are those which do not mention any only-knowing modalities $\boldsymbol{O}_i^{(k)}$. As before, we use TRUE to denote an objective tautology $\forall x.(x = x)$ and use FALSE to denote its negation.

6.1.2. Semantics

The semantics is given in terms of possible worlds, where a *world* w is defined as a set of ground atoms, i.e. $w \subseteq \mathcal{P}$. The set of all worlds is denoted as W.

In the logic $\mathcal{OBL}$, a distribution d is a function from W to non-negative real numbers $\mathbb{R}^{\geq 0}$ and an epistemic state is defined as a set of distributions.[1] We extend the notions to the multi-agent case: A distribution d defined above describes an agent's weighting to each possible world. For an agent i, to describe agent i's belief about other agents' beliefs on the actual world, the weight assigned to world w, namely $d(w)$, is partitioned and reassigned to tuples of form $(w, e_1, \ldots, e_{n-1})$ where each e_j is an $\mathcal{OBL}$ epistemic state describing the belief of an agent other than i. Formally, we have the following definition:

Definition 6.3 (Structure and Distribution)**.** For $k \geq 1$, k-structures, written as s^k, and k-distributions, written as d^k, are inductively defined as:

- $s^1 = (w, \underbrace{\{\varnothing\}, \ldots, \{\varnothing\}}_{n-1})$, where $w \in W$. We use $\mathcal{S}^1$ to denote the set of all 1-structures.
- $d^1 : \mathcal{S}^1 \to \mathbb{R}^{\geq 0}$. We use $\mathcal{D}^1$ to denote the set of all 1-distributions.
- For $k > 1$, $s^k = (w, e_1^{k-1}, \ldots, e_{n-1}^{k-1})$, where $w \in W$ and $e_j^{k-1} \subseteq \mathcal{D}^k$ for $j \in \{1, \ldots, n-1\}$. The set of all k-structures is denoted as $\mathcal{S}^k$.

[1] The idea to incorporate a set of distributions instead of a single distribution in the epistemic state derives from the philosophical stance that *de re* knowledge about degrees of belief should not be valid. Namely, if epistemic states only contain single distributions, formulae such as $\exists r.\boldsymbol{K}(\boldsymbol{B}(\phi\colon r))$ will be valid for any ϕ, which is counter-intuitive.

- $d^k : \mathcal{S}^k \to \mathbb{R}^{\geq 0}$. We use $\mathcal{D}^k$ to denote the set of all k-distributions.

When the context is clear, we omit the superscript and write d instead of d^k. We say e is an *epistemic state*, specifically a k-epistemic state, if e is a set of k-distributions, i.e. $e \subseteq \mathcal{D}^k$. The number k is said to be the depth of e, written as $dep[e] = k$.

Similar to $\mathcal{OBL}$, we appeal to three conditions Bnd, Eq, Norm to obtain probability distributions.

Definition 6.4. Let d be a k-distribution, $\mathcal{U}, \mathcal{V} \subseteq \mathcal{S}^k$ and $\mathcal{U} \subseteq \mathcal{V}$, r a real number. We define conditions Bnd, Eq and Norm as follows:

- $\textsc{Bnd}(d, \mathcal{U}, r)$ iff there is no $m \in \mathbb{N}$ and $\{s_1, \ldots, s_m\} \subseteq \mathcal{U}$ such that
$$\sum_{j=1}^{m} d(s_j) > r,$$
- $\textsc{Eq}(d, \mathcal{U}, r)$ iff $\textsc{Bnd}(d, \mathcal{U}, r)$ and no $r' < r$ s.t. $\textsc{Bnd}(d, \mathcal{U}, r')$;
- $\textsc{Norm}(d, \mathcal{U}, \mathcal{V}, r)$ iff there is $b \neq 0$ such that $\textsc{Eq}(d, \mathcal{U}, b \times r)$ and $\textsc{Eq}(d, \mathcal{V}, b)$.

$\textsc{Bnd}(d, \mathcal{U}, r)$ ensures the weight of $\mathcal{U}$ wrt d is bounded by r. $\textsc{Eq}(d, \mathcal{U}, r)$ ensures that the weight is bounded and r is the supremum. Given $\textsc{Norm}(d, \mathcal{U}, \mathcal{V}, r)$, r can be viewed as the normalized sum of the weight of worlds in $\mathcal{U}$ in relation to $\mathcal{V}$. Essentially, although the distribution d is defined over an uncountable domain, these conditions on d admit a well-defined summation, and the weights on worlds can indeed be interpreted as probabilities. Similar with [BLL16], we have

Theorem 6.5. Let d be a k-distribution, $\mathcal{U} = \{s \in \mathcal{S}^k \mid d(s) \neq 0\}$ and $\mathcal{U} \subseteq \mathcal{V}$. For any $b \geq 0$, if $\textsc{Bnd}(d, \mathcal{V}, b)$, then $\mathcal{U}$ is countable.

Proof. Let $\mathcal{U}_j = \{s \in \mathcal{U} \mid d(s) \geq 1/j\}$ for $j \in \mathbb{N}^+$. It is easy to see that $\mathcal{U} = \bigcup \mathcal{U}_i$. Suppose that $\mathcal{U}$ is uncountable, there is a $N > 0$ such that $\mathcal{U}_N$ is infinite (Otherwise we could enumerate $\mathcal{U}$ by enumerating $\mathcal{U}_j$ starting at $j = 1$). Consider any countably infinite sequence $\{s_j\}_{j=1}^{\infty}$ taken from $\mathcal{U}_N$. Since $d(s_j) \geq 1/N$ for

all j, the sum $\sum_{l=1}^{\infty} d(s_j)$ is unbounded, contradicting the assumption that $\mathcal{V}$ is bounded. □

By satisfying the BND conditions, a distribution assigns non-zero values to a countable support set $Supp(d) = \{s \mid d(s) \neq 0\}$. For the rest of this chapter, by summation of weights of a bounded distribution d, we always mean summation in its (countable) support set, i.e.

$$\sum_{s \in E} d(s) \doteq \sum_{s \in E \cap Supp(d)} d(s)$$

In $\mathcal{OBL}$, the necessitation rule does not hold for modality $\boldsymbol{B}$ (also $\boldsymbol{K}$). For instance, TRUE is valid in $\mathcal{OBL}$ yet $\boldsymbol{K}$TRUE is not. Besides, an empty epistemic state is legal in $\mathcal{OBL}$, which breaks the laws of probability (e.g. $\boldsymbol{B}(p\colon 0.1) \wedge \boldsymbol{B}(p\colon 0.2)$ will be satisfied by the empty state). Fortunately, these undesirable properties can be easily avoided by ruling out improper distributions and epistemic states.

Definition 6.6 (Regularity)**.** We inductively define **regular** k-distributions $\mathcal{D}^k$ and **regular** k-epistemic states $\mathcal{E}^k$: Let $\mathcal{E}^0 = \{\varnothing\}$. For any $k > 0$,

- $\mathcal{D}^k = \{d \mid \text{EQ}(d, \mathcal{S}^k, 1)\}$;
- $\mathcal{E}^k = 2^{\mathcal{D}^k} \backslash \{\varnothing\}$.

For the rest of this chapter, we consider only regular distributions and epistemic states. Now we define the truth of sentences. By a model, we mean a tuple $(w, \vec{e})$, where $\vec{e} = (e_1, \ldots, e_n)$ and each e_i is an epistemic state (e_i denotes the i-th argument of $\vec{e}$). We say a formula α and $\vec{e}$ are *compatible* if $dep[\alpha, i] \leq dep[e_i]$ f.a. $i \in \{1, \ldots, n\}$. The truth value of objective sentences is assigned as follows:

- $w, \vec{e} \models P(n)$ iff $P(n) \in w$;
- $w, \vec{e} \models t_1 = t_2$ iff t_1 and t_2 are identical standard name;
- $w, \vec{e} \models \neg\alpha$ iff not $w, \vec{e} \models \alpha$;
- $w, \vec{e} \models \alpha \wedge \beta$ iff $w, \vec{e} \models \alpha$ and $w, \vec{e} \models \beta$;

- $w, \vec{e} \models \forall x.\alpha$ iff $w, \vec{e} \models \alpha^x_m$ for any $m \in \mathcal{N}$;

Here α^x_m means the formula obtained by substituting each appearance of free variable x in α by a standard name m. The semantics for the objective fragment is identical to language $\mathcal{L}$ [LL01]. Supposing that $\vec{e}$ is compatible with $\boldsymbol{B}_i(\alpha : r)$, then

- $w, \vec{e} \models \boldsymbol{B}_i(\alpha : r)$ iff f.a. $d \in e_i$, $\textsc{Norm}(d, W^{i,e_i}_{\alpha}, W^{i,e_i}_{\textsc{True}}, r)$,

where for $i \in \mathsf{Ag}$, $e \in \mathcal{E}^k$ and formula φ, $W^{i,e}_{\varphi}$ is defined as

$$W^{i,e}_{\varphi} = \{s \in \mathcal{S}^k \mid s \cup_i e \models \varphi\} \tag{6.1}$$

When the context is clear, we write $W_{\textsc{True}}$ rather $W^{i,e}_{\textsc{True}}$.

Given an epistemic state of depth k, we can define the semantics of only-believing up to depth k:

- For $e_i \in \mathcal{E}^k$, $w, \vec{e} \models \boldsymbol{O}^{(k)}_i(\alpha : r)$ iff f.a. $d \in \mathcal{D}^k$,

$$d \in e_i \textbf{ iff } \textsc{Norm}(d, W^{i,e_i+d}_{\alpha_j}, W_{\textsc{True}}, r_j).$$

For a sentence α and a set of sentences Σ, we write $\Sigma \models \alpha$ (read as Σ logically entails α) to mean that for every model $(w, e_1, \ldots, e_n)$ compatible with α and all sentences in Σ, if $w, e_1, \ldots, e_n \models \alpha'$ f.a. $\alpha' \in \Sigma$, then $w, e_1, \ldots, e_n \models \alpha$. We say α is valid (written as $\models \alpha$) if $\{\} \models \alpha$. Furthermore, when α is objective, we write $w \models \alpha$; When α is subjective, we write $e_1, \ldots, e_m \models \alpha$ (or $\vec{e} \models \alpha$); For i-subjective α, we write $e_i \models \alpha$.

We still need to assign truth values for only-believing up to depth k, while the model is of a greater depth (otherwise sentences like $\boldsymbol{O}^{(1)}_1 p \wedge \boldsymbol{O}^{(2)}_1(p \wedge \boldsymbol{K}_2 p)$ are unsatisfiable). To do so, we need to capture a model's belief of lower depths and omit the depths greater than needed. This is achieved via the notion which we call "*(order) reduction*":

Definition 6.7 (Reduction)**.** We inductively define the reduction of any distributions, structures, and epistemic states:

- For $d \in \mathcal{D}^2$ and $d' \in \mathcal{D}^1$, d' is the reduction of d, written as $d' = \mathcal{R}[d]$, iff for all w,
$$d'(w, \{\varnothing\}, \ldots, \{\varnothing\}) = \sum_{e_1,\ldots,e_{n-1}} d(w, e_1, \ldots e_{n-1})$$

- For $e \in \mathcal{E}^2$ and $e' \in \mathcal{E}^1$, e' is the reduction of e, written as $e' = \mathcal{R}[e]$, iff $e' = \{\mathcal{R}[d] \,|\, d \in e\}$. For structure $s = (w, e_1, \ldots, e_{n-1})$, the reduction is defined as $\mathcal{R}[s] = (w, \mathcal{R}[e_1], \ldots, \mathcal{R}[e_{n-1}])$.

- For $d \in \mathcal{D}^{k+1}$ and $d' \in \mathcal{D}^k$, d' is the reduction of d, written as $d' = \mathcal{R}[d]$, iff for all $s' \in \mathcal{S}^k$,
$$d'(s') = \sum_{\{s \in \mathcal{S}^{k+1} \,|\, \mathcal{R}[s] = s'\}} d(s')$$

- For $e \in \mathcal{E}^{k+1}$ and $e' \in \mathcal{E}^k$, e' is the reduction of e, written as $e' = \mathcal{R}[e]$, iff $e' = \{\mathcal{R}[d] \,|\, d \in e\}$

By definition, for each $e \in \mathcal{E}^{k+1}$, there exists a unique $e' \in \mathcal{E}^k$ s.t. $e' = \mathcal{R}[e]$. We write $e' = \mathcal{R}^{(2)}[e]$ to mean $e' = \mathcal{R}[\mathcal{R}[e]]$. Analogously, $e' = \mathcal{R}^{(k)}[e]$ means $e' = \mathcal{R}[\cdots \mathcal{R}[e] \cdots]$ with k nested $\mathcal{R}$. The following lemma indicates that $\mathcal{R}[e]$ faithfully reflects the properties of e in lower depths. Therefore, for $e \in \mathcal{E}^{k+1}$, it is reasonable to assign truth values for only-believing up to depth k based on the truth assignment of $\mathcal{R}[e]$.

Lemma 6.8. Given $\vec{e}$ and $\vec{e}'$ s.t. $e'_i = \mathcal{R}[e_i]$ f.a. $i \in \{1, \ldots, n\}$. For any basic formula α compatible with $\vec{e}'$, $w, \vec{e} \models \alpha$ iff $w, \vec{e}' \models \alpha$.

Proof. We prove the lemma via induction.

Basis:

For objective α, the proof is trivial as the truth value is irrelevant to $\vec{e}$ and $\vec{e}'$.

Induction hypothesis:

Suppose that the statement holds for all $\vec{e}, \vec{e}'$ and α' s.t. $e'_i = \mathcal{R}[e_i]$, $e'_i \in \mathcal{E}^l$ for $l < k$ and α' compatible with $\vec{e}'$.

- Induction on $\wedge$, $\neg$ or $\forall$ is trivial.

- For $e_i \in \mathcal{E}^k$, $e_i \models \boldsymbol{B}_i(\alpha\colon r)$ iff f.a. $d \in e_i$,[2]

$$\sum_{s \in W_\alpha^{i,e_i}} d(s) = r \tag{\#}$$

We introduce an auxiliary function for the proof: For $s \in \mathcal{S}^k$ and $s' \in \mathcal{S}^{k-1}$ we define

$$\mathbb{I}(s, s') = \begin{cases} 1 & s' = \mathcal{R}[s]; \\ 0 & \text{otherwise.} \end{cases}$$

Since the reduction is unique, for each s there's only one s' with $\mathbb{I}(s, s') = 1$. Thus when s is fixed,

$$\sum_{s'} \mathbb{I}(s, s') = 1$$

Back to (#), let $d' = \mathcal{R}[d]$, we have

$$\begin{aligned} & \sum_{s \in W_\alpha^{i,e_i}} d(s) \\ = & \sum_{s \in W_\alpha^{i,e_i}} \big(d(s) \cdot \sum_{s'} \mathbb{I}(s, s')\big) && \text{(Eq. 6.1.2)} \\ = & \sum_{\{(s,s') \,|\, s \in W_\alpha^{i,e_i}\}} d(s) \cdot \mathbb{I}(s, s') \\ = & \sum_{\{(s,s') \,|\, s \in W_\alpha^{i,e_i},\ \mathcal{R}[s]=s'\}} d(s) && \text{(Def. } \mathbb{I}\text{)} \end{aligned}$$

Let $e_i' = \mathcal{R}[e_i]$. By induction hypothesis we have $s \cup_i e_i \models \alpha$ iff $s' \cup_i e_i' \models \alpha$, i.e. $s \in W_\alpha^{i,e_i}$ iff $s' \in W_\alpha^{i,e_i'}$. Therefore,

$$\begin{aligned} \sum_{s \in W_\alpha^{i,e_i}} d(s) = & \sum_{\{(s,s') \,|\, s' \in W_\alpha^{i,e_i'},\ \mathcal{R}[s]=s'\}} d(s) && \text{(IH)} \\ = & \sum_{s' \in W_\alpha^{i,e_i'}} \big(\sum_{\{s \,|\, \mathcal{R}[s]=s'\}} d(s')\big) \\ = & \sum_{s' \in W_\alpha^{i,e_i'}} d'(s') && \text{(Def. 6.7)} \end{aligned}$$

[2] By definition, the truth of $\boldsymbol{B}_i(\alpha\colon r)$ is irrelevant to w and epistemic states of other agents except i. Thus we write $e_i \models \boldsymbol{B}_i(\alpha\colon r)$ instead of $w, \vec{e} \models \boldsymbol{B}_i(\alpha\colon r)$.

Thus $\textsc{Norm}(d', W_\alpha^{i,\mathcal{R}[e_i]}, W_{\textsc{True}}^{i,\mathcal{R}[e_i]}, r)$ f.a. $d' \in \mathcal{R}[e_i]$, i.e. $\mathcal{R}[e_i] \models \boldsymbol{B}_i(\alpha : r)$

□

Now we can complete the semantics:

- For $e_i \in \mathcal{E}^l$ with $l > k$, $w, \vec{e} \models \boldsymbol{O}_i^{(k)}(\alpha : r)$ iff
  $w, e_1, \ldots, R[e_i], \ldots, e_n \models \boldsymbol{O}_i^{(k)}(\alpha : r)$.

The result of Lem.6.8 can be extended to all wffs including those with only-believing:

Theorem 6.9. Given $\vec{e}, \vec{e}'$ s.t. $e_i' = \mathcal{R}[e_i]$. For any formula α compatible with $\vec{e}'$, then $w, \vec{e} \models \alpha$ iff $w, \vec{e}' \models \alpha$

The proof is similar to Lem 6.8. The other difference is the induction on $\boldsymbol{O}$-formulae, which directly follows the definition of the semantics.

6.2. Properties of the Logic

As many accounts in the previous chapters, $\mathcal{OBL}_m$ satisfies the $K(D)45_n$ properties, the Barcan formula [HC68], and properties of belief and only-believing in the single-agent cases. We prove these to demonstrate that the logic $\mathcal{OBL}_m$ has reasonable properties, rather than seeing them as the contribution of this work.

6.2.1. Knowledge

$\mathcal{OBL}_m$ satisfies the $KD45_n$ properties:

- (Necessitation) If $\models \alpha$, then $\models \boldsymbol{K}_i\alpha$.

  Proof. Suppose that $\models \alpha$, then for all e_i and $d \in e_i$, $\sum_{s \in W_\alpha^{i,e_i}} d(s) = 1$.
  Thus $e_i \models \boldsymbol{K}_i\alpha$ □

- (Consistency) $\models \boldsymbol{K}_i\alpha \supset \neg\boldsymbol{K}_i\neg\alpha$

Proof. Suppose that $e_i \models \boldsymbol{K}_i\alpha$, then for any $d \in e_i$,

$$\begin{aligned}\sum_{s\in W^{i,e_i}_{\neg\alpha}} d(s) &= \sum_{s\in W_{\mathrm{TRUE}}\setminus W^{i,e_i}_{\alpha}} d(s)\\ &= 1 - \sum_{s\in W^{i,e_i}_{\alpha}} d(s) \neq 1\end{aligned}$$

Thus $e_i \models \neg\boldsymbol{K}_i\neg\alpha$.

□

- (Distribution) $\models \boldsymbol{K}_i\alpha \wedge \boldsymbol{K}_i(\alpha \supset \beta) \supset \boldsymbol{K}_i\beta$

  Proof. For any e_i compatible with $\boldsymbol{K}_i\alpha$ and $\boldsymbol{K}_i(\alpha \supset \beta)$, if $e_i \models \boldsymbol{K}_i\alpha \wedge \boldsymbol{K}_i(\alpha \supset \beta)$, then for all $d \in e_i$,

$$\begin{aligned}\sum_{s\in W^{i,e_i}_{\beta}} d(s) &= \sum_{s\in W^{i,e_i}_{\alpha}} d(s) + \sum_{s\in W^{i,e_i}_{\alpha\supset\beta}} d(s) - \sum_{s\in W^{i,e_i}_{\alpha\vee\alpha\supset\beta}} d(s)\\ &= (1 + 1 - 1) = 1\end{aligned}$$

  i.e. $e_i \models \boldsymbol{K}_i\beta$. □

- (Pos. Introspection) $\models \boldsymbol{K}_i\alpha \supset \boldsymbol{K}_i\boldsymbol{K}_i\alpha$

  Proof. Suppose that $e_i \models \boldsymbol{K}_i\alpha$, then for all s, $s \cup_i e_i \models \boldsymbol{K}_i\alpha$, i.e. $s \in W^{i,e_i}_{\boldsymbol{K}_i\alpha}$ iff $s \in W^{i,e_i}_{\mathrm{TRUE}}$. For all $d \in e_i$,

$$\sum_{s\in W^{i,e_i}_{\boldsymbol{K}_i\alpha}} d(s) = 1$$

  i.e. $e_i \models \boldsymbol{K}_i\boldsymbol{K}_i\alpha$ □

- (Neg. Introspection) $\models \neg\boldsymbol{K}_i\alpha \supset \boldsymbol{K}_i\neg\boldsymbol{K}_i\alpha$

  Proof. Suppose that $e_i \models \neg\boldsymbol{K}_i\alpha$. Similar to the proof above, $s \in W^{i,e_i}_{\neg\boldsymbol{K}_i\alpha}$ iff $s \in W^{i,e_i}_{\mathrm{TRUE}}$. For all $d \in e_i$,

$$\sum_{s\in W^{i,e_i}_{\neg\boldsymbol{K}_i\alpha}} d(s) = 1$$

i.e. $e_i \models \boldsymbol{K}_i \neg \boldsymbol{K}_i \alpha$. □

Barcan formula:

- $\models \forall x.\boldsymbol{K}_i \alpha \supset \boldsymbol{K}_i \forall x.\alpha$

  Proof. Suppose that $e_i \models \forall x.\boldsymbol{K}_i \alpha$. By definition we have $e_i \models \boldsymbol{K}_i \alpha^x_m$ for any $m \in \mathcal{N}$. Then f.a. $d \in e_i$,

  $$\sum_{s \in W^{i,e_i}_{\alpha^x_m}} d(s) = 1 \text{ f.a. } m \in \mathcal{N}$$

  Namely,

  $$\sum_{s \in W^{i,e_i}_{\neg\alpha^x_m}} d(s) = 0 \text{ f.a. } m \in \mathcal{N}$$

  Therefore

  $$\begin{aligned} \sum_{s \in W^{i,e_i}_{\forall x.\alpha}} d(s) &= 1 - \sum_{s \in W^{i,e_i}_{\exists x.\neg\alpha}} d(s) \\ &\geq 1 - \sum_{m \in \mathcal{N}} \sum_{s \in W^{i,e_i}_{\neg\alpha^x_n}} d(s) = 1 \end{aligned}$$

  Thus $e_i \models \boldsymbol{K}_i \forall x.\alpha$ □

  It is worth mentioning that the converse $\models \boldsymbol{K}_i \forall x.\alpha \supset \forall x.\boldsymbol{K}_i \alpha$ holds as well.

- $\models \exists x.\boldsymbol{K}_i \alpha \supset \boldsymbol{K}_i \exists x.\alpha$

  The proof is of the same spirit as the universal one.

- $\not\models \boldsymbol{K}_i \exists x.\alpha \supset \exists x.\boldsymbol{K}_i \alpha$.

  Let $w, w' \in W$ satisfy $P(n_1) \in w$, $P(n') \notin w$ f.a. $n' \neq n_1$, $P(n_2) \in w'$, $P(n') \notin w'$ f.a. $n' \neq n_2$. Let $d_i \in \mathcal{D}^1$ assign weight 0.5 to both w and w'. Other worlds are assigned 0. Then $\{d_i\}$ satisfies $\boldsymbol{K}_i \exists x.\alpha$ but not $\exists x.\boldsymbol{K}_i \alpha$.

6.2.2. Degree of Belief

For any $i \in \mathsf{Ag}$ and formula α, β,

- $\models \boldsymbol{B}_i(\alpha\colon r) \supset \neg \boldsymbol{B}_i(\alpha\colon r')$ for $r' \neq r$
- if $\models \alpha \equiv \beta$, then $\models \boldsymbol{B}_i(\alpha\colon r) \equiv \boldsymbol{B}_i(\beta\colon r)$ for any r.
- $\models \boldsymbol{B}_i(\alpha\colon r) \supset \boldsymbol{B}_i(\neg\alpha\colon 1-r)$
- $\models \boldsymbol{B}_i(\alpha \wedge \beta\colon r) \wedge \boldsymbol{B}_i(\alpha \wedge \neg\beta\colon r') \supset \boldsymbol{B}_i(\alpha\colon r + r')$
- $\models \boldsymbol{B}_i(\alpha\colon r) \wedge \boldsymbol{B}_i(\beta\colon r') \wedge \boldsymbol{B}_i(\alpha \wedge \beta\colon r'') \supset \boldsymbol{B}_i(\alpha \vee \beta\colon n)$ where n is a standard name of sort number and $n = r + r' - r''$.

Most of the properties can be proved in a way similar to those of the previous subsection. Here we only provide the proof of the last one:

Proof. Suppose that $e_i \models \boldsymbol{B}_i(\alpha\colon r) \wedge \boldsymbol{B}_i(\beta\colon r') \wedge \boldsymbol{B}_i(\alpha \wedge \beta\colon r'')$, then f.a. $d \in e_i$,

$$\sum_{s \in W_\alpha^{i,e_i}} d(s) = r; \qquad \sum_{s \in W_\beta^{i,e_i}} d(s) = r'; \qquad \sum_{s \in W_{\alpha\wedge\beta}^{i,e_i}} d(s) = r''. \tag{6.2}$$

By applying the laws of set,

$$W_{\alpha\vee\beta}^{i,e_i} = W_\alpha^{i,e_i} + W_{\neg\alpha\wedge\beta}^{i,e_i} = W_\alpha^{i,e_i} + W_\beta^{i,e_i} - W_{\alpha\wedge\beta}^{i,e_i}$$

Thus $\sum_{s \in W_{\alpha\vee\beta}^{i,e_i}} d(s) = (r + r' - r'')$ □

Results of introspection also hold for degrees of belief. Furthermore, it can be extended to arbitrary i-subjective sentences.

- $\models \boldsymbol{B}_i(\alpha\colon r) \supset \boldsymbol{K}_i \boldsymbol{B}_i(\alpha\colon r)$
- $\models \neg\boldsymbol{B}_i(\alpha\colon r) \supset \boldsymbol{K}_i \neg\boldsymbol{B}_i(\alpha\colon r)$
- For any i-subjective formula α, $\models \alpha \supset \boldsymbol{K}_i\alpha$

6.3. Only-believing with Higher-order Beliefs

Here we discuss the properties of only-believing. To begin with, what is only-believed should be believed at first:

Proposition 6.10. $\models \boldsymbol{O}_i^{(k)}(\alpha\colon r) \supset \boldsymbol{B}_i(\alpha\colon r)$

Proof. W.l.o.g. we consider $e_i \models \boldsymbol{O}_i^{(k)}(\alpha\colon r)$, i.e. for any $d \in \mathcal{D}^k$, $d \in e_i$ iff $\textsc{Norm}(d, W_\alpha^{i,e_i+d}, W_{\textsc{True}}^{i,e_i+d}, r)$. For $d \in e_i$, $e_i + d = e_i$, thus $e \models \boldsymbol{B}(\alpha\colon r)$. □

If an agent only believes α with degree r up to a certain depth, then she should also only believe it up to a lower (but compatible) depth. The converse result does not necessarily hold. Formally,

Proposition 6.11. $\models \boldsymbol{O}_i^{(k+1)}(\alpha\colon r) \supset \boldsymbol{O}_i^{(k)}(\alpha\colon r)$.

Proof. The proof is similar to Lem. 6.8 with function $\mathbb{I}$ as used in the lemma. Suppose that $e_i \models \boldsymbol{O}_i^{(k+1)}(\alpha\colon r)$, w.l.o.g. we assume that $e_i \in \mathcal{E}^{k+1}$. By Thm 6.9 it suffices to prove that $e_i' \models \boldsymbol{O}_i^{(k)}(\alpha\colon r)$ for $e_i' = \mathcal{R}[e_i]$. For $d \in \mathcal{D}^{k+1}$, let d' be the k-distribution s.t. $d' = \mathcal{R}[d]$. Then

$$
\begin{aligned}
\sum_{s \in W_\alpha^{i,e_i+d}} d(s) &= \sum_{s \in W_\alpha^{i,e_i+d}} \big(d(s) \cdot \sum_{s'} \mathbb{I}(s,s')\big) && \text{(Eq. 6.1.2)} \\
&= \sum_{\{(s,s') \,|\, s \in W_\alpha^{i,e_i+d}\}} d(s) \cdot \mathbb{I}(s,s') \\
&= \sum_{\{(s,s') \,|\, s \in W_\alpha^{i,e_i+d},\ s'=\mathcal{R}[s]\}} d(s) && \text{(Def. } \mathbb{I}\text{)} \\
&= \sum_{\{(s,s') \,|\, s' \in W_\alpha^{i,e_i'+d'},\ s'=\mathcal{R}[s]\}} d(s) && \text{(Thm. 6.9)} \\
&= \sum_{s' \in W_\alpha^{i,e_i'+d'}} \Big(\sum_{\{s \,|\, s'=\mathcal{R}[s]\}} d(s)\Big) \\
&= \sum_{s' \in W_\alpha^{i,e_i'+d'}} d'(s') && \text{(Def. 6.7)}
\end{aligned}
$$

Some explanation for the fourth step: $s \in W_\alpha^{i,e_i}$ iff $s \cup_i e_i \models \alpha$. Since $e_i' = \mathcal{R}[e_i]$ and $s' = \mathcal{R}[s]$, by Thm. 6.9 we have $s' \cup_i e_i' \models \alpha$, i.e. $s' \in W_\alpha^{i,e_i'}$.
For all $d' \in e_i'$, there is $d \in e_i$ such that $d' = \mathcal{R}[d]$, then $\sum_{s' \in W_\alpha^{i,e_i'+d'}} d'(s') = r$.
Likewise, for $d' \notin e_i'$, there is $d \notin e_i$ such that $d' = \mathcal{R}[d]$, i.e. $\sum_{s' \in W_\alpha^{i,e_i'+d'}} d'(s') \neq r$.
Thus $e_i' \models \boldsymbol{O}_i^{(k)}(\alpha : r)$. □

There exists α s.t. $\not\models \boldsymbol{O}_i^{(k)}(\alpha : r) \supset \boldsymbol{O}_i^{(k+1)}(\alpha : r)$. An example can be easily constructed: Let $e_i \in \mathcal{E}^2$ and $e_i \models \boldsymbol{O}_1^{(2)}(p \wedge \boldsymbol{K}_2 p)$, then $e_i \models \boldsymbol{O}_1^{(1)} p$ but $e_i \not\models \boldsymbol{O}_1^{(2)} p$.

The proposition demonstrates that from $\boldsymbol{O}_i^{(k)}$ to $\boldsymbol{O}_i^{(k+1)}$ it is indeed a more precise specification of an agent's only-believing.

In general, an $\boldsymbol{O}$-formula could be satisfied by more than one epistemic state. For example $\boldsymbol{O}_1^{(1)}\big((p \wedge \boldsymbol{K}_1 p) \vee (\neg p \wedge \boldsymbol{K}_1 \neg p)\big)$. For i-objective sentence α, however, there is a unique k-epistemic state $e_i \in \mathcal{E}^k$ which satisfies $\boldsymbol{O}_i^{(k)}\alpha$:

Lemma 6.12. Given i-objective α and k such that $\boldsymbol{O}_i^{(k)}\alpha$ is satisfiable, there is precisely one $e_i \in \mathcal{E}^k$ s.t. $e_i \models \boldsymbol{O}_i^{(k)}\alpha$.

Proof. Given $e_i \in \mathcal{E}^k$, by the definition, $e_i \models \boldsymbol{O}_i^{(k)}\alpha$ iff f.a. $d \in \mathcal{D}^k$, $d \in e_i$ iff

$$\sum_{s \in W_\alpha^{i,e_i+d}} d(s) = 1$$

Since α is i-objective, if $s \cup_i (e_i + d) \models \alpha$, then f.a. $\tilde{e}_i$, $s \cup_i \tilde{e}_i \models \alpha$. For any $\tilde{e}_i \in \mathcal{E}^k$ s.t. $\tilde{e}_i \models \boldsymbol{O}_i^{(k)}\alpha$, it is trivial that f.a. $d \in \mathcal{D}^k$, $d \in e_i$ iff $d \in \tilde{e}_i$, i.e. $e_i = \tilde{e}_i$. □

For a model of a depth greater than needed, the uniqueness of the model can be interpreted as follows: Suppose that $e_i \in \mathcal{E}^k$ is the unique model satisfying $\boldsymbol{O}_i^{(k)}\alpha$. Then for all $e_i^* \in \mathcal{E}^{k'}$ s.t. $k' > k$ and $e_i^* \models \boldsymbol{O}_i^{(k)}\alpha$, e_i^* will reduce to e_i after finite steps of reduction:

Lemma 6.13. Given i-objective α and k such that $\boldsymbol{O}_i^{(k)}\alpha$ is satisfiable. Let $e_i \in \mathcal{E}^k$, $e_i \models \boldsymbol{O}_i^{(k)}\alpha$. For any e_i^* such that $e_i^* \models \boldsymbol{O}_i^{(k)}\alpha$, $e_i = \mathcal{R}^{(j)}[e_i^*]$ for some j.

Proof. For any $k' > k$ and $e_i^* \in \mathcal{E}^{k'}$, e_i^* reduces to $\tilde{e}_i \in \mathcal{E}^k$ after $k' - k$ steps of reduction. Given $e_i^* \models \boldsymbol{O}_i^{(k)}\alpha$, by applying Thm. 6.9 we have $\tilde{e}_i \models \boldsymbol{O}_i^{(k)}\alpha$. By Lem. 6.12, $\tilde{e}_i$ is unique, i.e. $\tilde{e}_i = e_i$. □

In general, the results can be extended to the case with degree of belief:

Theorem 6.14. Given i-objective α and r such that $\boldsymbol{O}_i^{(k)}(\alpha\colon r)$ is satisfiable, there is precisely one $e_i \in \mathcal{E}^k$ such that $e_i \models \boldsymbol{O}_i^{(k)}(\alpha\colon r)$. For any e_i^* such that $e_i^* \models \boldsymbol{O}_i^{(k)}(\alpha\colon r)$, $e_i = \mathcal{R}^{(j)}[e_i^*]$ for some j.

The proof is similar to Lem. 6.12 and Lem. 6.13.

A nice property of the uniqueness is that, given what is only-believed by an agent, everything else being believed or not believed by the agent is logically implied. Formally, we have the following results:

Theorem 6.15. Given i-objective α s.t. $\boldsymbol{O}_i^{(k)}(\alpha\colon r)$ is satisfiable, and arbitrary β with $dep[\boldsymbol{B}_i(\beta\colon r'), i] \leq k$. Then for any $0 < r' < 1$, either $\boldsymbol{B}_i(\beta\colon r')$ or $\neg\boldsymbol{B}_i(\beta\colon r')$ is entailed by $\boldsymbol{O}_i^{(k)}(\alpha\colon r)$, and

- $\models \boldsymbol{O}_i^{(k)}(\alpha\colon r) \supset \boldsymbol{B}_i(\beta\colon r')$ iff $\models \boldsymbol{B}_i(\alpha\colon r) \supset \boldsymbol{B}_i(\beta\colon r')$
- $\models \boldsymbol{O}_i^{(k)}(\alpha\colon r) \supset \neg\boldsymbol{B}_i(\beta\colon r')$ iff $\not\models \boldsymbol{B}_i(\alpha\colon r) \supset \boldsymbol{B}_i(\beta\colon r')$

Proof. By Thm. 6.14, it exists a unique $e_i \in \mathcal{E}^k$ s.t. $e_i \models \boldsymbol{O}_i^{(k)}(\alpha\colon r)$.
If $\boldsymbol{B}_i(\alpha\colon r) \supset \boldsymbol{B}_i(\beta\colon r')$ is valid, by Prop. 6.10, $\boldsymbol{O}_i^{(k)}(\alpha\colon r)$ entails $\boldsymbol{B}_i(\beta\colon r')$. When $\boldsymbol{B}_i(\alpha\colon r) \supset \boldsymbol{B}_i(\beta\colon r')$ is invalid, there exists $d_i \in \mathcal{D}^k$ s.t. $\{d_i\} \models \boldsymbol{B}_i(\alpha\colon r) \wedge \neg\boldsymbol{B}_i(\beta\colon r')$. Thus, $d_i \in e_i$ and hencefore $e_i \models \neg\boldsymbol{B}_i(\beta\colon r')$, $\models \boldsymbol{O}_i^{(k)}(\alpha\colon r) \supset \neg\boldsymbol{B}_i(\beta\colon r')$.
For any $e_i' \in \mathcal{E}^{k'}$ s.t. $k' > k$, $e_i' \models \boldsymbol{O}_i^{(k)}(\alpha\colon r)$, By Thm. 6.9 and 6.14, $e_i' \models \boldsymbol{B}_i(\beta\colon r')$ iff $e_i \models \boldsymbol{B}_i(\beta\colon r')$. □

Theorem 6.16. Given i-objective α s.t. $\boldsymbol{O}_i^{(k)}(\alpha\colon r)$ is satisfiable, and arbitrary β with $dep[\boldsymbol{K}_i\beta, i] \leq k$, then either $\boldsymbol{O}_i^{(k)}(\alpha\colon r) \supset \boldsymbol{K}_i\beta$ or $\boldsymbol{O}_i^{(k)}(\alpha\colon r) \supset \neg\boldsymbol{K}_i\beta$ is valid, and

- $\models \boldsymbol{O}_i^{(k)}(\alpha\colon r) \supset \boldsymbol{K}_i\beta$ iff $\models \boldsymbol{B}_i(\alpha\colon r) \supset \boldsymbol{K}_i\beta$ or $\models \boldsymbol{O}_i^{(k)}(\alpha\colon r) \supset \beta$
- $\models \boldsymbol{O}_i^{(k)}(\alpha\colon r) \supset \neg\boldsymbol{K}_i\beta$ iff $\not\models \boldsymbol{B}_i(\alpha\colon r) \supset \boldsymbol{K}_i\beta$ and $\not\models \boldsymbol{O}_i^{(k)}(\alpha\colon r) \supset \beta$

Given distinct propositions p, q, to determine whether $\boldsymbol{O}_i^{(k)} p$ entails $\boldsymbol{K}_i \neg \boldsymbol{K}_i q$, it suffices to check the validity of $\boldsymbol{K}_i p \supset \boldsymbol{K}_i \neg \boldsymbol{K}_i q$ and $\boldsymbol{O}_i^{(k)} p \supset \neg \boldsymbol{K}_i q$. The latter is valid since $\not\models \boldsymbol{K}_i p \supset \boldsymbol{K}_i q$ and $\not\models \boldsymbol{O}_i^{(k)} p \supset q$. Thus $\models \boldsymbol{O}_i^{(k)} p \supset \boldsymbol{K}_i \neg \boldsymbol{K}_i q$.

Note that this is not true for $\boldsymbol{K}_i$, e.g. neither $\boldsymbol{K}_i p \supset \boldsymbol{K}_i \neg \boldsymbol{K}_i q$ nor $\boldsymbol{K}_i p \supset \neg \boldsymbol{K}_i \neg \boldsymbol{K}_i q$ is valid. Also, we would like to reiterate that the results of Thm.6.15 and Thm.6.16 hold only under the restriction of depth and do not hold when $dep[\boldsymbol{K}_i \beta, i] > k$. For example, both $\boldsymbol{O}_1^{(1)} p \wedge \boldsymbol{K}_1 (p \wedge \boldsymbol{K}_2 p)$ and $\boldsymbol{O}_1^{(1)} p \wedge \boldsymbol{K}_1 (p \wedge \neg \boldsymbol{K}_2 p)$ are satisfiable.

If β is also i-objective, we can reduce the use of the outermost modalities:

Corollary 6.17. Given i-objective α, β, s.t. $\boldsymbol{O}_i^{(k)} \alpha$ is satisfiable, then

- $\models \boldsymbol{O}_i^{(k)} \alpha \supset \boldsymbol{K}_i \beta$ iff $\models \alpha \supset \beta$
- $\models \boldsymbol{O}_i^{(k)} \alpha \supset \neg \boldsymbol{K}_i \beta$ iff $\not\models \alpha \supset \beta$

In the conference proceeding version of the logic [FL24], the modalities of only-believing are defined in the form $\boldsymbol{O}_i(\alpha_1 : r_1, \ldots, \alpha_m : r_m)$, where the pairs of formula and degree of belief are enumerated. The main reason for such a design is that, following the property which also exists in $\mathcal{OL}$ and $\mathcal{OBL}$, only-believing an i-subjective formula, e.g. $\boldsymbol{O}_i(\bigwedge_{j=1}^m \boldsymbol{B}_i(\varphi_j : r_j))$, can only be satisfied trivially, either by the complete state $\mathcal{D}^k$ or by the empty epistemic state. In the semantics presented in this thesis, such a restriction is lifted, and the results above can be extended to only-believing a collection of belief formulae:

Theorem 6.18. Let $\Phi = \bigwedge_{j=1}^m \boldsymbol{B}_i(\varphi_j : r_j)$ be the conjunction of a finite set of B_i-formulae, where for each j, φ_j is i-objective, and has no free variables other than x. Then Thm. 6.14, Thm. 6.15 and Thm 6.16 hold when replacing the i-objective formula α with Φ, and r with 1.

6.4. Discussion

The logical accounts of the closest relation to $\mathcal{OBL}_m$ are $\mathcal{OBL}$ [BLL16] and $\mathcal{ONL}_n$ [BL10]. $\mathcal{OBL}$ is a first-order, single-agent account of subjective probability and only-believing[BLL16], which is an extension of the logic $\mathcal{OL}$ proposed by Levesque [Lev90, LL01]. For cases where at most one agent is involved, $\mathcal{OBL}_m$ reduces to $\mathcal{OBL}$—with exceptions: Besides the difference in only-believing a subjective formula, which we just discussed in 6.3, the empty epistemic state is legal in $\mathcal{OBL}$, and distributions, where the weightings of worlds do not form a proper probability distribution, are not excluded. The former setting breaks the consistency, e.g. $\boldsymbol{K}\textsc{True} \wedge \boldsymbol{K}\textsc{False}$ is satisfiable in $\mathcal{OBL}$, and the latter breaks the necessitation rules, e.g. $\textsc{True}$ is valid in $\mathcal{OBL}$ but $\boldsymbol{K}\textsc{True}$ is not. We argue that omitting these structures in the semantics has no loss of expressiveness or generality. On the other hand, by the exclusion of these structures, the logic will obtain the complete $KD45_n$ properties, while only simple, straightforward restrictions are required.

$\mathcal{ONL}_n$ proposed by Belle and Lakemeyer[BL10, BL15b] is a first-order modal language for multi-agent only-knowing, where operators $\boldsymbol{L}_i, \boldsymbol{N}_i$ are used to describing what agent i "at least knows" and "at most knows". Only-knowing is considered as an abbreviation, e.g. $\boldsymbol{O}_i\alpha \equiv \boldsymbol{L}_i\alpha \wedge \boldsymbol{N}_i\neg\alpha$. The k-distributions or k-epistemic states in this paper are inspired by their work. One might notice that in $\mathcal{ONL}_n$, the superscript (k) is not required for only-knowing, since "validity is not affected if models of a depth greater than that needed are used". Namely, for a semantic proof, it could suffice to check only the properties of models of the minimal depth. A proof of the claim is given in Lemma 4 and Theorem 5 of [BL10]. However, the Lemma only holds for basic formulae. For non-basic formulae, counterexamples can be found:

Proposition 6.19. Let $\models^*$ be the satisfaction relation in $\mathcal{ONL}_n$, then there is an epistemic state e_a^2 for agent a such that $e_a^2 \not\models^* \boldsymbol{O}_a p$, but $e_a^2 \downarrow_1^2 \models^* \boldsymbol{O}_a p$

Proof. We define $W_p = \{w \mid w \models p\}$ and $W_q = \{w \mid w \models q\}$. Namely, W_p (W_q) is

the set of world, where p (q) is satisfied. We consider an $\mathcal{ONL}_n$ epistemic state $e_a^2 := \{(w, e_b^1) \mid w \in W_p, e_b^1 \subseteq W_q\}$. Then $e_a^2 \not\models^* \boldsymbol{O}_a p$ but $e_a \downarrow_1^2 \models^* \boldsymbol{O}_a p$ □

Whether Theorem 5 in [BL10] holds for non-basic formulae is still open. Hence, we restrict our account by specifying the depth of only-believing with the superscript. It also corresponds to the notion of ToM introduced in Sec. 5.2, that an agent is a limited reasoner which cannot reason about beliefs of unbounded orders. For $k \geq 2$, $\boldsymbol{O}_1^{(k)} p$ means that p is all agent 1 knows up to depth k, where the depth for $\boldsymbol{K}_1 \boldsymbol{K}_2 q$ is also included. We have $\models \boldsymbol{O}_1^{(k)} p \supset \neg \boldsymbol{K}_1 \boldsymbol{K}_2 q$.[3]

[3] I would like to take the opportunity to correct a statement in [FL24] that "neither $\boldsymbol{O}_1 p \supset \boldsymbol{K}_1 \boldsymbol{K}_2 q$ nor $\boldsymbol{O}_1 p \supset \neg \boldsymbol{K}_1 \boldsymbol{K}_2 q$ is valid in $\mathcal{ONL}_n$", which is wrong.

7 Conclusion

Based on a family of logical accounts with the set-based semantics, this thesis advances the research on probabilistic and higher-order beliefs. In terms of representation, the logics we proposed, including $\mathcal{OBL}_c$ and $\mathcal{PS}$, introduce measure-theoretic settings into the set-based semantics of modal logics, thereby enriching the expressive power of probabilistic beliefs. In terms of reasoning, in particular higher-order beliefs, we provide decidable solutions for both static and dynamic modal reasoning while, in parallel, deepen our understanding of the connections between dynamic modal logics, epistemic logics, and their corresponding fragments in conventional first-order logic. A unified account that integrates probabilistic and higher-order reasoning is also proposed, paving the way for generalizing tools for interfacing with symbolic and probabilistic knowledge bases.

Some possibilities for future work are as follows:

- While a unified formalism of probabilistic only-believing is given in the thesis, it lacks the mechanism of non-monotonic reasoning. As Lakemeyer and Levesque showed in [LL05a], the concept of only-knowing is capable of characterizing different variants of non-monotonic reasoning, including Reiter's default reasoning[Rei80], Moore's autoepistemic logic[Moo85], and the variant from Konolige[Kon88]. We believe that a framework for probabilistic default reasoning would be highly valuable, as it could bridge the commu-

nities of formal reasoning and machine learning, enabling us to augment a symbolic knowledge base with probabilistic defaults provided by statistical models.

- While the regression method, for the first time, provides a decidable solution to the multi-agent epistemic planning problem, regression is inefficient for long-lived agents. Additionally, as a backward-searching method, regression may explore redundant, unreachable states and is inconvenient for investigating the system's temporal properties. Motivated by these, we are interested in progression reasoning. Previous work, including [LL11, LF21], provides evidence for the possibility of a progression method for higher-order, even probabilistic higher-order reasoning.

- When a progression method is obtained, it is possible to encode the system into a logical program [LL22] and verify its temporal probability. This will further serve as the foundation for more complicated yet practical research areas, such as epistemic strategy/protocol synthesis and epistemic games.

Besides the aforementioned, straightforwardly feasible directions for further research, as a long-term goal I would like to respond to other relevant discplines: To integrate agents into society, we plan to explore the formalization of concepts in social science, such as Jürgen Habermas' *theory of communicative actions*[Hab84], or the *Universalization Principle* of moral norm in *discourse ethics*[Hab90, Hab92]:

"(A norm is valid if only if) all affected can accept the consequences and the side effects its general observance can be anticipated to have for the satisfaction of everyone's interests, and these consequences are preferred to those of known alternative possibilities" (Jürgen Habermas, [Hab90])

This could be achieved by introducing additional modalities (for notions such as acceptance and preference), and by formulating the principle within an epistemic variant of the Situation Calculus.

Bibliography

[AB13] Guillaume Aucher and Thomas Bolander. *Undecidability in epistemic planning*. PhD thesis, INRIA, 2013. (Cited on page 128.)

[AB15] Guillaume Aucher and Vaishak Belle. Multi-agent only knowing on planet kripke. In *Proceedings of the 24th International Conference on Artificial Intelligence*, pages 2713–2719, 2015. (Cited on pages 13 and 134.)

[AE89] Yves Auffray and Patrice Enjalbert. Modal theorem proving: an equational viewpoint. In *Proceedings of the 11th international joint conference on Artificial intelligence-Volume 1*, pages 441–445, 1989. (Cited on page 83.)

[Auc11] Guillaume Aucher. Del-sequents for progression. *Journal of Applied Non-Classical Logics*, 21(3-4):289–321, 2011. (Cited on pages 9 and 109.)

[Auc12] Guillaume Aucher. Del-sequents for regression and epistemic planning. *Journal of Applied Non-Classical Logics*, 22(4):337–367, 2012. (Cited on pages 9 and 109.)

[BA11] Thomas Bolander and Mikkel Birkegaard Andersen. Epistemic planning for single-and multi-agent systems. *Journal of Applied Non-Classical Logics*, 21(1):9–34, 2011. (Cited on pages 9, 108 and 109.)

[Bac89] Fahiem Ivor Bacchus. Representing and reasoning with probabilistic knowledge, 1989. (Cited on page 8.)

[BBvDM17] Chitta Baral, Thomas Bolander, Hans van Ditmarsch, and Sheila McIlrath. Epistemic planning (dagstuhl seminar 17231). *Dagstuhl Reports*, 7(6):1–47, 2017. (Cited on page 108.)

[Bel23] Vaishak Belle. Actions, continuous distributions and meta-beliefs. In *The 22nd International Conference on Autonomous Agents and Multiagent Systems, 2023*, 2023. (Cited on page 74.)

[BG07] Torben Brauner and Silvio Ghilardi. First order modal logic. *STUDIES IN LOGIC AND PRACTICAL REASONING*, 3:549–620, 2007. (Cited on pages 103 and 106.)

[BGPS22] Chitta Baral, Gregory Gelfond, Enrico Pontelli, and Tran Cao Son. An action language for multi-agent domains. *Artificial Intelligence*, 302:103601, 2022. (Cited on page 124.)

[BHL99a] Fahiem Bacchus, Joseph Y Halpern, and Hector J Levesque. Reasoning about noisy sensors and effectors in the situation calculus. *Artificial Intelligence*, 111(1-2):171–208, 1999. (Cited on pages 10, 46 and 115.)

[BHL99b] Fahiem Bacchus, Joseph Y Halpern, and Hector J Levesque. Reasoning about noisy sensors and effectors in the situation calculus. *Artificial Intelligence*, 111(1-2):171–208, 1999. (Cited on page 52.)

[BJS15] Thomas Bolander, Martin Holm Jensen, and François Schwarzentruber. Complexity results in epistemic planning. In *24th International Joint Conference on Artificial Intelligence*, pages 2791–2797. AAAI Press, 2015. (Cited on page 109.)

[BL10] Vaishak Belle and Gerhard Lakemeyer. Multi-agent only-knowing revisited. In *Twelfth International Conference on the Principles of Knowledge Representation and Reasoning*, pages 49–59, Toronto, 2010. AAAI Press. (Cited on pages 5, 7, 13, 84, 86, 134, 152 and 153.)

[BL14a] Vaishak Belle and Gerhard Lakemeyer. Multiagent only knowing in dynamic systems. *Journal of Artificial Intelligence Research*, 49:363–402, 2014. (Cited on pages 109, 110 and 112.)

[BL14b] Vaishak Belle and Hector Levesque. How to progress beliefs in continuous domains. In *Proceedings of the Fourteenth International Conference on the Principles of Knowledge Representation and Reasoning*, pages 438–447, Vienna, 2014. AAAI Press. (Cited on page 11.)

[BL15a] Vaishak Belle and Gerhard Lakemeyer. Only knowing meets common knowledge. In *Proceedings of the Twenty-Fourth International Joint Conference on Artificial Intelligence, IJCAI 2015, Buenos Aires, Argentina, July 25-31, 2015*, pages 2755–2761, 2015. (Cited on pages 5 and 13.)

[BL15b] Vaishak Belle and Gerhard Lakemeyer. Semantical considerations on multiagent only knowing. *Artificial Intelligence*, 223:1–26, 2015. (Cited on pages 7, 84, 86, 88, 102, 103, 112 and 152.)

[BL17] Vaishak Belle and Gerhard Lakemeyer. Reasoning about

probabilities in unbounded first-order dynamical domains. In *Twenty-Sixth International Joint Conference on Artificial Intelligence*, pages 828–836, Melbourne, 2017. IJCAI. (Cited on pages 8, 10, 13, 46, 53, 66, 70 and 79.)

[BL18] Vaishak Belle and Hector J Levesque. Reasoning about discrete and continuous noisy sensors and effectors in dynamical systems. *Artificial Intelligence*, 262:189–221, 2018. (Cited on pages 10, 13, 41, 46 and 52.)

[BL20] Vaishak Belle and Hector J Levesque. Regression and progression in stochastic domains. *Artificial Intelligence*, 281:103247, 2020. (Cited on page 10.)

[BLL16] Vaishak Belle, Gerhard Lakemeyer, and Hector Levesque. A first-order logic of probability and only knowing in unbounded domains. In *Proceedings of the Thirtieth AAAI Conference on Artificial Intelligence*, page 893–899, Phoenix, Arizona, 2016. AAAI Press. (Cited on pages 8, 31, 78, 135, 139 and 152.)

[Bol17] Thomas Bolander. A gentle introduction to epistemic planning: The del approach. *arXiv preprint arXiv:1703.02192*, 2017. (Cited on pages 9 and 82.)

[Cat91] Laurent Catach. Tableaux: A general theorem prover for modal logics. *Journal of automated reasoning*, 7:489–510, 1991. (Cited on page 82.)

[Che80] Brian F Chellas. *Modal logic: an introduction*. Cambridge university press, 1980. (Cited on pages 6, 17, 81, 82, 84 and 89.)

[Cla13] Jens Claßen. *Planning and verification in the agent language Golog*. PhD thesis, Aachen, Techn. Hochsch., Diss., 2013, 2013. (Cited on page 113.)

[DDN05] Stéphane Demri and Hans De Nivelle. Deciding regular grammar logics with converse through first-order logic. *Journal of Logic, Language and Information*, 14:289–329, 2005. (Cited on page 83.)

[Den89] Daniel C Dennett. *The intentional stance*. MIT press, 1989. (Cited on page 1.)

[DGMR+16] Giuseppe De Giacomo, Aniello Murano, Sasha Rubin, Antonio Di Stasio, et al. Imperfect-information games and generalized planning. In *IJCAI*, pages 1037–1043, 2016. (Cited on page 108.)

[DNSH00] Hans De Nivelle, Renate A Schmidt, and Ulrich Hustadt. Resolution-based methods for modal logics. *Logic Journal of the IGPL*, 8(3):265–292, 2000. (Cited on page 82.)

[End01] Herbert B Enderton. *A mathematical introduction to logic*. Elsevier, 2001. (Cited on pages 103 and 105.)

[FH94] Ronald Fagin and Joseph Y Halpern. Reasoning about knowledge and probability. *Journal of the ACM (JACM)*, 41(2):340–367, 1994. (Cited on page 31.)

[FHM90] Ronald Fagin, Joseph Y Halpern, and Nimrod Megiddo. A logic for reasoning about probabilities. *Information and computation*, 87(1-2):78–128, 1990. (Cited on page 8.)

[FHMV04] Ronald Fagin, Joseph Y Halpern, Yoram Moses, and Moshe Vardi. *Reasoning about knowledge*. MIT press, Cambridge, 2004. (Cited on page 36.)

[Fin75] Kit Fine. Some connections between elementary and modal logic. In *Studies in Logic and the Foundations of Mathematics*, volume 82, pages 15–31. Elsevier, 1975. (Cited on page 83.)

[FL13] Liangda Fang and Yongmei Liu. Multiagent knowledge and belief change in the situation calculus. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 27, pages 304–312, 2013. (Cited on page 109.)

[FL24] Qihui Feng and Gerhard Lakemeyer. Probabilistic multi-agent only-believing. In *Proceedings of the 23rd International Conference on Autonomous Agents and Multiagent Systems*, pages 571–579, 2024. (Cited on pages 14, 151 and 153.)

[FLBL23] Qihui Feng, Daxin Liu, Vaishak Belle, and Gerhard Lakemeyer. A logic of only-believing over arbitrary probability distributions. In *The 22nd International Conference on Autonomous Agents and Multiagent Systems, 2023*, 2023. (Cited on page 68.)

[FLLB23] Qihui Feng, Daxin Liu, Gerhard Lakemeyer, and Vaishak Belle. A logic of only-believing over arbitrary probability distributions. In *Proceedings of the 2023 International Conference on Autonomous Agents and Multiagent Systems*, page 355–363. International Foundation for Autonomous Agents and Multiagent Systems, May 2023. (Cited on pages 14, 25, 30 and 36.)

[FLW15] Liangda Fang, Yongmei Liu, and Ximing Wen. On the progression of knowledge and belief for nondeterministic actions in the situation calculus. In *IJCAI*, pages 2955–2963, 2015. (Cited on page 109.)

[FWKL25] Qihui Feng, Hannah Wilk, Shakil M Khan, and Gerhard Lakemeyer. Translating multi-agent modal logics of knowledge and belief into decidable first-order fragments. In *Proceedings of the 24th International Conference on Autonomous Agents and*

Multiagent Systems, pages 740–748, 2025. (Cited on pages 111, 112 and 114.)

[GL07] Alfredo Gabaldon and Gerhard Lakemeyer. Esp: A logic of only-knowing, noisy sensing and acting. In *Proceedings of the Twenty-Second AAAI Conference on Artificial Intelligence*, pages 974–979, Vancouver, 2007. AAAI Press. (Cited on page 29.)

[GMV99] Harald Ganzinger, Christoph Meyer, and Margus Veanes. The two-variable guarded fragment with transitive relations. In *Proceedings. 14th Symposium on Logic in Computer Science (Cat. No. PR00158)*, pages 24–34. IEEE, 1999. (Cited on page 83.)

[GO99] Erich Grädel and Martin Otto. On logics with two variables. *Theoretical computer science*, 224(1-2):73–113, 1999. (Cited on page 83.)

[Grä99] Erich Grädel. On the restraining power of guards. *The Journal of Symbolic Logic*, 64(4):1719–1742, 1999. (Cited on page 107.)

[Hab84] Jürgen Habermas. *The Theory of Communicative Action, Volume 1: Reason and the Rationalization of Society*. Beacon Press, Boston, MA, 1984. English edition of *Theorie des kommunikativen Handelns*. (Cited on page 156.)

[Hab90] Jürgen Habermas. *Moral Consciousness and Communicative Action*. Studies in Contemporary German Social Thought. MIT Press, Cambridge, MA, 1990. Develops discourse ethics and the universalization principle. (Cited on page 156.)

[Hab92] Jürgen Habermas. *Between Facts and Norms: Contributions to a Discourse Theory of Law and Democracy*. MIT Press,

Cambridge, MA, 1992. Extends communicative action theory into law and democracy. (Cited on page 156.)

[Hal90] Joseph Y Halpern. An analysis of first-order logics of probability. *Artificial intelligence*, 46(3):311–350, 1990. (Cited on page 8.)

[Hal93] Joseph Y Halpern. Reasoning about only knowing with many agents. In *Proceedings of the eleventh national conference on Artificial intelligence*, pages 655–661, 1993. (Cited on page 133.)

[Hal95] Joseph Y Halpern. The effect of bounding the number of primitive propositions and the depth of nesting on the complexity of modal logic. *Artificial Intelligence*, 75(2):361–372, 1995. (Cited on pages 85 and 129.)

[HC68] George Edward Hughes and M. J. Cresswell. *An Introduction to Modal Logic*. Methuen, London, England, 1968. (Cited on pages 6, 17 and 144.)

[Hin62] Jaakko Hintikka. Knowledge and belief: An introduction to the logic of the two notions, 1962. (Cited on pages 3, 5, 6 and 82.)

[HL01] Joseph Y Halpern and Gerhard Lakemeyer. Multi-agent only knowing. *Journal of Logic and Computation*, 11(1):41–70, 2001. (Cited on pages 7, 13 and 133.)

[HM85] Joseph Y Halpern and Y Moses. A guide to the modal logics of knowledge and belief: Preliminary draft. *contract*, 39(82-C):0250, 1985. (Cited on page 7.)

[HM90] Joseph Y Halpern and Yoram Moses. Knowledge and common knowledge in a distributed environment. *Journal of the ACM (JACM)*, 37(3):549–587, 1990. (Cited on pages 82 and 88.)

[HM92] Joseph Y Halpern and Yoram Moses. A guide to completeness and complexity for modal logics of knowledge and belief. *Artificial intelligence*, 54(3):319–379, 1992. (Cited on pages 6, 81, 82, 88, 89 and 92.)

[HSC22] Mark K Ho, Rebecca Saxe, and Fiery Cushman. Planning with theory of mind. *Trends in Cognitive Sciences*, 26(11):959–971, 2022. (Cited on page 109.)

[Kon88] Kurt Konolige. On the relation between default and autoepistemic logic. *Artificial intelligence*, 35(3):343–382, 1988. (Cited on pages 7 and 155.)

[Kri63] Saul A Kripke. Semantical analysis of modal logic i. normal modal propositional calculi. *Mathematical Logic Quarterly*, 9(5-6):67–96, 1963. (Cited on pages 3, 5, 6, 7, 17, 82 and 89.)

[KV13] Laura Kovács and Andrei Voronkov. First-order theorem proving and vampire. In *International Conference on Computer Aided Verification*, pages 1–35. Springer, 2013. (Cited on page 83.)

[Lad77] Richard E Ladner. The computational complexity of provability in systems of modal propositional logic. *SIAM journal on computing*, 6(3):467–480, 1977. (Cited on pages 6, 82 and 88.)

[Lak93] Gerhard Lakemeyer. All they know: A study in multi-agent autoepistemic reasoning. In *ijcai*, volume 93, pages 376–381, 1993. (Cited on page 133.)

[LC24] Daxin Liu and Jens Claßen. First-order progression beyond local-effect and normal actions. In *33rd International Joint Conference on Artificial Intelligence*, pages 3475–3483. IJCAI Organization, 2024. (Cited on page 11.)

[LC25] Daxin Liu and Jens Claßen. On action theories with iterable first-order progression. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pages 15041–15048, 2025. (Cited on page 11.)

[Lev82] Hector Joseph Levesque. A formal treatment of incomplete knowledge bases, 1982. (Cited on pages 5, 7 and 86.)

[Lev90] Hector J Levesque. All i know: a study in autoepistemic logic. *Artificial intelligence*, 42(2-3):263–309, 1990. (Cited on pages 5, 7, 13, 86, 133 and 152.)

[LF21] Daxin Liu and Qihui Feng. On the progression of belief. In *Proceedings of the International Conference on Principles of Knowledge Representation and Reasoning*, volume 18, pages 465–474, online event, 2021. AAAI Press. (Cited on pages 8, 13, 39, 46, 53 and 156.)

[LF23] Daxin Liu and Qihui Feng. On the progression of belief. *Artificial Intelligence*, 322:103947, 2023. (Cited on pages 5, 10 and 113.)

[LFBL23] Daxin Liu, Qihui Feng, Vaishak Belle, and Gerhard Lakemeyer. Concerning measures in a first-order logic with actions and meta-beliefs. In *20th International Conference on Principles of Knowledge Representation and Reasoning*, pages 451–460. IJCAI Inc, 2023. (Cited on pages 14 and 25.)

[LL01] Hector J Levesque and Gerhard Lakemeyer. *The logic of knowledge bases*. MIT Press, Cambridge, 2001. (Cited on pages 13, 17, 18, 32, 36, 37, 43, 69, 76, 82, 86, 102, 103, 105, 133, 141 and 152.)

[LL04] Gerhard Lakemeyer and Hector J Levesque. Situations, si!

situation terms, no! In *KR*, pages 516–526, 2004. (Cited on pages 10 and 13.)

[LL05a] Gerhard Lakemeyer and Hector J Levesque. Only-knowing: Taking it beyond autoepistemic reasoning. In *AAAI*, volume 5, pages 633–638, 2005. (Cited on pages 7, 13 and 155.)

[LL05b] Yongmei Liu and Hector J Levesque. Tractable reasoning with incomplete first-order knowledge in dynamic systems with context-dependent actions. In *IJCAI*, volume 5, pages 522–527, 2005. (Cited on page 11.)

[LL06] Gerhard Lakemeyer and Hector J Levesque. Towards an axiom system for default logic. In *proceedings of the national conference on artificial intelligence*, volume 21, page 263. Menlo Park, CA; Cambridge, MA; London; AAAI Press; MIT Press; 1999, 2006. (Cited on pages 7 and 13.)

[LL10] Yongmei Liu and Gerhard Lakemeyer. On first-order definability and computability of progression for local-effect actions and beyond. Schloss Dagstuhl–Leibniz-Zentrum für Informatik, 2010. (Cited on page 11.)

[LL11] Gerhard Lakemeyer and Hector J Levesque. A semantic characterization of a useful fragment of the situation calculus with knowledge. *Artificial Intelligence*, 175(1):142–164, 2011. (Cited on pages 5, 10, 13, 66, 118, 156 and 187.)

[LL21] Daxin Liu and Gerhard Lakemeyer. Reasoning about beliefs and meta-beliefs by regression in an expressive probabilistic action logic. In *Proceedings of the Thirtieth International Joint Conference on Artificial Intelligence*, pages 1951–1958, Montreal, 2021. IJCAI. (Cited on pages 8, 10, 37, 39, 46, 52 and 53.)

[LL22] Daxin Liu and Gerhard Lakemeyer. On the verification of belief programs. *arXiv preprint arXiv:2204.12562v3*, 2022. (Cited on page 156.)

[LPW11] Benedikt Löwe, Eric Pacuit, and Andreas Witzel. Del planning and some tractable cases. In *Logic, Rationality, and Interaction: Third International Workshop, LORI 2011, Guangzhou, China, October 10-13, 2011. Proceedings 3*, pages 179–192. Springer, 2011. (Cited on page 109.)

[LR97] Fangzhen Lin and Ray Reiter. How to progress a database. *Artificial Intelligence*, 92(1-2):131–167, 1997. (Cited on pages 9 and 11.)

[McC10] William McCune. Prover9 and mace4. http://www.cs.unm.edu/ mccune/prover9/, 2005–2010. (Cited on page 83.)

[MH81] John McCarthy and Patrick J Hayes. Some philosophical problems from the standpoint of artificial intelligence. In *Readings in artificial intelligence*, pages 431–450. Elsevier, 1981. (Cited on pages 9 and 109.)

[Moo85] Robert C Moore. Semantical considerations on nonmonotonic logic. *Artificial intelligence*, 25(1):75–94, 1985. (Cited on pages 7, 43 and 155.)

[Mor76] Charles G Morgan. Methods for automated theorem proving in nonclassical logics. *IEEE Transactions on Computers*, 100(8):852–862, 1976. (Cited on page 83.)

[Non93] Andreas Nonnengart. First-order modal logic theorem proving and functional simulation. In *Proceedings of the 13th*

international joint conference on Artifical intelligence-Volume 1, pages 80–85, 1993. (Cited on page 83.)

[Pel23] Claudie Peloquin. *Representing Minds Representing Minds: An Examination of the Association between Recursive Mental State Attributions and Executive Processes.* PhD thesis, Open Access Te Herenga Waka-Victoria University of Wellington, 2023. (Cited on page 109.)

[Pie03] Ahti-Veikko Pietarinen. What do epistemic logic and cognitive science have to do with each other? *Cognitive systems research*, 4(3):169–190, 2003. (Cited on page 108.)

[Rei80] Raymond Reiter. A logic for default reasoning. *Artificial intelligence*, 13(1-2):81–132, 1980. (Cited on pages 7, 43 and 155.)

[Rei91] Raymond Reiter. The frame problem in the situation calculus: A simple solution (sometimes) and a completeness result for goal regression. *Artificial intelligence and mathematical theory of computation: papers in honor of John McCarthy*, 50:359–380, 1991. (Cited on pages 9 and 10.)

[Rei01] Raymond Reiter. *Knowledge in action: logical foundations for specifying and implementing dynamical systems.* MIT press, Cambridge, 2001. (Cited on pages 9, 10, 26, 72 and 109.)

[Rud21] Walter Rudin. Principles of mathematical analysis. 2021. (Cited on page 18.)

[Sch99] Renate A Schmidt. Decidability by resolution for propositional modal logics. *Journal of Automated Reasoning*, 22:379–396, 1999. (Cited on page 82.)

[SH03] Renate A Schmidt and Ullrich Hustadt. A principle for incorporating axioms into the first-order translation of modal formulae. In *International Conference on Automated Deduction*, pages 412–426. Springer, 2003. (Cited on page 83.)

[SH07] Renate A Schmidt and Ullrich Hustadt. The axiomatic translation principle for modal logic. *ACM Transactions on Computational Logic (TOCL)*, 8(4):19–es, 2007. (Cited on page 83.)

[SH13] Renate A Schmidt and Ullrich Hustadt. First-order resolution methods for modal logics. In *Programming Logics: Essays in Memory of Harald Ganzinger*, pages 345–391. Springer, 2013. (Cited on page 82.)

[SL03] Richard B Scherl and Hector J Levesque. Knowledge, action, and the frame problem. *Artificial Intelligence*, 144(1-2):1–39, 2003. (Cited on page 10.)

[SM73] Larry J Stockmeyer and Albert R Meyer. Word problems requiring exponential time (preliminary report). In *Proceedings of the fifth annual ACM symposium on Theory of computing*, pages 1–9, 1973. (Cited on page 7.)

[SS09] Elias M Stein and Rami Shakarchi. *Real analysis: measure theory, integration, and Hilbert spaces*. Princeton University Press, Princeton, 2009. (Cited on pages 18, 19, 20, 40 and 60.)

[StC13] Luc Segoufin and Balder ten Cate. Unary negation. *Logical Methods in Computer Science*, 9, 2013. (Cited on page 83.)

[TCS11] Balder Ten Cate and Luc Segoufin. Unary negation. In *28th International Symposium on Theoretical Aspects of Computer*

Science (STACS 2011). Schloss-Dagstuhl-Leibniz Zentrum für Informatik, 2011. (Cited on page 83.)

[Tur36] Alan Mathison Turing. On computable numbers, with an application to the entscheidungsproblem. *J. of Math*, 58:345–363, 1936. (Cited on pages 27 and 49.)

[Var96] Moshe Y Vardi. Why is modal logic so robustly decidable? *Descriptive complexity and finite models*, 31:149–184, 1996. (Cited on page 83.)

[vB77] Johan van Benthem. Modal correspondence theory. 1977. (Cited on page 83.)

[vB97] Johan van Benthem. Dynamic bits and pieces. 1997. (Cited on page 101.)

[VDvdHHK15] Hans Van Ditmarsch, Wiebe van der Hoek, Joseph Y Halpern, and Barteld Kooi. *Handbook of epistemic logic*. College Publications, 2015. (Cited on page 82.)

[Vit05] Giuseppe Vitali. Sul problema della misura dei gruppi di punti di una retta, bologna, tip. *Camberini e Parmeggiani*, 1905. (Cited on page 58.)

[VL+07] Stavros Vassos, Hector J Levesque, et al. Progression of situation calculus action theories with incomplete information. In *IJCAI*, volume 7, pages 2024–2029, 2007. (Cited on page 11.)

[VL13] Stavros Vassos and Hector J Levesque. How to progress a database iii. *Artificial Intelligence*, 195:203–221, 2013. (Cited on page 11.)

[WP83] Heinz Wimmer and Josef Perner. Beliefs about beliefs: Representation and constraining function of wrong beliefs in

young children's understanding of deception. *Cognition*, 13(1):103–128, 1983. (Cited on page 124.)

[YWL13] Quan Yu, Ximing Wen, and Yongmei Liu. Multi-agent epistemic explanatory diagnosis via reasoning about actions. In *IJCAI*, pages 1183–1190, 2013. (Cited on page 109.)

A

Appendix

A.1. Long Proofs

A.1.1. Proofs about $\mathcal{KL}_n$ translation

While the validity of the translation between the propositional modal logic $\mathcal{ML}_n$ and FOL is proved, the translation between the first-order modal logic $\mathcal{KL}_n$ and FOL is more involved. A proof of the $\mathcal{KL}_n$ translation, originally developed by my collaborator and previous student Hannah Louise Wilk, is given here.

We construct a canonical FOL model M^c as follows:

C1 Domain of discourse $\mathcal{D} = \mathcal{D}_o \cup \mathcal{D}_w$, where the domain of sort object $\mathcal{D}_o$ is isomorphic to the set of standard name $\mathcal{N}$, and the domain of sort state $\mathcal{D}_w$ is isomorphic to the set of $\mathcal{KL}_n$ models $\mathbb{M}$. We use $f_o \colon \mathcal{N} \to \mathcal{D}_o$ and $f_w \colon \mathbb{M} \to \mathcal{D}_w$ to denote the functions of isomorphism and use f_o^{-1} and f_w^{-1} for the inverse functions. In particular, $n^{M^c} = f_o(n)$ for any $n \in \mathcal{N}$.

C2 Equality is interpreted in the usual sense, i.e. $d =^{M^c} d$ and $d \neq^{M^c} d'$ for any distinct elements d, d'.

For j-ary $\mathcal{KL}_n$ predicate P and function g, we consider the interpretation of P and f in FOL as follows:

- $P \subseteq \underbrace{\mathcal{D}_o \times \cdots, \times \mathcal{D}_o}_{j} \times \mathcal{D}_w$ and $g : \underbrace{\mathcal{D}_o \times \cdots, \times \mathcal{D}_o}_{j} \times \mathcal{D}_w \to \mathcal{D}_o$

C3 f.a. $\vec{d} \in (\mathcal{D}_o)^j$ and $\omega \in \mathcal{D}_w$, $(\vec{d}, \omega) \in P^{M^c}$ iff $\vec{n} \in w[P]$ where $(w, \vec{e}\,') = f_w^{-1}(\omega)$ with some $\vec{e}\,'$ and $n_i = f_o^{-1}(d_i)$ for $i \in \{1, \ldots, j\}$

C4 f.a. $\vec{d} \in (\mathcal{D}_o)^j$ and $\omega \in \mathcal{D}_w$, $g^{M^c}(\vec{d}, \omega) = d'$ iff $w[g](\vec{n}) = n'$ where $(w, \vec{e}\,') = f_w^{-1}(\omega)$ with some $\vec{e}\,'$, $n_i = f_o^{-1}(d_i)$ and $n' = f_o^{-1}(d')$

Given $\sigma \in \mathsf{Ag}^+$, the interpretation of predicates E_σ is as follows:[1]

C5 Let $end(\sigma) = i$. For $\omega, \omega' \in \mathcal{D}_w$ s.t. $f_w^{-1}(\omega) = (w, \vec{e})$ and $f_w^{-1}(\omega') = (w', \vec{e}\,')$, $(\omega, \omega') \in E_\sigma^{M^c}$ iff $e_i = e_i'$ and $(w, \vec{e}_{-i}) \in e_i'$.

Where $\vec{e}_{-i}$ is the tuple obtained via removing the i-th argument from $\vec{e}$. Namely, for $\vec{e} = (e_1, \ldots, e_n)$, $\vec{e}_{-i} = (e_1, \ldots, e_{i-1}, e_{i+1}, \ldots, e_n)$.

Theorem A.1. Let M^c be constructed as above, consider models $(w, \vec{e})$ and $(w', \vec{e}\,')$ such that $e_i = e_i'$ and $(w, \vec{e}_{-i}) \in e_i'$, then for any sentence $\alpha \in \mathcal{KL}_n$,

- Given μ s.t. $\mu(u) = f_w(w, \vec{e})$, then $w, \vec{e} \models \alpha$ iff $M^c, \mu \models_{FO} \mathcal{R}[\alpha, \epsilon, u, v]$
- Given μ s.t. $\mu(u) = f_w(w, \vec{e})$, $\mu(v) = f_w(w', \vec{e}\,')$, then for any $\sigma \in \mathsf{Ag}^*$,

$$w, \vec{e} \models \alpha \text{ iff } M^c, \mu \models_{FO} \mathcal{R}[\alpha, \sigma \cdot i, u, v]$$

With this theorem, by applying a step of Skolemization, we prove Theorem 5.29. To prove A.1, we first show that it holds for the equality of any ground terms, as well as for atomic formula of form $P(t_1, \ldots, t_n)$:

Lemma A.2. Given $\mathcal{KL}_n$ model $(w, \vec{e})$, let M^c be constructed as above. Considering σ, variables u, v and variable map μ s.t. $\mu(u) = f_w(w, \vec{e})$,

1. For any ground terms t_1, t_2,

$$w, \vec{e} \models t_1 = t_2 \text{ iff } M^c, \mu \models_{FO} \mathcal{R}[t_1 = t_2, \sigma, u, v] \tag{A.1}$$

[1] In σ, the adjacent letters must be distinct, i.e. $\sigma \notin (\mathsf{Ag}^* \cdot i \cdot i \cdot \mathsf{Ag}^*)$ for any $i \in \mathsf{Ag}$.

2. For any atomic formula $P(t_1, \ldots, t_j)$,

$$w, \vec{e} \models P(t_1, \ldots, t_j) \text{ iff } M^c, \mu \models_{FO} \mathcal{R}[P(t_1, \ldots, t_j), \sigma, u, v] \tag{A.2}$$

Proof. We show the proof of part 1 by induction, where part 2 can be proved similarly.
Base case: Suppose that $t = F(n_1, \ldots, n_k)$ and $t' = F'\left(n'_1, \ldots, n'_{k'}\right)$, we have

$$\begin{aligned}
& w, \vec{e} \models \left(t = t'\right) \\
\Longleftrightarrow & w\left[F(n_1, \ldots, n_k)\right] = w\left[F'\left(n'_1, \ldots, n'_{k'}\right)\right] && (\mathcal{KL}_n \text{ semantics}) \\
\Longleftrightarrow & F^{M^c}\left(f_o(n_1), \ldots, f_o(n_k), \mu(u)\right) = {F'}^{M^c}\left(f_o(n'_1), \ldots, f_o(n'_{k'}), \mu(u)\right) && (\text{C4}) \\
\Longleftrightarrow & M^c, \mu \models F'\left(n_1, \ldots, n_k, u\right) = F'\left(n'_1, \ldots, n'_{k'}, u\right) && (\text{FOL semantics}) \\
\Longleftrightarrow & M^c, \mu \models (t = t') \uparrow_\sigma^u && (\text{Def. } \uparrow_\sigma^u)
\end{aligned}$$

Constants are considered function symbols without arguments. When t or t' is a standard name, the proof is similar but simpler.
Inductive step:

Given $t = F(t_1, \ldots, t_k)$ and $t' = F'(t'_1, \ldots t'_{k'})$, as induction hypotheses:

1 For $i \in \{1, \ldots, k\}$, $w, \vec{e} \models t_i = n_i$ iff $M^c, \mu \models (t_i = n_i) \uparrow_\sigma^u$;

2 For $j \in \{1, \ldots, k'\}$, $w, \vec{e} \models t'_j = n'_j$ iff $M^c, \mu' \models (t'_j = n'_j) \uparrow_\sigma^u$

3 For $\tilde{t} = F(n_1, \ldots, n_k)$ and $\tilde{t'} = F'(n'_1, \ldots, n'_{k'})$,
$w, \vec{e} \models \tilde{t} = \tilde{t'}$ iff $M^c, \mu \models (\tilde{t} = \tilde{t'}) \uparrow_\sigma^u$

Let $\hat{t_i}$ be the term on the LHS in $(t_i = n_i) \uparrow^u_\sigma$ and $\hat{t'_j}$ on the LHS in $(t'_j = n'_j) \uparrow^u_\sigma$,

$$
\begin{aligned}
& w, \vec{e} \models (t = t') \\
\Longleftrightarrow & w[F(t_1, \ldots, t_k)] = w[F'(t'_1, \ldots, t'_{k'})] && (\mathcal{KL}_n \text{ semantics}) \\
\Longleftrightarrow & w[F(n_1, \ldots, n_k)] = w[F(n'_1, \ldots, n'_{k'})] \text{ where} \\
& \quad n_i = w[t_i] \text{ and } n'_j = w[t'_j] \text{ f.a. } i, j && (\text{denotation}) \\
\Longleftrightarrow & w, \vec{e} \models F(n_1, \ldots, n_k) = F(n'_1, \ldots, n'_{k'}) \text{ where} \\
& \quad w, \vec{e}' \models n_i = t_i \text{ and } w, \vec{e}' \models n'_j = t'_j \text{ f.a. } i, j && (\text{semantics}) \\
\Longleftrightarrow & M^c, \mu \models F(n_1, \ldots, n_k, u) = F'(n'_1, \ldots, n'_{k'}, u) \text{ where} \\
& \quad w, \vec{e}' \models n_i = t_i \text{ and } w, \vec{e}' \models n'_j = t'_j \text{ f.a. } i, j && (\text{I.H. } 3) \\
\Longleftrightarrow & M^c, \mu \models F(n_1, \ldots, n_k, u) = F'(n'_1, \ldots, n'_{k'}, u) \text{ where} \\
& \quad M^c, \mu \models (t_i = n_i) \uparrow^u \text{ and } M^c, \mu \models (t'_j = n'_j) \uparrow^u \text{ f.a. } i, j && (\text{I.H. } 1,2) \\
\Longleftrightarrow & M^c, \mu \models F(n_1, \ldots, n_k, u) = F'(n'_1, \ldots, n'_{k'}, u) \text{ where} \\
& \quad M^c, \mu \models \hat{t_i} = n_i \text{ and } M^c, \mu \models \hat{t'_j} = n'_j \text{ f.a. } i, j && (\text{Def. } \hat{t_i}, \hat{t'_j}) \\
\Longleftrightarrow & M^c, \mu \models F(\hat{t_1}, \ldots, \hat{t_k}, u) = F'(\hat{t'_1}, \ldots, \hat{t'_{k'}}, u) && (\text{FOL semantics}) \\
\Longleftrightarrow & M^c, \mu \models (F(t_1, \ldots, t_k) = F'(t'_1, \ldots, t'_{k'})) \uparrow^u && (\text{Def. } \uparrow^u) \\
\Longleftrightarrow & M^c, \mu \models (t = t') \uparrow^u \\
\Longleftrightarrow & M^c, \mu \models \mathcal{R}[t = t', \sigma, u, v]
\end{aligned}
$$

□

Now we complete the proof of Theorem A.1:

Proof. Given $(w, \vec{e})$ and $(w', \vec{e}')$ such that $e_i = e'_i$ and $(w, \vec{e}_{-i}) \in e'_i$, base cases with primitive atoms are proved in the previous lemma.
Induction on $\neg, \wedge, \forall$:

- $\alpha \stackrel{def}{=} \neg\beta$

$$
\begin{aligned}
& w, \vec{e} \models \neg\beta \text{ for } \beta \text{ compatible with } (w, \vec{e}) \\
\Longleftrightarrow & w, \vec{e} \not\models \beta \\
\Longleftrightarrow & M^c, \mu \not\models \mathcal{R}[\beta, \sigma, u, v] && \text{(I.H.)} \\
\Longleftrightarrow & M^c, \mu \models \neg\mathcal{R}[\beta, \sigma, u, v] \\
\Longleftrightarrow & M^c, \mu \models \mathcal{R}[\neg\beta, \sigma, u, v] && \text{(Def. } \mathcal{R}[\cdot]\text{)}
\end{aligned}
$$

- $\alpha \stackrel{def}{=} \beta \wedge \gamma$

$$
\begin{aligned}
& w, \vec{e} \models \beta \wedge \gamma \\
\Longleftrightarrow & w, \vec{e} \models \beta \text{ and } w, \vec{e} \models \gamma && \text{(Semantics } \mathcal{KL}_n\text{)} \\
\Longleftrightarrow & M^c, \mu \models \mathcal{R}[\beta, \sigma, u, v]) \text{ and } M^c, \mu \models \mathcal{R}[\gamma, \sigma, u, v] && \text{(I.H.)} \\
\Longleftrightarrow & M^c, \mu) \models \mathcal{R}[\beta, \sigma, u, v] \wedge \mathcal{R}[\gamma, \sigma, u, v] \\
\Longleftrightarrow & M^c, \mu) \models \mathcal{R}[\beta \wedge \gamma, \sigma, u, u'] && \text{(Def. } \mathcal{R}[\cdot]\text{)}
\end{aligned}
$$

- $\alpha \stackrel{def}{=} \forall x.\beta$, where β has no free variable other than x,

$$
\begin{aligned}
& w, \vec{e} \models \forall x.\beta \\
\Longleftrightarrow & w, \vec{e} \models \beta^x_n \text{ f.a. } n \in \mathcal{N} && \text{(Semantics } \mathcal{KL}_n\text{)} \\
\Longleftrightarrow & M^c, \mu \models \mathcal{R}[\beta^x_n, \sigma, u, v] \text{ f.a. } n \in \mathcal{N} && \text{(I.H.)} \\
\Longleftrightarrow & M^c, \mu \models \mathcal{R}[\beta, \sigma, u, v]^x_n \text{ f.a. } n \in \mathcal{N} \\
\Longleftrightarrow & M^c, \mu' \models \mathcal{R}[\beta, \sigma, u, v] \text{ f.a. } n \in \mathcal{N}, \mu' \sim_x \mu, \mu'(x) = f_o(n) \\
\Longleftrightarrow & M^c, \mu' \models \forall x.\mathcal{R}[\beta, \sigma, u, u'] && \text{(FOL semantics)} \\
\Longleftrightarrow & M^c, \mu' \models \mathcal{R}[\forall x.\beta, \sigma, u, u'] && \text{(Def. } \mathcal{R}[\cdot]\text{)}
\end{aligned}
$$

For induction on $\boldsymbol{K}_i$ operator, as induction hypothesis we assume that for any $(w, \vec{e})$ and $(w', \vec{e}')$ such that $e_i = e'_i$ and $(w', \vec{e}'_{-i}) \in e_i$, for μ' such that $\mu'(u') = f_w(w', \vec{e}')$ and $\mu'(u) = f_w(w, \vec{e})$, let $\sigma \in \mathsf{Ag}^*$ and $end(\sigma) \neq i$.

$$(w', \vec{e}'_{-i}) \cup_i e_i \models \alpha \text{ iff } M^c, \mu' \models \mathcal{R}[\alpha, \sigma \cdot i, u', u] \tag{A.3}$$

Then

$$\begin{aligned}
& w, \vec{e} \models \boldsymbol{K}_i\alpha \\
\Longleftrightarrow & \text{F.a. } (w', \vec{e}'_{-i}) \in e_i, (w', \vec{e}'_{-i}) \cup_i e_i \models \alpha && (\mathcal{KL}_n \text{ semantics}) \\
\Longleftrightarrow & \text{F.a. } (w', \vec{e}'_{-i}) \in e_i \text{ and } \mu' \text{ s.t. } \mu'(u') = f_w((w', \vec{e}'_{-i}) \cup_i e_i), \\
& \qquad \mu'(u) = f_w(w, \vec{e}), M^c, \mu' \models \mathcal{R}[\alpha, \sigma \cdot i, u', u] && (\text{ I.H.}) \\
\Longleftrightarrow & \text{F.a. } \mu' \text{ s.t. } \mu'(u) = f_w(w, \vec{e}) \text{ and } M^c, \mu' \models E_{\sigma \cdot i}(u', u), \\
& \qquad M^c, \mu' \models \mathcal{R}[\alpha, \sigma \cdot i, u', u] && (\text{ cond. C5}) \\
\Longleftrightarrow & M^c, \mu \models \forall u' E_{\sigma \cdot i}(u', u) \supset \mathcal{R}[\alpha, \sigma \cdot i, u', u] && (\text{FOL semantics}) \\
\Longleftrightarrow & M^c, \mu \models \mathcal{R}[\boldsymbol{K}_i\alpha, \sigma, u, v] && (\text{Def. } \mathcal{R}[\cdot])
\end{aligned}$$

Considering the case $\sigma = \epsilon$, then we proved for μ s.t. $\mu(u) = f_w(w, \vec{e})$,

$$w, \vec{e} \models \boldsymbol{K}_i\alpha \text{ iff } M^c, \mu \models \mathcal{R}[\boldsymbol{K}_i\alpha, \epsilon, u, v] \tag{A.4}$$

Which is the first case in Thm A.1. To complete the induction on operator $\boldsymbol{K}_i$ of the second case, it remains to prove that for any $\sigma \in \mathsf{Ag}^*$,

$$w, \vec{e} \models \boldsymbol{K}_i\alpha \text{ iff } M^c, \mu \models_{FO} \mathcal{R}[\boldsymbol{K}_i\alpha, \sigma \cdot i, u, v] \tag{A.5}$$

$$\begin{aligned}
& M^c, \mu \models_{FO} \mathcal{R}[\boldsymbol{K}_i\alpha, \sigma \cdot i, u, v] \\
\Longleftrightarrow & M^c, \mu \models \forall u'.E_{\sigma \cdot i}(u', v) \supset \mathcal{R}[\alpha, \sigma \cdot i, u', v] && (\text{Def. } \mathcal{R}[\cdot]) \\
\Longleftrightarrow & M^c, \mu' \models E_{\sigma \cdot i}(u', v) \supset \mathcal{R}[\alpha, \sigma \cdot i, u', v] \text{ f.a. } \mu' \sim_{u'} \mu && (\text{FOL semantics}) \\
\Longleftrightarrow & M^c, \mu' \models \mathcal{R}[\alpha, \sigma \cdot i, u', v] \text{ f.a. } \mu' \sim_{u'} \mu, \mu'(u') = f_w(w'', \vec{e''}) \text{ s.t.} \\
& \qquad (w'', \vec{e''}_{-i}) \in e_i, e''_i = e_i && (\text{Cond. C5}) \\
\Longleftrightarrow & w'', \vec{e''} \models \alpha \text{ f.a. } (w'', \vec{e''}) \text{ s.t. } (w'', \vec{e''}_{-i}) \in e_i, e''_i = e_i && (\text{I.H. Eq. A.3}) \\
\Longleftrightarrow & w, \vec{e} \models \boldsymbol{K}_i\alpha && (\mathcal{KL}_n \text{ semantics})
\end{aligned}$$

□

In $\mathcal{KL}$ the unique name assumption is made regarding the domain of quantification. Correspondingly, we need the same assumption when translating a $\mathcal{KL}_n$ formula into FO-formula:

- Σ_{eq} is the following infinite set of formulas:
  - $\forall x(x = x)$
  - $\forall x \forall y.(x = y) \supset (y = x)$
  - $\forall x \forall y \forall z((x = y) \wedge (y = z)) \supset (x = z)$
  - for any function symbol F:
    $\forall x_1 \ldots \forall x_k \forall y_1 \ldots \forall y_k((x_1 = y_1) \wedge \ldots \wedge (x_k = y_k)) \supset F(x_1, \ldots, x_k) = F(y_1, \ldots, y_k)$
  - for any predicate symbol P:
    $\forall x_1 \ldots \forall x_k \forall y_1 \ldots \forall y_k((x_1 = y_1) \wedge \ldots \wedge (x_k = y_k)) \supset P(x_1, \ldots, x_k) \equiv P(y_1, \ldots, y_k)$
- $\Sigma_{una} = \{(n \neq n') \mid n, n' \text{ are distinct standard names in } \mathcal{N}\}$

Intuitively, Σ_{eq} means that the symbol $=$, essentially as a binary predicate, is interpreted in its usual sense. Σ_{una} requires that the domain of discourse must be infinite.

In this direction, we aim to prove:

The proof of the theorem consists of the following steps: Given $\alpha \in \mathcal{KL}_n$

1. If $\mathcal{R}[\alpha, \varepsilon, u, v] \cup \Sigma_{eq} \cup \Sigma_{una}$ is FO-satisfiable, then $\mathcal{R}[\alpha, \varepsilon, u, v]$ is satisfiable in $\mathcal{M}^{e,u}$
2. If $\mathcal{R}[\alpha, \varepsilon, u, v]$ is satisfiable in $\mathcal{M}^{e,u}$ then $\mathcal{R}[\alpha, \varepsilon, u, v]$ is satisfiable in $\mathcal{M}^{e,u,o}$
3. If $\mathcal{R}[\alpha, \varepsilon, u, v]$ is satisfiable in $\mathcal{M}^{e,u,o}$ then $M^c, \mu \models \mathcal{R}[\alpha, \varepsilon, u, v]$ for some μ

Here $\mathcal{M}^{e,u}$ denotes the set of all models which satisfy $\Sigma_{eq} \cup \Sigma_{una}$, $\mathcal{M}^{e,u,o}$ denotes the set of all models in $\mathcal{M}^{e,u}$ of form $M = \langle(\mathcal{D}_o^c, \mathcal{D}_w), I\rangle$ and for any $d \in \mathcal{D}_o^c$

it exists n s.t. $n^I = d$. Namely, it shares the same domain of object with the canonical model M^c, and every element of sort object is named. When items 1-3 are proven, then we prove Thm. 5.31.

Among the three steps above, the first one is trivial and the second is proved by applying the Löwenheim-Skolem theorem, renaming elements in $\mathcal{D}_o^M$ and define elements in $\mathcal{D}'_w$ such that they preserve the truth values.

Proof. Supposing that $M, \mu \models \mathcal{R}[\alpha, \epsilon, u, v]$ for some $M \in \mathcal{M}^{e,u}$. Since $\mathcal{M}^{e,u}$ satisfies the unique name assumption Σ_{una}, the domain of discourse is infinite. By the Löwenheim-Skolem theorem, any first-order satisfiable theory with an infinite model also has a countable model. Therefore, we can assume that both $\mathcal{D}_w$ and $\mathcal{D}_o$ are countable. Thus a bijection $\pi : \mathcal{D}_o^M \to \mathcal{N}$ can be found. In addition, for finitely many standard names in α (i.e. in $\mathcal{R}[\alpha, \epsilon, u, v]$), w.l.o.g. we can require that $\pi(n_i^M) = n_i$.
Now we define $M' = \langle \mathcal{D}_o^c, \mathcal{D}'_w, \Phi \rangle$ as follows: For each $\omega \in \mathcal{D}_w$, we introduce a new element ω' in $\mathcal{D}'_w$ s.t. for each symbols P_σ or F_σ occurs in $\mathcal{R}[\alpha, \epsilon, u, v]$,

- $P_\sigma^{M'}(\vec{d'}, \omega') = 1$ iff $P_\sigma^M(\vec{d}, \omega) = 1$ where $d'_i = f_o(\pi(d_i))$
- $F_\sigma^{M'}(\vec{d'}, \omega') = d'_0$ iff $F_\sigma^M(\vec{d}, \omega) = d_0$ where $d'_i = f_o(\pi(d_i))$

Here f_o is the bijection used in defining the canonical model M^c, and $\phi :$ denote the bijection $\mathcal{D}_w \to \mathcal{D}'_w$. For each E_σ occurs in $\mathcal{R}[\alpha, \epsilon, u, v]$, we define the interpretation that f.a. $\omega'_1, \omega'_2 \in \mathcal{D}'_w$, $E_\sigma^{M'}(\omega'_1, \omega'_2) = 1$ iff $E_\sigma^M(\phi^{-1}(\omega'_1), \phi^{-1}(\omega'_2)) = 1$.
Let μ' satisfies $\mu'(u) = \phi(\mu(u))$ and $\mu'(v) = \phi(\mu(v))$, then it is straightforward that $M', \mu' \models \mathcal{R}[\alpha, \epsilon, u, v]$ and $\mathcal{M}' \in \mathcal{M}^o$. □

For the third item, the proof is achieved by constructing a set of mappings between $\mathcal{D}_w^M$ and $\mathcal{D}_w^c$ such that it preserves the truth value for formulas up to a certain depth: Given $M \in \mathcal{M}^{e,u,o}$ and $k \in \mathbb{N}$, for any $\sigma \in \mathsf{Ag}^*$ such that $|\sigma| \leq k$, we inductively construct mapping $\phi^{(\sigma)}$ and set S_σ^ω as follows:

1. When $|\sigma| = k$, f.a. $\omega \in \mathcal{D}_w$, we select a state $\omega^c \in \mathcal{D}_w^c$ such that ,

1.i) $(\vec{d}, \omega) \in P_\sigma^M$ iff $(\vec{d}, \omega^c) \in P_\sigma^{M^c}$ for any P and $\vec{d}$;

1.ii) $F_\sigma^M(\vec{d}, \omega) = F_\sigma^{M^c}(\vec{d}, \omega^c)$ for any F and $\vec{d}$.

We use $\phi^{(\sigma)}$ to denote the mapping from ω to the selected state, i.e. $\phi^{(\sigma)}(\omega) = \omega^c$. Given ω, we define the set $S_\sigma^\omega = \{\omega' \,|\, (\omega', \omega) \in E_\sigma^M\}$. For simplicity, we write $\phi^{(\sigma)}(S_\sigma^\omega)$ to mean $\{\phi^{(\sigma)}(\omega') \,|\, \omega' \in S_\sigma^\omega\}$.

2. When $|\sigma| < k$, f.a. $\omega \in \mathcal{D}_w$, we select a state $\omega^c \in \mathcal{D}_w^c$ such that

2.i) $(\vec{d}, \omega) \in P_\sigma^M$ iff $(\vec{d}, \omega^c) \in P_\sigma^{M^c}$ for any P and $\vec{d}$;

2.ii) $F_\sigma^M(\vec{d}, \omega) = F_\sigma^{M^c}(\vec{d}, \omega^c)$ for any F and $\vec{d}$;

2.iii) F.a. $i \neq end(\sigma)$, $(\omega', \omega^c) \in E_{\sigma\cdot i}^{M^c}$ iff $\omega' \in \phi^{(\sigma\cdot i)}(S_{\sigma\cdot i}^\omega)$

$\phi^{(\sigma)}$ and S_σ^ω are defined identically as the first case.

Note that for the mapping ϕ^σ with $\sigma = \sigma' \cdot j \cdot i$, i.e., j is the second last letter in the word, we always select ω^c such that $(w, \vec{e}) = f_w^{-1}(\omega^c)$, where the depth of e_j is higher by one than any $j' \neq j$, and $(w, \vec{e}_{-j}) \in e_j$.

Apparently, for $|\sigma| < k$, $\phi^{(\sigma)}$ and S_σ^ω are defined based on the definition of $\phi^{(\sigma\cdot j)}$ and $S_{\sigma\cdot j}^\omega$ with $j \neq end(\sigma)$. Since the value assignments in 1.i) - 2.iii) are independent, we can prove that for any model M and finite k, such a set of mappings always exists.

Proof. We prove the statement by induction on the structure of formula α.

- Basis: For objective α, by 1.i), 1.ii), 2.i) and 2.ii) we prove that the statement holds f.a. σ and μ

- Induction:

– Negation: Suppose that the lemma holds for α,

$$\begin{aligned}
& M, \mu \models \mathcal{R}[\neg\alpha, \sigma, u, v] \\
\Longleftrightarrow & M, \mu \models \neg\mathcal{R}[\alpha, \sigma, u, v] && \text{(Def. } \mathcal{R}[\cdot]) \\
\Longleftrightarrow & M, \mu \not\models \mathcal{R}[\alpha, \sigma, u, v] && \text{(FOL semantics)} \\
\Longleftrightarrow & M^c, \tilde{\mu} \not\models \mathcal{R}[\alpha, \sigma, u, v] && \text{(I.H.)} \\
\Longleftrightarrow & M^c, \tilde{\mu} \models \neg\mathcal{R}[\alpha, \sigma, u, v] && \text{(FOL semantics)} \\
\Longleftrightarrow & M^c, \tilde{\mu} \models \mathcal{R}[\neg\alpha, \sigma, u, v] && \text{(Def. } \mathcal{R}[\cdot])
\end{aligned}$$

– Conjunction: Suppose that the lemma holds for α and β,

$$\begin{aligned}
& M, \mu \models \mathcal{R}[\alpha \wedge \beta, \sigma, u, v] \\
\Longleftrightarrow & M, \mu \models \mathcal{R}[\alpha, \sigma, u, v] \wedge \mathcal{R}[\beta, \sigma, u, v] && \text{(Def. } \mathcal{R}[\cdot]) \\
\Longleftrightarrow & M, \mu \models \mathcal{R}[\alpha, \sigma, u, v] \text{ and } M, \mu \models \mathcal{R}[\beta, \sigma, u, v] && \text{(FOL semantics)} \\
\Longleftrightarrow & M^c, \tilde{\mu} \models \mathcal{R}[\alpha, \sigma, u, v] \text{ and } M^c, \tilde{\mu} \models \mathcal{R}[\beta, \sigma, u, v] && \text{(I.H.)} \\
\Longleftrightarrow & M^c, \tilde{\mu} \models \mathcal{R}[\alpha, \sigma, u, v] \wedge \mathcal{R}[\beta, \sigma, u, v] && \text{(FOL semantics)} \\
\Longleftrightarrow & M^c, \tilde{\mu} \models \mathcal{R}[\alpha \wedge \beta, \sigma, u, v] && \text{(Def. } \mathcal{R}[\cdot])
\end{aligned}$$

– Universal quantification of sort object: Suppose that the lemma holds for any α_n^x with $n \in \mathcal{D}_o^c$

$$\begin{aligned}
& M, \mu \models \mathcal{R}[\forall x.\alpha, \sigma, u, v] \\
\Longleftrightarrow & M, \mu \models \forall x.\mathcal{R}[\alpha, \sigma, u, v] && \text{(Def. } \mathcal{R}[\cdot]) \\
\Longleftrightarrow & M, \mu \models \mathcal{R}[\alpha, \sigma, u, v]_n^x \text{ f.a. } n \in \mathcal{N} && (\#) \\
\Longleftrightarrow & M, \mu \models \mathcal{R}[\alpha_n^x, \sigma, u, v] \\
\Longleftrightarrow & M^c, \tilde{\mu} \models \mathcal{R}[\alpha_n^x, \sigma, u, v] \text{ f.a. } n \in \mathcal{N} && \text{(I.H.)} \\
\Longleftrightarrow & M^c, \tilde{\mu} \models \mathcal{R}[\alpha, \sigma, u, v]_n^x \text{ f.a. } n \in \mathcal{N} \\
\Longleftrightarrow & M^c, \tilde{\mu} \models \forall x.\mathcal{R}[\alpha, \sigma, u, v] && (\#) \\
\Longleftrightarrow & M^c, \tilde{\mu} \models \mathcal{R}[\forall x.\alpha, \sigma, u, v]
\end{aligned}$$

To explain the steps labeled with (#): By the definition of $\mathcal{M}^{e,u,o}$, the sub-domain of sort object is countable and each element is named, therefore the quantification can be interpreted substitutionally.

– For sentence $\boldsymbol{K}_i\alpha$ s.t. $i \neq end(\sigma)$

$$
\begin{aligned}
& M, \mu \models \mathcal{R}[\boldsymbol{K}_i\alpha, \sigma, u, v] \\
\Longleftrightarrow & M, \mu \models \forall u'.E_{\sigma\cdot i}(u', u) \supset \mathcal{R}[\alpha, \sigma \cdot i, u', u] && \text{(Def. } \mathcal{R}[\cdot]) \\
\Longleftrightarrow & M, \mu' \models \mathcal{R}[\alpha, \sigma \cdot i, u', u] \text{ f.a. } \mu' \sim_{u'} \mu \text{ s.t. } (\mu'(u'), \mu'(u)) \in E^M_{\sigma\cdot i} \\
& && \text{(FOL semantics)} \\
\Longleftrightarrow & M, \mu' \models \mathcal{R}[\alpha, \sigma \cdot i, u', u] \text{ f.a. } \mu' \sim_{u'} \mu \text{ s.t. } \mu'(u') \in S^{\mu'(u)}_{\sigma\cdot i} \\
& && \text{(Def. } S^{\omega}_{\sigma}) \\
\Longleftrightarrow & M^c, \tilde{\mu}' \models \mathcal{R}[\alpha, \sigma \cdot i, u', u] \text{ f.a. } \tilde{\mu}' \text{ and } \mu' \sim_{u'} \mu \text{ s.t. } \mu'(u') \in S^{\mu'(u)}_{\sigma\cdot i}, \\
& \quad \phi^{(\sigma\cdot i)}(\mu'(u')) = \tilde{\mu}'(u') \text{ and } \phi^{(\sigma)}(\mu'(u)) = \tilde{\mu}'(u) && \text{(I.H.)}
\end{aligned}
$$

Since $\mu' \sim_{u'} \mu$, by $\mu'(u') \in S^{\mu'(u)}_{(\sigma\cdot i)}$ and $\phi^{(\sigma\cdot i)}(\mu'(u')) = \tilde{\mu}'(u')$ we derive $(\tilde{\mu}'(u'), \tilde{\mu}'(u)) \in E^{M^c}_{\sigma\cdot i}$. Since $\mathcal{R}[\alpha, \sigma \cdot i, u', u]$ has no free variables other than u, u', $\tilde{\mu}'(u) = \phi^{(\sigma)}(\mu'(u)) = \phi^{(\sigma)}(\mu(u)) = \tilde{\mu}(u)$, w.l.o.g. $\tilde{\mu}' \sim_{u'} \tilde{\mu}$.

$$
\begin{aligned}
\Longleftrightarrow & M^c, \tilde{\mu}' \models \mathcal{R}[\alpha, \sigma \cdot i, u', u] \text{ f.a. } \tilde{\mu}' \sim_{u'} \tilde{\mu} \text{ s.t. } (\tilde{\mu}'(u'), \tilde{\mu}'(u)) \in E^{M^c}_{\sigma\cdot i} \\
\Longleftrightarrow & M^c, \tilde{\mu} \models \mathcal{R}[\boldsymbol{K}_i\alpha, \sigma, u, v] && \text{(Def. } \mathcal{R}[\cdot])
\end{aligned}
$$

– For sentence $\boldsymbol{K}_i\alpha$ s.t. $i = end(\sigma)$

$$
\begin{aligned}
& M, \mu \models \mathcal{R}[\boldsymbol{K}_i\alpha, \sigma, u, v] \\
\Longleftrightarrow & M, \mu \models \forall u'.E_\sigma(u', v) \supset \mathcal{R}[\alpha, \sigma, u', v] && \text{(Def. } \mathcal{R}[\cdot]) \\
\Longleftrightarrow & M, \mu' \models \mathcal{R}[\alpha, \sigma, u', v] \text{ f.a. } \mu' \sim_{u'} \mu \text{ s.t. } (\mu'(u'), \mu'(v)) \in E^M_\sigma \\
& && \text{(FOL semantics)} \\
\Longleftrightarrow & M, \mu' \models \mathcal{R}[\alpha, \sigma, u', v] \text{ f.a. } \mu' \sim_{u'} \mu \text{ s.t. } \mu'(u') \in S^{\mu'(v)}_\sigma && \text{(Def. } S^{\omega}_\sigma) \\
\Longleftrightarrow & M^c, \tilde{\mu}' \models \mathcal{R}[\alpha, \sigma, u', v] \text{ f.a. } \tilde{\mu}' \text{ and } \mu' \sim_{u'} \mu \text{ s.t. } \mu'(u') \in S^{\mu'(v)}_\sigma, \\
& \phi^{(\sigma)}(\mu'(u')) = \tilde{\mu}'(u') \text{ and } \phi^{(\sigma^-)}(\mu'(v)) = \tilde{\mu}'(v) && \text{(I.H.)}
\end{aligned}
$$

Since $\mu' \sim_{u'} \mu$, by $\mu'(u') \in S^{\mu'(v)}_{\sigma^-}$ and $\phi^{(\sigma)}(\mu'(u')) = \tilde{\mu}'(u')$ we derive $(\tilde{\mu}'(u'), \tilde{\mu}'(v)) \in E^{M^c}_\sigma$. Since $\mathcal{R}[\alpha, \sigma, u', v]$ has no free variables other than v, u' and $\tilde{\mu}'(v) = \phi^{(\sigma^-)}(\mu'(v)) = \phi^{(\sigma^-)}(\mu(v)) = \tilde{\mu}(v)$, w.l.o.g. $\tilde{\mu}' \sim_{u'} \tilde{\mu}$.

$$
\begin{aligned}
\Longleftrightarrow & M^c, \tilde{\mu}' \models \mathcal{R}[\alpha, \sigma, u', v] \text{ f.a. } \tilde{\mu}' \sim_{u'} \tilde{\mu} \text{ s.t. } (\tilde{\mu}'(u'), \tilde{\mu}'(v)) \in E^{M^c}_\sigma \\
\Longleftrightarrow & M^c, \tilde{\mu} \models \mathcal{R}[\boldsymbol{K}_i\alpha, \sigma, u, v] && \text{(Def. } \mathcal{R}[\cdot])
\end{aligned}
$$

□

A.1.2. Proofs about Regression

We first recap Theorem 5.50 and Lemma 5.51:

- **Thm. 5.50**: For any theory Γ, φ regressable wrt Γ, for all z and $i \in \mathsf{Ag}$,

$$dep[\mathcal{R}[\Gamma, z, \varphi], i] \leq \mathsf{Ord}(\Gamma, i)$$

- **Lem. 5.51**: Given action theory Γ, for all Γ-model $(w, \vec{e})$, $z \in Z$ and static formula α regressable wrt Γ,

$$(w, \vec{e}) \models [z]\alpha \text{ iff } (w, \vec{e}) \models \mathcal{R}[\Gamma, z, \alpha]$$

We proceed a triple-induction, respectively for the theorem and the lemma: The outer-induction is on the order of the theory Γ, the mid-induction is on the length of action sequence z and the inner-induction is on the construction of formula α. The following proposition is also proved via induction. It will be used in proving the aforementioned theorem and lemma:

Proposition A.3. For any Γ, $z = a_1 \cdots a_m$, and φ regressable wrt. Γ,

$$\mathcal{R}[\Gamma, z, \boldsymbol{K}_i\varphi] = \boldsymbol{K}_i[\bigwedge_{a'_1 \ldots a'_m} (\bigwedge_{j=1}^{m} \mathcal{R}[\Gamma^i, a'_1 \cdots a'_{j-1}, Poss(a'_j) \wedge I(a_j, a'_j)]) \rightarrow \mathcal{R}[\Gamma^i, a'_1 \cdots a'_m, \varphi]] \tag{A.6}$$

Proof of Theorem 5.50: **Base Case 1**: For 0-order action theory Γ, namely, Γ is objective. Then a formula regressable wrt. Γ must be objective as well. Thus the regression will always result in an objective formula, i.e. $dep[\mathcal{R}[\Gamma, z, \varphi], i] = 0$.
I.H. 1: Suppose that Eq. (5.12) holds for all Γ' such that $\mathsf{Ord}(\Gamma') < k$, all $z \in Z$ and regressable static α. Now we consider theories of order k:
Base Case 2: For $\mathsf{Ord}(\Gamma) = k$ and $z = \langle\rangle$, we consider all cases as follows:

- $\mathcal{R}[\Gamma, \langle\rangle, p] := p$ has depth 0.
- $\mathcal{R}[\Gamma, \langle\rangle, \alpha \wedge \beta] := \mathcal{R}[\Gamma, \langle\rangle, \alpha] \wedge \mathcal{R}[\Gamma, \langle\rangle, \beta]$ does not increase the depth.
- $\mathcal{R}[\Gamma, \langle\rangle, Poss(a)] := \pi_a^\Gamma$. By the definition of Ord, $dep[\pi_a^\Gamma, i] \leq \mathsf{Ord}(\Gamma, i)$.
- $\mathcal{R}[\Gamma, \langle\rangle, I(a, a')] := \phi_{a,a'}^\Gamma$. By the definition, $dep[\phi_{a,a'}^\Gamma, i] \leq \mathsf{Ord}(\Gamma, i)$.
- $\mathcal{R}[\Gamma, \langle\rangle, \boldsymbol{K}_i\alpha] := \boldsymbol{K}_i(\mathcal{R}[\Gamma^i, \langle\rangle, \alpha])$. When $i \neq j$,

$$\begin{aligned} dep[(\mathcal{R}[\Gamma^i, \langle\rangle, \alpha]), j] &\leq \mathsf{Ord}(\Gamma^i, j) && \text{(I.H.1)} \\ &< \mathsf{Ord}(\Gamma, i) && \text{(Lem. 5.45)} \end{aligned}$$

  and when $i = j$,

$$\begin{aligned} dep[(\mathcal{R}[\Gamma^i, \langle\rangle, \alpha]), j] &\leq \mathsf{Ord}(\Gamma^i, j) && \text{(I.H.1)} \\ &= \mathsf{Ord}(\Gamma, i) && \text{(Lem. 5.45)} \end{aligned}$$

Hence

$$dep[\boldsymbol{K}_i(\mathcal{R}[\Gamma^i, \langle\rangle, \alpha]), i] \leq \max\{\mathsf{Ord}(\Gamma, i), \max_{j \neq i}\{\mathsf{Ord}(\Gamma^j, i) + 1\}\} \leq \mathsf{Ord}(\Gamma, i).$$

Cases for other static formulae can be proved by formula construction.

I.H. 2: Suppose that for any Γ' s.t. $\mathsf{Ord}(\Gamma') = k$ and z' s.t. $|z'| < m$, Eq. (5.12) holds for all regressable α. Now we consider z of length m. W.l.o.g. let $z = \sigma \cdot a$,

Base Case 3: For $p \in \mathsf{Ap}$, $\mathcal{R}[\Gamma, z, p] := \mathcal{R}[\Gamma, \sigma, \gamma^{\Gamma}_{p,a}]$, where $\gamma^{\Gamma}_{p,a}$ is regressable wrt. Γ and therefore has be proved in I.H.2.

Induction on operators $\wedge$ and $\neg$ is trivial. For induction on modality $\boldsymbol{K}_i$:

$$\begin{aligned}
&\mathcal{R}[\Gamma, z, \boldsymbol{K}_i\varphi] \\
=&\boldsymbol{K}_i\{\bigwedge_{a'_1 \ldots a'_m}(\bigwedge_{j=1}^{m} \mathcal{R}[\Gamma^i, a'_1 \cdots a'_{j-1}, Poss(a'_j) \wedge I(a_j, a'_j)]) \rightarrow \mathcal{R}[\Gamma^i, a'_1 \cdots a'_m, \varphi]\}
\end{aligned}$$

(Prop. A.3)

Conjunctions and implication will not increase the depth. Similar as in Base case 2, we prove that

- $dep[\mathcal{R}[\Gamma^i, a'_1 \cdots a'_{j-1}, Poss(a'_j) \wedge I(a_j, a'_j)], i] \leq \mathsf{Ord}(\Gamma, i)$;
- $dep[\mathcal{R}[\Gamma^i, a'_1 \cdots a'_m, \varphi], i] \leq \mathsf{Ord}(\Gamma, i)$;
- $dep[\mathcal{R}[\Gamma^i, a'_1 \cdots a'_{j-1}, Poss(a'_j) \wedge I(a_j, a'_j)], i'] < \mathsf{Ord}(\Gamma, i)$ for $i' \neq i$; and
- $dep[\mathcal{R}[\Gamma^i, a'_1 \cdots a'_m, \varphi], i'] < \mathsf{Ord}(\Gamma, i)$ for $i' \neq i$.

By the definition of *dep* we prove that Eq. (5.12) holds.

Proof of Lemma 5.51:

Base Case 1: For 0-order action theory Γ, namely, Γ is objective. Then a formula regressable wrt. Γ must be objective as well. We refer to the work from Lakemeyer and Levesque [LL11] for regression of objective formulae.

I.H. 1: Suppose that Eq. (5.13) holds for all Γ' such that $\mathsf{Ord}(\Gamma') < k$, all $z \in Z$ and regressable static α. Now we consider theories of order k:

Base Case 2: For $\mathsf{Ord}(\Gamma) = k$ and $z = \langle\rangle$, the proof is trivial as $\mathcal{R}[\Gamma, \langle\rangle, \alpha] = \alpha$ for any static α which does not mention $Poss$ and I. Formulae with $Poss$ (or I) can be reduced to the former cases with rule 3(4) in Def. 5.49.

I.H. 2: Suppose that for any Γ' s.t. $\mathsf{Ord}(\Gamma') = k$ and z' s.t. $|z'| < m$, Eq. 5.13 holds for all regressable α. Now we consider z of length m. W.l.o.g. let $z = \sigma \cdot a$,

Base Case 3: For $p \in \mathsf{Ap}$ and Γ-model $(w, \vec{e})$,

$$
\begin{aligned}
(w,\vec{e}) \models [z]p \Longleftrightarrow & (w,\vec{e}) \models [\sigma]([a]p) & \\
\Longleftrightarrow & (w,\vec{e}) \models [\sigma]\gamma^{\Gamma}_{p,a} & ((w,\vec{e}) \models \Gamma_{ssa}) \\
\Longleftrightarrow & (w,\vec{e}) \models \mathcal{R}[\Gamma, \sigma, \gamma^{\Gamma}_{p,a}] & \text{(I.H. 2)} \\
\Longleftrightarrow & (w,\vec{e}) \models \mathcal{R}[\Gamma, \sigma \cdot a, p] & \text{(Def. } \mathcal{R}[\cdot]) \\
\Longleftrightarrow & (w,\vec{e}) \models \mathcal{R}[\Gamma, z, p] &
\end{aligned}
$$

Induction on operators $\wedge$ and $\neg$ is trivial. For induction on modality $\boldsymbol{K}_i$:

I.H. 3: Suppose that Eq. 5.13 holds for all k-order Γ', z of length m and a fixed,

regressable formula φ, let $z = a_1 \cdots a_m$,

$$(w, \vec{e}) \models \mathcal{R}[\Gamma, z, \boldsymbol{K}_i\varphi]$$

$$\Longleftrightarrow (w, \vec{e}) \models \boldsymbol{K}_i\{\bigwedge_{a'_1 \ldots a'_m} (\bigwedge_{j=1}^{m} \mathcal{R}[\Gamma^i, a'_1 \cdots a'_{j-1}, Poss(a'_j) \wedge I(a_j, a'_j)]) \rightarrow \mathcal{R}[\Gamma^i, a'_1 \cdots a'_m, \varphi]\}$$

(Prop. A.3)

$\Longleftrightarrow$For any $s \in e_i$,

$$s \cup_i e_i \models \bigwedge_{a'_1 \ldots a'_m} (\bigwedge_{j=1}^{m} \mathcal{R}[\Gamma^i, a'_1 \cdots a'_{j-1}, Poss(a'_j) \wedge I(a_j, a'_j)]) \rightarrow \mathcal{R}[\Gamma^i, a'_1 \cdots a'_m, \varphi]$$

(Semantics)

$\Longleftrightarrow$For any $s \in e_i$ and $a'_1 \ldots a'_m$, if $s \cup_i e_i \models \mathcal{R}[\Gamma^i, a'_1 \cdots a'_{j-1}, Poss(a'_j) \wedge I(a_j, a'_j)]$

for all $j \leq m$, then $s \cup_i e_i \models \mathcal{R}[\Gamma^i, a'_1 \cdots a'_m, \varphi]$ (Semantics)

$\Longleftrightarrow$For any $s \in e_i$ and $a'_1 \ldots a'_m$, if $s \cup_i e_i \models [a'_1 \cdots a'_{j-1}]Poss(a'_j) \wedge I(a_j, a'_j)$

for all $j \leq m$, then $s \cup_i e_i \models \mathcal{R}[\Gamma^i, a'_1 \cdots a'_m, \varphi]$ (I.H. 2)

$\Longleftrightarrow$For any $s \in e_i, z' \in Z$, if $z' \in \Pi_{w_s}, z \sim_{w_s} z'$, then $s \cup_i e_i \models \mathcal{R}[\Gamma^i, z', \varphi]$

(Semantics of $Poss, I$)

$\Longleftrightarrow$For any $s \in e_i, z' \in Z$, if $z' \in \Pi_{w_s}, z \sim_{w_s} z'$, then $s \cup_i e_i \models [z']\varphi$ (#)

$\Longleftrightarrow$For any $s' \in e_{i,z}, s' \cup_i e_{i,z} \models \varphi$ (Def. $e_{i,z}$)

$\Longleftrightarrow e_{i,z} \models \boldsymbol{K}_i\varphi$ (Semantics)

$\Longleftrightarrow e_i \models [z]\boldsymbol{K}_i\varphi$ (Semantics)

where w_s stands for the world in structure s. It could be two cases in (#): When $\mathsf{Ord}(\Gamma^i) < k$, then I.H 1 is applied as $(s \cup_i e_i)$ is a Γ^i-model. Otherwise $\mathsf{Ord}(\Gamma^i) = k$, then I.H.3 is applied.