Towards Theory-of-Mind agents using Automated Planning and Dynamic Epistemic Logic

Andersen, Mikkel Birkegaard; Bolander, Thomas

Publication date:
2015

Document Version
Publisher's PDF, also known as Version of record

Towards Theory-of-Mind agents using Automated Planning and Dynamic Epistemic Logic

Mikkel Birkegaard Andersen

Kongens Lyngby 2014
PhD-2014-351

Technical University of Denmark
Department of Applied Mathematics and Computer Science
Matematiktorvet, Building 303 B, DK-2800 Kongens Lyngby, Denmark
Phone +45 45253031
compute@compute.dtu.dk
www.compute.dtu.dk PhD-2014-351

Abstract

This thesis is part of a growing body of work in what we call epistemic planning. Epistemic planning is situated at the intersection of automated planning and what can broadly be called dynamic logics. Both are part of the much larger field of Artificial Intelligence.

Automated Planning has been around since at least the 1970s. It is a diverse collection of methods, models, algorithms and specification languages for giving autonomous agents the ability to come up with plans for proactively achieving goals. Autonomous agents can be understood as independent actors, given a purpose by their designer. Whether they are in a software system, connected to the real world with sensors and actuators, or used as a tool for modelling people, for instance in economics, they need to be able to imagine (or predict) outcomes of actions in order to form plans.

The feature that most distinguishes planning from other decision making methods, is that the planner does not know the full system from the beginning. Most of the time it would simply be too big to store in memory! Instead of being given the entire "game", they use a specification of actions and the initial state to generate only a fraction of the full search space. This means that what an agent can plan for depends crucially on what domains we can describe. This is where logic comes into the picture.

For most of its more than 2500 year long history, logic has been mostly interested in the study of valid reasoning. In later years (in the scheme of things), more attention has been given to studying when reasoning fails in humans. Like using calculus to analyse and simulate both when, for instance, a bridge holds and when it collapses, we can use logic to analyse and simulate reasoning both when it is sound and when it isn't.

The subbranch of logic applied in this work is Dynamic Epistemic Logic. The epistemic part concerns the formalisation of knowledge and belief (mainly) in multi-agent settings. We can describe situations in which many agents are present and have different knowledge and beliefs about the world and each others' knowledge and belief. Adding the dynamic part of Dynamic Epistemic Logic to our arsenal, we can describe how situations change when, broadly speaking, things happen. In the application to Automated Planning, we let these things be actions of the agents in the system. In doing so we derive new planning formalisms that allow agents to plan under consideration of how what they do changes both the world and knowledge and belief about the world.

In this thesis we give new planning formalisms for single-agent planning and new results for the model theory of multi-agent models. The first of the two fully developed planning formalisms is conditional (single-agent) epistemic planning, allowing an agent to plan with what it knows now and what it knows it will come to know. Though this is nothing new in Automated Planning, it sets the stage for later work.

The second planning formalism extends conditional epistemic planning with beliefs, letting the agent have expectations, without probabilities, of how things will turn out. Our radically different notions of bisimulation for the multi-agent versions of these models are particularly interesting for logicians, as are surprising expressivity results for well known logics on such models.

The final part of the thesis describes ideas on extending the second formalism to a multi-agent setting. With a view towards the practical implementation of agents, we shall also see how an agent can discard the parts of its model that it does not believe to be the case. While this is not necessary for analysing reasoning agents, it does seem a requirement for practical implementations. There are simply too many possibilities for a resource-bounded agent to keep track of. If the agent does discard unlikely possibilities, it must be able to do belief revision if it later turns out to be wrong. Such a procedure is also described.

The long term potential of multi-agent aware planning algorithms is that agents that can predict and understand others in order to plan cooperation, communication, and/or competition. It is the slow edging towards a general framework for multi-agent planning that is the underlying motivation, and some of the main results, of this thesis. While regrettably we haven't gotten there yet, we're considerably closer than when we started.

Resumé

Denne afhandling placerer sig i et voksende felt som vi kalder epistemisk planlægning. Epistemisk planlægning befinder sig i et krydsfeltet mellem automatisk planlægning og det, der bredt kan kaldes dynamisk logik. Begge er indehold i det større emnefelt Kunstig Intelligens.

Automatisk planlægning har eksisteret siden 70erne, og er mangfoldig samling metoder, modeller, algoritmer og specifikationssprog beregnet til, at give autonome agenter evnen til at lave planer for proaktivt at opnå deres mål. Autonome agenter can forstås som uafhængige aktører, som bliver givet et formål eller en opgave af den der designer dem. Om de findes i et stykke software, er forbundet til den virkelige verden med sensorer og motorer, eller bruges som værktøj til at modellere mennesker, f.eks. i økonomi, så har de brug for at kunne forestille og forudse udfaldet af deres handlinger så de kan lave planer.

Nok den egenskab der mest adskiller planlægning fra andre metoder til beslutningstagen er, at planlægningsalgoritmen ikke kender det fulde system fra starten. Oftest vil det være alt for stort overhovedet at lagre i hukommelsen! I stedet for at være givet det fulde "spil", så bruger de en specifikation af handlinger og starttilstand til kun at generere en brøkdel af det fulde søgerum. Derfor afgør specifikationssproget i allerhøjeste grad hvad en agent kan planlægge for at opnå. Det er her logik kommer ind i billedet.

I det meste af dets mere end 2500 år lange historie har logikken mest drejet sig om at studere valid ræsonering. I de senere år (I lyset af den lange historie) har studiet af fejlslagen ræsonering fået mere opmærksomhed. På samme måde som vi kan bruge differentialligninger til at analysere og simulere at en bro holder eller styrter sammen, så kan logik bruges til at analysere og simulere både sund og usund ræsonering.

Den undergren af logikken som bliver anvendt i det nærværende arbejde er Dynamisk Epistemisk Logik. Den epistemiske del omhandler formalisering af tro og

viden, hovedsagligt i multi-agent systemer. Vi kan beskrive situationer hvori mange agenter er tilstede og har forskellig viden og tro om hinandens viden og tro. Når den dynamiske det af Dynamisk Epistemisk Logik føjes til vores værktøjskasse, så kan vi beskrive hvordan situationer ændrer sig når, groft sagt, ting sker. I kontekst af automatisk planlægning, så lader vi disse ting være agenternes handlinger. Ved at kombinere Dynamisk Epistemisk Logik og Automatisk planlægning, opnår vi nye planlægningsformalismer der gør det muligt for agenter at planlægge med hensyn til hvordan deres handlinger ændrer både verden og tro og viden om verden.

I denne afhandling præsenteres nye planlægningsformalismer for enkelt-agent planlægning og nye resultater om multi-agent modellers egenskaber. Den første fuldt udviklede formalisme er betinget (enkelt-agent) epistemisk planlægning, hvormed en agent kan planlægge både med hvad den ved, men også hvad den ved at den vil komme til at vide. Selvom dette er set før i automatisk planlægning, så gøder dette jorden til det efterfølgende arbejde.

Den anden formalisme udvider betinget epistemisk planlægning med tro, der lader agenten have forventninger (uden brug af sandsynligheder), om hvordan tingene vil forløbe. Vores radikalt anderledes bisimulationsbegreb for multi-agent udgaven af disse modeller er særligt interessant for logikere, ligesom vores overraskende resultater om velkendte logikkers udtrykskraft på sådanne modeller er det.

I den sidste del af afhandlingen beskrives ideer om udvidelse af den anden formalisme til en multi-agent version. Med henblik på praktisk implementation af agenter, så vises også hvordan en agent can skære dele af sin model væk, hvis den ikke tror at de er tilfældet. Selvom dette ikke er nødvendigt for at analysere ræsonerende agenter, så virker det som et krav til praktisk implementation. Der er ganske enkelt for mange muligheder til at en agent med begrænsede ressourcer kan holde styr på dem alle. Når en agent kaster muligheder bort, så må den også kunne revidere sin model hvis det senere viser sig, at den tog fejl. Sådan en procedure beskrives også.

På den lange bane er potentialet ved planlægningsalgoritmer der tager hensyn til mange agenter, at agenter kan forudse og forstå andre, og dermed tage samarbejde, kommunikation og/eller konkurrence med i deres planlægning. Det er det lange, seje træk frem mod sand multi-agent planlægning der er den grundlæggende motivation, og nogle af hovedresultaterne af denne afhandling. Selvom vi ikke har løst problemet endnu, så er vi meget tættere på end da vi startede.

Preface

This thesis is the end product of my PhD studies in the Algorithms, Logic and Graphs section at DTU Compute, Technical University of Denmark, from February 2011 to October 2014. My supervisor during the PhD was Associate Professor Thomas Bolander. From September 2011 to December 2011 I visited the Institute for Logic, Language and Computation at the University of Amsterdam, under the supervision of Alexandru Baltag and Johan van Benthem.

The thesis consists of three joint publications and an extended chapter on future work. It is part of the formal requirements to obtain the PhD degree at the Technical University of Denmark.

Lyngby, 15-October-2014

Mikkel Birkegaard Andersen

Acknowledgements

The biggest thanks of all must go to my supervisor Thomas Bolander. Thomas has been the singlemost important influence on my development as an engineer and an academic (if that is indeed what I have become). From my first experience of him as a supervisor in a specialisation course in January 2007, and later in my Bachelor's thesis, my Master's thesis and now my PhD, Thomas has been a dedicated and knowledgeable supervisor and one of the most pleasant people I've had the good fortune to meet.

Thank you to Martin Holm Jensen, with whom I've collaborated closely throughout my PhD. He has been an excellent partner in crime. Without Martin and Thomas, my PhD work would have been half as good and less than half as fun. Thank you for the work, the travels, the drum solos, and the beers.

Thank you also to Hans van Ditmarsch for a fruitful collaboration and for helping me get to New Zealand in the middle of the cold and dark Danish winter.

Thanks to colleagues at DTU and elsewhere, who have provided invaluable guidance and feedback. If I leave someone out it is my fault, not yours: Jens Ulrik Hansen, Rasmus Rendsvig, Valentin Goranko, Johan van Benthem, Alexandru Baltag, Steen Vester, and the entire Algolog crew.

Finally, thank you to family and friends old and new. I've had wonderful discussions with many of you, on topics both closely and not very closely related to the ideas and interests put forth in this work.

And to Fredrik, Thor and Peter: Soon I will be a PhD too.

Contents

Abstract i

Resumé iii

Preface v

Acknowledgements vii

1 Introduction 1

1.1 Agents . . . 3

1.2 Automated Planning . . . 5

1.2.1 Classical Planning . . . 6

1.2.2 Nondeterminism & Partial Observability . . . 10

1.3 Dynamic Epistemic Logic in Automated Planning . . . 12

1.3.1 States and Actions in DEL . . . 12

1.4 Social Reasoning . . . 15

1.4.1 ToM – What is it good for? . . . 17

1.5 Other Related Work . . . 19

1.5.1 Implementations of Social Reasoning . . . 21

1.6 Outline of Thesis . . . 23

2 Conditional Epistemic Planning 25

Abstract . . . 27

2.1 Introduction . . . 27

2.2 Dynamic Epistemic Logic . . . 29

2.3 Conditional Plans in DEL . . . 30

2.3.1 States and Actions: The Internal Perspective . . . 31

2.3.2 Applicability, Plans and Solutions . . . 32

2.4 Plan Synthesis . . . 35

2.4.1 Planning Trees . . . 36

2.4.2 Strong Planning Algorithm . . . 42

2.4.3 Weak Planning Algorithm . . . 43

2.5 Related and Future Work . . . 44

3 Don't Plan for the Unexpected: Planning Based on Plausibility Models 47
Abstract 49
3.1 Introduction 50
3.2 Dynamic Logic of Doxastic Ontic Actions 51
3.3 Plausibility Planning 56
3.3.1 The Internal Perspective On States 57
3.3.2 Reasoning About Actions 57
3.3.3 Planning 61
3.4 Plan Synthesis 65
3.4.1 Bisimulations, contractions and modal equivalence 65
3.4.2 Planning Trees 69
3.4.3 Planning Algorithm 77
3.5 Related and Future Work 78

4 Bisimulation and expressivity for conditional belief, degrees of belief, and safe belief 81
Abstract 83
4.1 Introduction 84
4.2 Plausibility models and bisimulation 87
4.3 Logical language and semantics 101
4.4 Bisimulation characterisation for L^C, L^D and L^S 105
4.4.1 Bisimulation correspondence for conditional belief 105
4.4.2 Bisimulation correspondence for degrees of belief 110
4.4.3 Bisimulation correspondence for safe belief 118
4.5 Expressivity 120
4.5.1 Expressivity of Safe Belief 121
4.5.2 Expressivity of Degrees of Belief 126
4.5.3 Mapping Out the Relative Expressive Power 129
4.5.4 Reflection on bisimulation characterization and expressivity . 130
4.6 Comparison and applications 133

5 Multi-agent plausibility planning 137
5.1 Planning so far 137
5.2 Models & Language 139
5.2.1 Buying Records – A Running Single-Agent Example 148
5.3 Plausibility Planning with Multi-agent Models 151
5.3.1 Restrictions & Revisions 155
5.3.2 Synthesising Default Plans 161
5.4 Discussion 163

6 Conclusion and Future Work 167

Bibliography 169

Chapter 1

Introduction

Along with “Considering how likely we all are to be blown to pieces by it within the next five years, the atomic bomb has not roused so much discussion as might have been expected.” and “We were somewhere around Barstow on the edge of the desert when the drugs began to take hold.”, one of my favourite essay openers is “I propose to consider the question, ‘Can machines think?’”[1] It comes from Alan Turing’s seminal essay *Computing Machinery and Intelligence* in which he both poses and dismisses the question within the very first paragraph, later writing that the question is “too meaningless to deserve discussion.” His reasons were that qualities like “thinking” and “understanding” are much too vague and emotionally charged to be useful.

Instead he suggests setting up a game in which two players interact with a judge via text. One player is a human, one is a machine and the judge must guess which is which. *The Imitation Game*, as he dubbed this test, is won by the machine if the judge cannot guess correctly more than chance permits. Though one should be careful about quarrelling with Turing, I still believe that the question “Can machines think?” is a potent and meaningful one, well worthy of discussion.

With the question about thinking machines being both plainly understandable and mildly provocative, I find it very useful. Experience tells me that the interesting and animated discussion does not follow from saying “I do research in artificial intelligence” when someone asks me what I do for a living. But tell them “I make thinking machines” and look at them go!

[1] Considering the recent hype regarding Google DeepMind and Vicarious I might also add “One of the most salient features of our culture is that there is so much bullshit.” to the list.

Would I then rather that we call the discipline “Thinking Machine”-ology? No. “Thinking machines” seems to suggest the actual construction of machinery. As argued in [Levesque, 2014], Artificial Intelligence (AI) can be approached as both science and as engineering:

> [T]he science of AI studies intelligent behaviour, not *who* or *what* is producing the behaviour. It studies natural language understanding, for instance, not natural language understanders. This is what makes AI quite different from the study of people (in neuroscience, psychology, cognitive science, evolutionary biology, and so on).

This is contrasted with AI as engineering, which asks how we can construct computer systems (i.e. machines) that exhibit intelligent behaviour. Ideally, AI is both. He goes on to say that we need not nail down any particular definition of intelligent behaviour. The way to progress is for different researchers to focus on different types of behaviour.

In the same way we should refrain from trying to understand everything at once, Marvin Minsky cautions that we should not try to build one system that does everything at once [Bicks, 2010]:

> The basic idea I promote is that you mustn’t look for a magic bullet. You mustn’t look for one wonderful way to solve all problems. Instead you want to look for 20 or 30 ways to solve different kinds of problems. And to build some kind of higher administrative device that figures out what kind of problem you have and what method to use.

Rather than attempting the possibly impossible task of building a single tool that handles all problems, we build a Swiss Army knife of many different tools, each specialised for a specific class of problems.

The question my co-authors and I are interested in is how to understand and build agents that plan and act in an uncertain and social world. It is this question that frames my thesis. In the work contained in these pages, we edge our way towards *an* answer to this problem.

The remainder of this chapter presents the concepts underlying (explicitly or implicitly) the ideas in this thesis. In Section 1.1 I give a definition of a simplified type of autonomous agent as a way to understand the context our methods are to be used in. Section 1.2 provides an introduction to automated planning, the basic method an agent can use to figure out how to act. Section 1.3 gives a brief account of dynamic epistemic logic (DEL). As DEL is treated in detail in later chapters, I here focus on providing a brief historical overview and context for what is to come. Section 1.4 discusses the concept of a Theory of Mind, a crucial component of social reasoning, and gives some examples motivating why we want to build agents that have one. Section 1.5 discusses and relates epistemic planning to other frameworks for reasoning about actions in uncertain and/or multi-agent settings. It also touches upon some implementations of such systems. Finally, Section 1.6 gives an overview of the ideas and results of chapters 2 through 5.

1.1 Agents

An agent is an entity situated in an environment over which it has limited control. It can affect the environment by doing actions. Sensors provides partial information about the environment. Apart from us being able to model them, this concept of agent does not make any special demands of the kind of actions that can be done. The agent can be a car, with actions such as turning the engine on, braking, and steering; its sensors being sat navs and fuel gauges. Or maybe the agent is a botanist on a field expedition, an intelligent house lived in by a family of five, a teenager's smartphone, a network of military computers or something else entirely. What is important is the basic structure of its interaction with its surroundings. This basic structure is shown in Algorithm 1. It is a perpetual cycle in which: 1) based on her internal model, the agent choses an action to do next, 2) it does the action, changing the external world, 3) the world is (perhaps partially) observed and 4) the agent's model updated.

Algorithm 1 AgentLoop

```
$M \leftarrow M_0$
while true do
    $A \leftarrow nextAction(M)$
    $do(A)$
    $\rho \leftarrow perceiveEnvironment()$
    $M \leftarrow updateModel(M, A, \rho)$
end while
```

This algorithm is a stripped down version of an agent loop for a BDI—Beliefs, Desires, Intentions—agent [Wooldridge, 2000]. It effects the four informally outlined steps. M_0 is the agent's initial view of the world, in this formulation including information about what the agent is trying to achieve. The current internal model (or mental model) is M. This is the *beliefs* part. Though usually logic based, Bayesian approaches have been investigated in the recent past [Fagundes et al., 2009].

Based on the agent's current model, *nextAction* invokes a decision procedure that determines what to do next. Keeping in line with Minsky's warning against one-size-fits-all thinking, we can imagine that *nextAction* can invoke a number of different procedures depending on what the agent is currently doing. Is it playing chess? Invoke a game tree search. Doing mathematics? Invoke a theorem prover. Playing Jeopardy? Call Watson [Ferrucci, 2012]. Doing an impressionist painting? Run The Painting Fool [Colton, 2012].

Another thing we let be encapsulated in *nextAction* is reevaluating incoming information to find out if the agent should change what it would like to do and what it has decided to do. This is the *desires* and *intentions* part. Desires are loose commitments to a number of different, possibly conflicting goals. They are goals that the agent can chose to do, but which it hasn't yet committed to. Desires are evaluated against other desires, some being promoted to intentions. Adopting an intention indicates a resolve to making a real effort at achieving it: Imagine that I would like another cup of coffee and that I would also like to finish writing this paragraph. Both are desires. If choose to finish the paragraph, that desire becomes an intention. If stop writing in the middle this sentence and go get coffee, I was never really intent on finishing writing first. I also shouldn't be blindly committed to an intention. If the table catches fire as I'm writing, I would quite the fool if I kept tapping away at the keyboard.

It is generally understood that an agent cannot achieve all its desires, not even if its desires don't conflict [Wooldridge, 1996]. There simply isn't enough time in the day to do all one wants to! Picking out desires for promotion to intentions is in a sense an optimisation problem. An agent with limited resources (e.g. time, money, energy) must try to choose a number of goals that it honestly believes it can achieve, in such a way as to maximise its overall utility (or happiness or reward). Keeping in line with Minsky's admonitions, it is expected that this optimisation is difficult and highly application-specific. Indeed, the BDI literature seldomly attempts to give definite procedures for how to manage desires and intentions, instead providing frameworks for programming and verifying agents [Dennis et al., 2007, Hepple et al., 2007].

Next in the loop is *do* and *perceiveEnvironment*. These two subroutines connect the agent to the external world, allowing it to carry out actions and receive input from sensors. We do not worry about specifics, except expecting that *do* does the action and *perceiveEnviroment* returns information in some suitable format. In the final stage of the loop, *updateModel* takes the incoming sensory information and the chosen action and revises the internal model. With this done, the loop can start over, going through action selection, action execution and model updating once again.

On our agenda is defining suitable models for believing and acting in uncertain environments, including when that uncertainty is of a social nature. We are focused on how to build Automated Planning procedures for use in $nextAction$ and, in Chapter 5 where I return to the agent loop, how an agent can maintain a suitable model of the world. More specifically we: 1) identify and adapt models that are suitable for modelling agent's mental models of the world when it is partially observable, nondeterministic and includes other (possibly acting) agents 2) investigate how to model actions that change agent's mental models and use these to capture how knowledge and beliefs change when the world changes 3) use these models of the world and models of actions to create procedures that lets an agents figure out what to do next 4) investigate the properties of these models and the logical languages that can be interpreted on them.

There are, of course, things not (yet) on our agenda. For one, we do not deal with generating desires and choosing intentions, instead presupposing that our agents are supplied with meaningful goals from the outset. Neither do we deal with how agents interact with the environment. We are interested in what goes on inside the head of agents as they plan, act, and perceive, not how the inside and outside connect. The final major issue that we have sidestepped is that of specification. While the models we use are very useful for the kind of issues we wish to deal with, they can be difficult for the uninitiated to construct. All three issues, but particularly the first and the third, are important if we want our methods to see mainstream acceptance and use.

1.2 Automated Planning

A powerful analogy for what it means to reason and plan in an uncertain and social world can be found in games. I will return to this analogy throughout what remains of this chapter, so it useful to elaborate on it a bit.

Possibly the most common category of objects evoked by the term ‘game’ are classical adversarial games like Chess, Connect Four and Backgammon. These are two-player games with precisely defined rules dictating how a game begins, what moves are legal and how they change the board, and how a game is won or lost.

There are also games for just one player, more commonly called puzzles. Examples are Peg Solitaire, Sliding Puzzles and Rubik's Cubes. Again they have precise rules governing legal moves and solutions. A Rubik's Cube, for instance, is “played” by first scrambling the cube and then recovering the initial configuration such that each 3×3 face has the same colours. The mechanics of a Rubik's Cube enforce the the rules, because the way the cube can be twisted coincides with the allowed moves. Though it is possible with some Cubes to peel off coloured stickers and

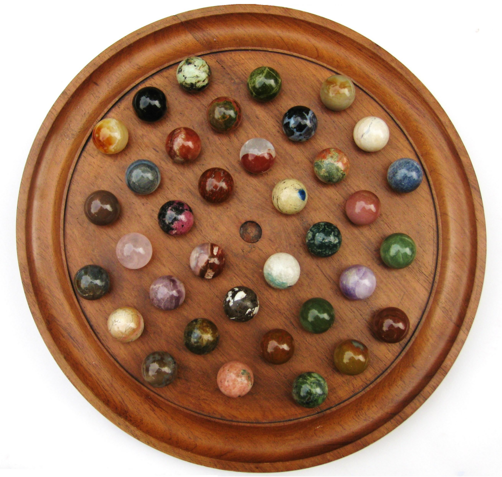

Figure 1.1: The initial board in the European version of Peg Solitaire.

putting them back onto the cube in the correct positions, this is not a solution. The move that peels stickers off and puts them back on falls outside the rules.

Games can have more than two players. Like Risk, they can encourage the formation of (possibly temporary) coalitions, or, like Pandemic, be cooperative with players winning or loosing against the game itself. They can be partially observable as most card games are. They can contain nondeterminism like throwing dice or drawing from a deck.

1.2.1 Classical Planning

A puzzle such as Peg Solitaire is a good place to begin this exposition on planning. The initial board setup in the European version of Peg Solitaire can be seen in Figure 1.1. It is a game for one player, in which the objective is to clear the board of marbles. A marble is cleared off the board when an adjacent marble jumps over it, which it can only do when there is a free position on the other. Initially the only free space is the middle position, which is also where the solitary final marble must end up.

Peg Solitaire is an example of a problem that can be modelled model with the the simplest and earliest kind of automated planning, aptly called *classical planning*. A generic problem solver for classical planning, the Stanford Research Institute Problem Solver (STRIPS) was introduced in 1971 [Fikes and Nilsson, 1971]. In establishing the classical planning formalism, it became possible to make computers solve all problems sharing the same essential features (given enough time and memory), without having to program them from scratch for each problem. All that was required was a specification of the "rules of the game" using a suitable description language. The specification language became know as STRIPS and was initially very complex, including arbitrary first-order formulas. It turned out to be difficult to give precise semantics for this language, so instead a formulation based on propositional logic was adopted [Ghallab et al., 2004].

The canonical way to formulate a planning problem in automated planning is analogous to the starting position, legal moves and win/loss condition(s) of games as identified earlier. A planning problem consists of a description of the initial state, an action library, and a goal formula. Figure 1.2 shows a problem description of a 1-dimensional version of Peg Solitaire in the now standard Planning Domain Definition Languge (PDDL) [Ghallab et al., 1998]. Initially there are marbles at position 1, 3 and 4, and the goal is to have only a marble at position 3.

Any planner capable of parsing PDDL and will be able to solve the problem, coming up with the plan [`move(4,3,2), move(1,2,3)`]. This plan is a sequence of moves transforming the initial state into a state satisfying the goal.

PDDL makes no assumptions about how the planner models states and actions, but the simplest model for classical planning is the set theoretic one. In the set theoretic formulation, a state is set of propositions $S \subseteq P$, where P is the set of all propositions in the domain. Propositions which never change their truth value are called rigids. For this particular example the rigids are `(IN-LINE 1 2 3)` and `(IN-LINE 2 3 4)`. As they never change their truth value, they are left out for visual clarity.

Example 1.1. With prettier notation the initial state is the set $\{occ_1, occ_3, occ_4\}$. For a given state, the planner can check which actions can be done (are applicable) by testing its preconditions. Here there is only one action and checking if it is applicable is a matter of finding an instantiation of the parameters `from`, `over` and `to`, such that positive preconditions are in the state and negative ones are not. Letting `from` $= 4$, `over` $= 3$ and `to` $= 2$ and naming the initial state S_0 we have that $\{occ_4, occ_3\} \subseteq S_0$ and $\{occ_2\} \cap S_0 = \emptyset$. This means that $move(4,3,2)$ is applicable. The results of doing $move(4,3,2)$ in S_0 is the set of propositions we get when first removing all negative effects and then adding the positives. For $move(4,3,2)$, the new state is $S_1 = S_0 \setminus \{occ_4, occ_3\} \cup \{occ_2\} = \{occ_1, occ_2\}$. Checking whether the goal has been achieved is done in the same way as applicability: Letting g^+ be the set of positive goal literals and g^- be the negative ones, g holds in S if $g^+ \subseteq S$ and $g^- \cap S = \emptyset$. Here we have $\{occ_3\} \not\subseteq S_1$, so the goal hasn't been achieved. In S_1, an applicable instantiation of $move$ is $move(1,2,3)$. Applying $move(1,2,3)$ we get the new state $S_2 = S_1 \setminus \{occ_1, occ_2\} \cup \{occ_3\} = \{occ_3\}$. Now we have $\{occ_3\} \subseteq S_2$

```
(define (domain pegsolitaire)
    (:requirements :strips :typing)
    (:types location)
    (:predicates
        (IN-LINE ?x ?y ?z - location)
        (occupied ?l - location)
    )

    (:action move
     :parameters (?from - location ?over - location ?to - location)
     :precondition (and
                       (or
                        (IN-LINE ?from ?over ?to)
                        (IN-LINE ?to ?over ?from))
                       (occupied ?from)
                       (occupied ?over)
                       (not (occupied ?to))
                   )
     :effect (and
                 (not (occupied ?from))
                 (not (occupied ?over))
                 (occupied ?to)))
)

(define (problem pegsolitaire-1)
    (:domain pegsolitaire)
    (:init (IN-LINE 1 2 3) (IN-LINE 2 3 4)
     (occupied 1) (occupied 3) (occupied 4) )
    (:goal and (
        (not (occupied 1))
        (not (occupied 2))
        (occupied 3)
        (not (occupied 4))))
```

Figure 1.2: PDDL specification of a one-dimensional Peg Solitaire problem.

and $\{occ_1, occ_2, occ_4\} \cap S_1 = \emptyset$. The problem has been solved! Returning the plan for solving the problem is then just a matter of backtracking through the generated states. ■

We see how the initial state and action descriptions induces a graph where nodes are states and edges are actions. Such a planning scheme is called state-space search and it is by far the most common way of solving planning problems [Ghallab et al., 2004]. For the set theoretic representation of classical planning, the transition from one state to the next is defined as follows:

Definition 1.2. Let $S \subseteq P$ be a state and P a finite set of propositions. With A denoting an action, we let $pre^{+}(A) \subseteq P$ and $pre^{-}(A) \subseteq P$ be the positive and negative preconditions, and $post^{+}(A) \subseteq P$ and $post^{-}(A) \subseteq P$ be the positive and negative effects. We require that positive and negative preconditions/effects do not contain the same literal when instantiated, i.e. $pre^+(A) \cap pre^-(A) = \emptyset$ and $post^+(A) \cap post^-(A) = \emptyset$.

The action A is applicable in state S if

$$pre^+(A) \subseteq S \text{ and } pre^-(A) \cap S = \emptyset$$

and the results of doing A in S is then

$$\gamma(S,A) = \begin{cases} S \setminus post^-(A) \cup post^+(A) & \text{if } A \text{ is applicable in } S \\ \text{undefined} & \text{otherwise} \end{cases}$$

This definition enforces a number of limiting assumptions that lie at the heart of classical planning.

Full observability The value of a proposition is always known. Either $p \in S$ and it is true, or $p \notin S$ and it is false.

Determinism For any state S and any applicable action A, $\gamma(S,A)$ gives a single new state. The outcome of an action is alway deterministic.

Static Only the planning agent's action change the environment.

Offline While not a consequence of the definition of the transition function per se, classical planning is offline planning. This means that the planning agent generates the full plan before ever acting. The other three assumptions makes this possible because *nothing unexpected ever happens*.

While a state-space can be seen as a fairly standard labelled transition system, it is important to recognise that the transition system is not the input. Arguably *the* defining feature of automated planning is that the state-space being searched is *generated during the search*. This is a marked difference from related formalisms, like game theory and formal verification of games, where the transition system comes *ex nihilo* (e.g. as in [Osborne and Rubinstein, 1994, Alur et al., 1998]). It also invites caution when investigating decidability and complexity for planning.

If we allow negative preconditions and effects, the problem of deciding whether a plan exists in classical planning with the set theoretic representation is PSPACE-complete [Ghallab et al., 2004]. Just the number of distinct states is $O(2^{|P|})$, where P is given in the description of the planning problem. This is far larger than the input. This stands in sharp contrast with for instance [van der Hoek and Wooldridge, 2002] where a procedure for multi-agent planning with knowledge goals is called tractable. The reason for this incongruence is that they measure complexity in terms of the size of the transition system, not in the size of the problem description [van der Hoek et al., 2006]. As we are interested in algorithms solving a general class of problems from a compact description, we must measure complexity in the size of the description of the problem. After all people (and computers) play chess by knowing the rules, not by memorising all possible board positions.

While planning at first glance might seem very expensive, it isn't all bad. That the state-space is much larger than the description tells us that we can describe very large systems very compactly. Turning the complexity discrepancy between planning and e.g. the transition systems of [Alur et al., 1998, van der Hoek et al., 2006] on its head, planning might boast that its domain descriptions are exponentially more succinct.

Another silver lining is that useful information about the dynamics of the generated transition system is contained in the description. This has allowed great progress to be made in the development of heuristics for state-space search [Helmert and Geffner, 2008, Helmert, 2004].

1.2.2 Nondeterminism & Partial Observability

Relaxing the assumptions of determinism and full observability allows the modelling of far more complex and interesting problems. Uncertainty in states cannot be modelled a subset of P. Sets of states called *belief states* are used instead, i.e. $B \subseteq 2^P$. The meaning is that a proposition is known to be true if it is in all the states in a beliefs state, false if it is not in any of them, and not known if it is in some, but not in others. In [Bertoli et al., 2003] an extension to PDDL called NPDDL is given, supplying constructs for modelling uncertainty about the initial state and action effects as well as constructs for observation variables or rules.

Example 1.3. With $P = \{p, q, r\}$ and the belief state $B_0 = \{\{p, q\}, \{q\}\}$ we have that q is known to be true, r is known to be false, and the value of p is unknown. The initial state and actions are described using $unknown(p)$—p is either true or false—and $oneof(p, q)$—either p is true or q is true, but not both. If we have an action A with precondition q and effects $unknown(r)$ (as nondeterminism is modelled in NPDDL), then A is applicable because q holds in all states in B_0, and the result is the set of states resulting from applying A to each individual state. For a suitably redefined transition function γ, we get $\gamma(B_0,A) = \{\{p, q, r\}, \{p, q\}, \{q, r\}, \{q\}\} = B_1$.

Rules of (roughly) the form $observation((O_r \Leftarrow (r \wedge q)) \wedge (\neg O_r \Leftarrow (\neg r \wedge q))$ define observation variables like O_r and the conditions under which their values are determined. If q does not hold, then the value of O_r is undetermined. If q does hold, then the value of r determines whether O_r is true or false. The value of O_r then determines a partitioning of a belief state into a set of belief states agreeing on O_r. For B_1 the partitioning is $\{\{p,q,r\},\{q,r\}\}$ where O_r is true and $\{\{q,r\},\{q\}\}$ where O_r is false. ■

Applying an action in planning with belief states is then a two-step process, where a set of belief states is first computed from a belief state and an action, and then partitioned according to values of the observation variables (see [Ghallab et al., 2004] for formal definitions). Again we see how the problem description induces a graph, this time corresponding to a nondeterministic transition system. Nodes are belief states, actions are again edges, but now an action can label several outgoing edges, each one corresponding to one possible outcome.

There is then the question of what a plan for such a problem is. A *weak solution* is one in which the plan *might* lead to a goal state. Depending on the particulars of the domain being modelled, it may be a worthy strategy for an agent to find only a weak solution. If things turn out differently than expected, the agent can do online replanning. A related replanning strategy for coping with partial observability is to plan only until a branching point is encountered. Executing the plan up until that point, the agent finds out which of the branches become reality and only has to deal with that one.

For the careful agent, a *strong solution* is one that *guarantees* reaching a goal state, by taking all possible branches into account. Finally, a *strong cyclic solution* is one which is guaranteed to *eventually* reach a goal state, provided that the state-transitions are fair (all transitions have a non-zero chance of happening). Finding out whether a plan exists for nondeterministic planning with full observability (where belief states are singleton sets) is EXP-complete, EXPSPACE-complete for partial observability without branching (actions may be nondeterministic, but plans must be sequences of actions), called *conformant planning* and 2EXP-complete for partial observability with nondeterminism and branching [Rintanen, 2004].

The single-agent planning formalism that is presented in chapter 2 is another approach to the informally defined partially observable and nondeterministic problem. There, observability is incorporated into actions rather than separately defined observation variables. On the one hand it means that sensing can be done without changing the state. On the other hand it means that all sensing must be defined in actions themselves. Well worth mentioning is that [Jensen, 2013b] shows that *epistemic planning* (an umbrella term for kind of planning presented in this thesis) has the same complexities for the above variations of single-agent planning as do the traditional belief state representations.

1.3 Dynamic Epistemic Logic in Automated Planning

Epistemic logic is an area of (modal) logic concerned with representing and reasoning about the knowledge and beliefs of agents. It is of particular relevance in multi-agent systems, where agents have knowledge and beliefs about the knowledge and beliefs of other agents. In fact, knowledge and belief statements can be nested arbitrarily many times giving indeed very long (or infinite) chains of introspection into the mind of others. The multi-agent aspect means that epistemic logic is well-suited for dealing with Theory of mind issues. We touch upon these in Section 1.4.

Work on epistemic logic was initiated by [von Wright, 1951], and gained mainstream recognition in the logic and philosophy communities with [Hintikka, 1962], where the now standard possible worlds semantics was proposed. In the possible worlds semantics, an agent knows or believes the formula ϕ at a world w, if ϕ holds in all the worlds connected to w. Conditions on how worlds are connected to each other determine whether the concept being treated is knowledge or belief.

1.3.1 States and Actions in DEL

The original epistemic logics dealt only with static representations of epistemic (knowledge) and doxastic (belief) situations. Such static situations are modelled as *Kripke models*, basically a labelled graph where nodes are possible worlds and an edge from a world w_1 to another world w_2 determines that at w_1, w_2 is considered possible. With a given set of symbols P, each world is labelled by a valuation V determining which symbols are true there. That p is true at w if $p \in V(w)$ and false if $p \notin V(w)$ should begin to make the parallel between epistemic Kripke models and states in automated planning clear. Making the parallel even clearer, consider that belief states are sets of propositions about the world, each set representing a possible world. With edges between worlds designating indistinguishability (the agent does not know which of the connected worlds is the actual one), an epistemic model can express all that a belief state can [Löwe et al., 2011a, Bolander and Andersen, 2011].

Example 1.4. The belief state $B_0 = \{\{p, q\}, \{q\}\}$ from earlier corresponds to the single-agent epistemic model consisting of two worlds w_1 and w_2 where $V(w_1) = \{p, q\}$ and $V(w_2) = \{q\}$. The accessibly relation R says which worlds the agent considers possible. For this model we would have $R(w_1) = \{w_1, w_2\}$, meaning that if the actual world is w_1 then the agent cannot tell the difference between w_1 and w_2. Similarly we would have $R(w_2) = \{w_1, w_2\}$. In B_0, q is known because q is in every set of the belief state. In the epistemic model M, q is known at a world w, written $M, w \models Kq$ if it holds in all worlds in $R(w)$. Thus Kq holds at both w_1 and w_2, and therefore q is known in both B_0 and M. ■

By associating an accessibility relation R_i to each agent i, an epistemic model can express higher-order knowledge and belief of multiple agents. In this way we can encode knowledge and beliefs about others' knowledge and beliefs ad infinitum, allowing statements like *a knows that p is true, but doesn't know whether b knows that p is true*: $K_a p \wedge \neg K_a K_b p \wedge \neg K_a \neg K_b p$. This cannot be done with belief states. This is the first indication that epistemic concepts are a powerful addition to planning. But without actions we get nowhere.

It wasn't until [Baltag et al., 1998] that a general framework for expressing the dynamics of knowledge and belief became available. What they introduce has been called many names over the years, but the name *event models* seems to have been settled upon. An event model is much like a Kripke model, but describes how one epistemic model changes into another. The original formulation of event models only encoded changing knowledge and beliefs. Later postconditions were added by [van Ditmarsch and Kooi, 2008] making it possible for actions to change atomic facts, in addition to changing knowledge and belief about such facts. Instead of worlds they contain events with preconditions and postconditions (corresponding to the preconditions and effects of classical planning actions). As with epistemic models, event models describe indistinguishability for agents with a relation Q_i for each of them. An operation called the *product update* produces new epistemic models by doing a kind of multiplication of an epistemic model and an event model.

Example 1.5. The action A from the previous section can be represented as an event model with two events e_1 and e_2, with preconditions $pre(e_1) = pre(e_2) = q$ and postconditions $post(e_1) = r$ and $post(e_1) = \neg r$. Further $Q(e_1) = \{e_1, e_2\}$ and $Q(e_2) = \{e_1, e_2\}$. The agent cannot tell which of the two events happen, so the effect corresponds to $unknown(r)$. The sensing of the observation variable O_r can be encoded as another event model with three distinguishable events. They make no ontic changes, as changes of atomic facts are called, so all have postconditions $\top$. Their preconditions encode the sensing: One has $r \wedge q$ corresponding to O_r being true, one has $\neg r \wedge q$ corresponding to O_r being false, and one has $\neg q$ corresponding to O_r being unknown. Because all events of both event models have their preconditions satisfied in a world in the epistemic model, both event models are applicable. ■

Note that, though it is possible to match the belief state formulation of planning completely, event models as we use them, and as they are generally used, will not separate factual changes and epistemic changes.

In work predating the the results in this thesis, we showed how epistemic planning [Bolander and Andersen, 2011] encompasses both single-agent classical planning, conformant planning and their generalisation to multi-agent epistemic models. In epistemic planning, the planning problem is defined in the same way as for the rest of automated planning: Given an epistemic model encoding the initial state, a library of event models representing available actions, and a goal formula, find a (possibly branching) plan composed of event models such that following the plan produces a state in which the goal formula holds. All these notions are defined formally in chapter 2.

In [Baltag and Smets, 2008b] epistemic models and event models are amended with a plausibility ordering on worlds and events. These allow encoding what each agent believes. The smaller a world or event is in the ordering, the more plausible the more it is believed. Adding postconditions to event models with plausibilities, Chapter 3 presents a new and more fine grained understanding of the dynamics, letting the planning agent distinguish between different degrees of ordinary and extraordinary outcomes. In chapter 3, we present a formalism for the partially observable nondeterministic case with plausibilities. This allows the planning agent to plan for only those outcomes it expects and then do *online replanning* if things went the extraordinary way. Such solutions are called weak or strong *plausibility solutions*, depending on whether they achieve the goal for *at least one* or *all* of the expected outcomes.

One peculiar effect of adding ontic change to event models is that the product update may produce models with superfluous worlds. An epistemic model containing two worlds both having the same valuation V makes all the same formulas true as the epistemic model with only a single world with valuation V. The structural notion of bisimulation which can be computed in finite time on finite models (as we use for planning) lets us identify models making all the same formulas true, even though there are infinitely formulas [Blackburn et al., 2001]. Because it ensures that the state-space contains only a finite number of distinct states when taking bisimulation into account, this is a crucial property if we want decidability for the two single-agent frameworks we present: While we might not be able to find a plan even for a finite search-space, we can at least know when there's nothing more to check.

Bisimulations also lets us compute the contraction of a model as the smallest possible logically equivalent model. Whether the framework in question is decidable or not, this is obviously an attractive ability to have. General multi-agent epistemic planning, even without plausibilities, is shown in [Bolander and Andersen, 2011] to be undecidable. More recently [Aucher and Bolander, 2013] showed this to be the case even without factual change, while [Yu et al., 2013] and [Löwe et al., 2011a] identify decidable fragments. Despite general multi-agent planning being undecidable, it is possible to come up with problems where a strong solution (or at least a strong cyclic one) can be found if bisimulations are part of the picture. Bisimulations are crucial, even if they don't give decidability.

For the framework in chapter 2, it suffices to use standard bisimulation. When we get to the plausibility planning framework in chapter 3 this notion ceases working. We therefore come up with a new type of bisimulation for plausibility models. With this new notion, decidability is reestablished. Chapter 4 generalises this notion to multi-agent plausibility models in a not very straight forward way.

How we employ the models of DEL warrants a caution for those readers familiar with the formalism. The vantage point in modal logic is usually that of an external observer that sees all that goes on inside the system and in the minds of agents. This is not the case for us. The formalisms we present are intended for the agent(s) the system—the agents are modelling themselves. This difference in the external vs. internal perspectives was treated in [Aucher, 2010] and [Bolander and Andersen, 2011], and it discussed again in later chapters.

Note that our definition of planning is not shared by all. One already mentioned dissenting voice is [van der Hoek and Wooldridge, 2002]. Another is [Aucher, 2012], wherein epistemic planning is the problem of finding a single event model under restrictions, that transforms one epistemic model into another. While a procedure for finding the missing link between particular epistemic models has applications, e.g. for an agent learning action descriptions, it is not what my co-authors and I (and many others) mean by planning.

1.4 Social Reasoning

From our earliest work in epistemic planning we have been interested in endowing agents with the capability to reason about the higher-order knowledge and beliefs of others [Bolander and Andersen, 2011]. Agents able to do such social reasoning are commonly considered to have a Theory of mind (ToM). The concept comes from cognitive science where [Premack and Woodruff, 1978] describes it as the ability to attribute mental states like belief, desires and intentions to others, particularly mental states *different from one's own*.

The canonical example ToM is the Sally-Anne test from [Baron-Cohen et al., 1985], where the test subject is a child being gauged for his or her ability to reason about others. The test presents the child with a doll, Sally, who leaves a marble in a basket, before departing the scene. In Sally's absence another doll, Anne, moves the marble to a box. Children are asked to predict where Sally will look for the marble when she returns. Figure 1.3 illustrates the story.

[Baron-Cohen et al., 1985] reports subjecting 61 children, some normally developing, some with Down's syndrome, some diagnosed with autism, to the Sally-Anne

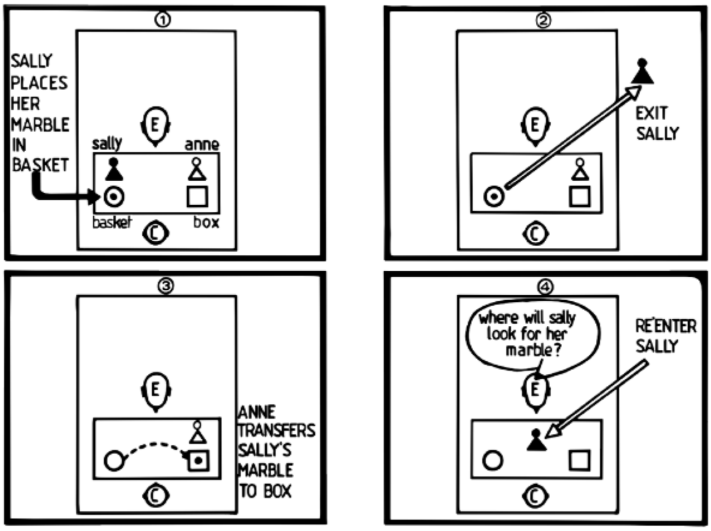
①
SALLY PLACES HER MARBLE IN BASKET
sally
anne
basket
box
②
EXIT SALLY
③
ANNE TRANSFERS SALLY'S MARBLE TO BOX
④
where will sally look for her marble?
RE'ENTER SALLY

Figure 1.3: Diagram of the story of Sally and Anne.

test by asking them three questions: The memory question of *where the marble was initially*, the reality question of *where the marble is now*, and the false belief question of *where Sally will look for her marble*. The correct answer to the latter is the basket. All children answered the memory and reality questions correctly. For the false belief question, the passing rate was roughly 85% for normally developing children, 86% for children with Down's syndrome, and just 20% for autistic children.

The Sally-Anne test as a test of ToM-capabilities has been called into question by, among others, [Bloom and German, 2000]. They argue that is problematic to conclude whether or not a child has a ToM based on just a false belief task like this, giving reasons why "there is more to passing the false belief task than theory of mind" and that "there is more to theory of mind than passing the false belief task". They provide a reasonable augment that linguistic and other cognitive impairments may be at play in children who fail false belief tasks.

Another point of contention is how ToM behaviour is produced. The competing views of whether ToM is a *theory theory* or a *simulation theory* are discussed in [Michlmayr, 2002]. In the theory theory view, the understanding of minds is produced by a framework of rules. For theory theorists, the ability to answer the false belief task comes from rules like "people will tend to look for a things where they remember they were last." In this view, the ability to correctly answer the false belief question comes from connecting the correct answer to the memory question with the look-where-they-last-saw-it rule.

For simulation theorists, answering the false belief question correctly comes from an ability to *imagine being someone else*. In this view, arriving at the correct answer comes from taking on the perspective of Sally and simulating the unfolding events from her point of view. Because simulating the effects of sequences of actions is exactly what planning is about, and for epistemic planning particularly the effects on the higher-order beliefs of several agents, simulation theory is compatible with our research agenda. It is hard to see that theory theory is.

I must, of course, concede that this is not an argument for the correctness of simulation theory. In fact, evidence seems to suggest that a hybrid of simulation theory and theory theory is needed to account for ToM [Michlmayr, 2002, Vogeley et al., 2001]. In the hybrid view, we use theory theory in situations that resemble what has previously been encountered. In situations that are sufficiently different from what we have seen before, we make use of more cognitively demanding simulation processes.

The hybrid view is similar to the 'System 1' and 'System 2' thinking of [Kahneman, 2011]. System 1 thinking is fast, instinctual, subconscious and cheap. It is what we use most of the time and in known situations. System 2 thinking is slow, logical, conscious and expensive. In unknown situations we use System 2 thinking, because our experience cannot help (much) in forming thoughts and opinions about what to do and believe. After repeated exposure to a particular System 2-requiring situation, a recipe of sorts sinks down to the System 1-level. There it lies ready for future use at much lower cognitive cost. The idea of ready made plans is not new to agent systems. For instance, the agent model in [Wooldridge, 1996] uses beliefs about the current state to retrieve a plan for the current goal from a database supplied by a designer, instead of computing from scratch. To my knowledge, a system that computes plans in new situations and stores them for later reuse has never been proposed.

Approaching AI as engineers, we sidestep the problem of how ToM works in people by admitting that we are interested in creating machines that exhibit ToM behaviour, not whether people think in that way.

1.4.1 ToM – What is it good for?

At this point, a pertinent question to address is what a ToM is good for. One answer is that a ToM helps make sense of how others behave and expect oneself to behave.

As an analogy consider that we have an understanding of how the physical world

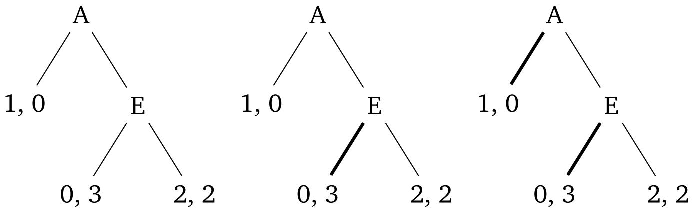

Figure 1.4: Game tree where backwards induction leads to suboptimal outcomes for both players. The example comes from [Van Benthem, 2011].

behaves. Generally things fall down, warm liquids cool, cool liquids warm. Even a very rudimentary understanding of physics grounded in nothing more than everyday experience helps us cope with complexity by allowing quite accurate predictions of how objects behave. We don't take brownian motion into account for everyday-sized objects, nor do we include contingencies in our plans in case things start falling up. By not considering such odd and implausible behaviour saves enormous amounts of cognitive resources. In addition, our theory of object behaviour needs only be useful at the level where we most often need to predict and act. A more scientifically precise formulation of the "things fall down" rule is that two things fall towards their common centre of gravity. At the everyday level however, such distinctions make no difference. What "things fall to the ground" lacks in precision is made up for in simplicity.

Similarly, a ToM is an understanding of "social physics". It lets us form beliefs about the beliefs of others, thereby allowing us predict their behaviour pretty accurately. Our expectations of what others are going to do greatly reduces the cognitive burden of being in social situations. Same as a theory of physical-object behaviour, we have a theory of things-with-minds behaviour. If others are going to act, and they most likely are, then the extra cognitive burden of having a ToM seems well worth it.

Moreover, a ToM enables cooperation and fruitful negotiation. Figure 1.4 shows a game tree where the canonical backwards induction algorithm[2] produces suboptimal outcomes for both players. The numbers show payoffs for player A and E respectively. The thick edges show backwards induction iteratively picking optimal moves from the bottom up. With backwards induction, player A reasons as follows: "The best E can do is picking his left move giving a payoff of 3. For this move I get 0. Therefore my payoff if I pick my right move is going to be 0. The payoff for my left move is 1. The best I can do is then choosing the left move, giving me a payoff of 1 and E a payoff of 0." If instead A and E cooperated, they would both be better off! Similarly E can reason that if A picks the right move, then A did so with the

[2] See [Osborne and Rubinstein, 1994] for a through treatment.

expectation of cooperation. If E chooses the left move, giving E 3 and A 0, then E should not expect A to play nice in the future.

Through extensive experimentation in a negotiation game with incomplete information, [de Weerd et al., 2013] confirms that agents with a ToM perform better than agents without one. They also show that when second-order ToM agents negotiate, neither agent has any incentive to deviate from outcomes that maximise their collective gain. Deviating from mutually beneficial moves leads to "distrust" (though not explicitly modelled) and makes even the defecting agent worse off in the long run.

Revisiting the analogy of games, the famous philosophical concept of language-games suggest that language broadly, and specifically dialogue, can be understood as a game with moves and objectives [Wittgenstein, 1953]:

> Let us imagine a language... The language is meant to serve for communication between a builder A and an assistant B. A is building with building-stones; there are blocks, pillars, slabs and beams. B has to pass the stones, and that in the order in which A needs them. For this purpose they use a language consisting of the words 'block', 'pillar', 'slab', 'beam'. A calls them out; –B brings the stone which he has learnt to bring at such-and-such a call. – Conceive this as a complete primitive language.

Later words like "this", "that", and "there" and gestures like pointing are added, but the basic idea remains: A has moves that changes the state of the game, and B responds to the new state with other appropriate moves.

When playing a language game, a negotiation game, or some other game (in the broadest of definitions), which includes higher-order knowledge and belief, it is worthwhile to attempt ToM modelling.

1.5 Other Related Work

The previously cited [van der Hoek and Wooldridge, 2002], uses Alternating Time Epistemic Logic (ATEL) [van der Hoek and Wooldridge, 2003], an extension of Alternating Time Logic (ATL) [Alur et al., 2002] with partial observability. As shown by [Goranko and Jamroga, 2004], ATL and ATEL subsumes the similar Coalition Logic of [Pauly, 2002]. ATL/ATEL are logics for reasoning about the abilities of agent coalitions, interpreted on the concurrent game structures of [Alur et al., 2002] (though the authors have proposed two earlier semantic structures). These are transition systems in which each agent has a set of allowed moves at each state. The joint move of the coalition of is simply the set of moves of each individual agent. While in principle deterministic, nondeterminism can be modelled by adding an agent in place of the environment. The main modality in ATL is $\langle\langle A\rangle\rangle\phi$, expressing that the coalition of agents A has a strategy (a choice of moves in the given state) for enforcing ϕ, regardless of what agents not in A do. In ATL, the group choice operator $\langle\langle A\rangle\rangle\phi$ must be followed by a temporal operator; either the next state operator X, the always operator G or a ϕ until ϕ' operator. In ATL*, no such restrictions are imposed.

Adding partial observability to ATL, thereby yielding ATEL, is just a matter of adding an indistinguishability relation on states for each agent. Though technically simple, this has two profound implications for model checking ATEL, the second being a consequence of the first. For one, if we are to interpret $\langle\langle A\rangle\rangle$ as saying that agents have a strategy, then we must reconcile the difference between agents knowing that a strategy exists, and knowing what the strategy is. The latter of these seems the right choice [Jamroga and Ågotnes, 2007]. You do not have a winning strategy if you do not know what that strategy is. To capture with this distinction [Jamroga and van der Hoek, 2004] have proposed ATL variants that are explicit about the distinction between perfect and imperfect information (i.e full vs. partial observability), and perfect and imperfect recall (i.e. how much of the history can be remembered by agents).

The important consequence of this distinction is that in the imperfect information setting with perfect recall (which most closely resembles planning), model checking ATL/ATL*-formulas is undecidable [Bulling et al., 2010]. This is in line with the undecidability results of epistemic planning in [Bolander and Andersen, 2011], and as a contrast to the tractable planning of [van der Hoek and Wooldridge, 2002]. The point is important, so I reiterate it again: Multi-agent planning is only decidable if agents need just need to find out *if* they have a strategy (as in [van der Hoek and Wooldridge, 2002]). The problem becomes undecidable when agents need to know *what* the strategy is [Jamroga and Ågotnes, 2007, Bolander and Andersen, 2011, Bulling et al., 2010].

The algorithms for checking satisfiability of ATL/ATL* of these sources, and those of [Gorankoa and Vester, 2014, Vester, 2013] (including the case of finite memory) are constructive, meaning that they can be used to explicitly generate strategies. More than this; the complexities of model checking the different variants when taking the *size of the transitions system as input*, though not directly relatable, correspond to results for single-agent planning [Bulling et al., 2010]. Also worth mentioning is that the finite memory case where strategies that can only depend on the last k states, corresponds to the problem of deciding whether a plan of length k exists. This problem is decidable in planning, as it indeed is in ATL/ATL* [Vester, 2013].

In certain approaches—-knowing *what* the strategy is, letting the input be the *size* of the transition system, limiting the plan length—multi-agent epistemic planning and ATL/ATL* model checking are very much in the same boat.

An initial investigation of the relation between DEL and Epistemic Temporal Logic (ETL) comes due to [van Benthem et al., 2007]. ETL adds epistemic indistinguishability to temporal logic with linear or branching time. With public announcements (in DEL encoded as event models) they show many commonalities between the two approaches, though also showing that asynchronous ETL systems cannot be generated by DEL protocols (sequences of public announcement event models under constraints). It is unclear what the consequences of adding postconditions and applicability would be for the relation between epistemic planning and ETL. It does seem that true asynchronicity would be difficult to achieve.

The formalism in [Herzig et al., 2003] is equivalent to single-agent epistemic planning (as in Chapter 2), though states, called Complex Knowledge States, are syntactically described and epistemic and factual change are separate notions (an action changes one or the other, not both). Progressing actions, and therefore plans, is done by by syntactic manipulation of CKSs. With a CKS represented as a disjunction of K-formulas, the more is known, the bigger the representation. This is the reverse of the relationship between the size of epistemic models and uncertainty. In epistemic models, the less is known, the bigger the model. As with belief states mentioned earlier, there does not seem to be an elegant way to extend the framework to the multi-agent case. At least in comparison with this formalism, we think that DEL is the way to go.

1.5.1 Implementations of Social Reasoning

Situated speech is used in [Gorniak and Roy, 2005] to improve speech processing in a computer game where the player gives orders to a virtual assistant by speaking into a microphone. They present a simple game containing the player avatar, his assistant, a bear, and a barrel. When the player says "attack this [something that sounds like a cross between bear and barrel]", the speech recognition system generates a number of possible hypotheses about what was said. The highest ranked hypothesis is "attack this bear". If the player avatar is standing next to the barrel, their system rejects the *heard* phrase "attack this bear" in favour of the *understood* phrase "attack this barrel". By allowing language to be interpreted in a specific context—in this scenario where the player is standing—they drastically improve the quality of user interaction.[3] Further developments in [Gorniak, 2005] use plan recognition to disambiguate objects: In a second game, the player has to navigate

[3] See `http://www.media.mit.edu/cogmac/projects/games.html` for a video.

a number of rooms separated by doors, each operated by a remote lever. By understanding the player's plan, including the preconditions of the next action he wants to perform, the assistant correctly interprets the ambiguous "pull the lever", opening the right doors by pulling the right levers at the right times. [Foster and Petrick, 2014] also deals with the problem of sensor uncertainty, letting their *robot bartender* plan to ask for clarification if faced with ambiguous or low-confidence sensing hypotheses.

Though neither of the above systems have an explicit ToM, the idea that further quality improvements in speech recognition, natural language understanding, and human-computer interaction can come from ToM seems reasonable. Suppose that the speaker knows about only one object, say a key, while the listener knows about two keys. When the speaker says "(could you) hand me the key" the noun "key" is ambiguous from the listeners perspective. However, if the listener can take the speakers perspective, the ambiguity disappears: The speaker cannot be referring to the key he does not know exists. This is speech situated not in a physical context, but in a social context of knowledge about knowledge about the physical context.

In [Brenner and Nebel, 2009] multiple autonomous agents inhabit a dynamic environment. The environment is partially observable and agents have only limited influence over it. Agents can perform limited reasoning about the beliefs of other agents (which are like the belief states mentioned earlier, except with multi-valued variables). They introduce the novel plan construct *assertions* which are meta-actions used when an agent knows that it will later come to know how to act. When executing an assertion, a sub-plan is created that handles the uncertainty. In this way agents can plan for later information acquisition, safe in the knowledge that they will most likely be able to whatever happens. By experimentation in a multi-agent grid world, they show that planning with assertions makes agents with limited memory and sensing perform almost as well as agents with full observability. I see their scenario as being an exemplary test bed for agents with ToM.

The "Planning with Knowledge and Sensing planning system" (PKS) (used by the robot bartender mentioned earlier) is an implementation of a single-agent planning system explicitly modelling the agent's knowledge [Petrick and Bacchus, 2002, Petrick and Bacchus, 2004]. In PKS, an agent's knowledge consists of four databases translatable into formulas in epistemic logic (plus some first-order extensions), with actions modifying these databases separately. As argued in [Bolander and Andersen, 2011] this is a more natural approach to observability than using observation variables. Though the pragmatic approach they have taken means that some epistemic models cannot be represented, PKS seems to be the only serious epistemic planner for practical applications.

1.6 Outline of Thesis

Here follows an overview of the contents and main results of the remaining chapters, reiterating some of what has been mentioned already. Not mentioned below is Chapter 6, where I reflect on what has been achieved so far, and point out what I believe are the most pressing and promising avenues for future work on epistemic planning.

Chapter 2: This chapter is the long version of [Andersen et al., 2012]. This first foray into single-agent epistemic planning shows how to use epistemic models as states, event models as actions and epistemic formulas as goal formulas. We give a language of conditional plans and show a translation of weak and strong solutions into DEL. Finally we give a terminating, sound and complete algorithm for synthesising weak and strong solutions.

Chapter 3: This chapter has previously been published as [Andersen et al., 2014]. It expands the pure knowledge approach of the single-agent epistemic planning of Chapter 2 to one with plausibilties and therefore beliefs. This lets us define weak and strong plausibility solutions corresponding to weak and strong solutions for expected outcomes. We give a terminating, sound and complete algorithm for synthesising weak and strong plausibility solutions.

Chapter 4: This chapter is the first printing of joint work with Martin Holm Jensen, Thomas Bolander, and Hans van Ditmarsch on extending [Andersen et al., 2013] to the multi-agent case. It defines bisimulation on multi-agent plausibility models and investigates expressivity for three logics. The bisimulation conditions we give are in terms of a relation derived from the bisimulation relation itself. This is necessary because the conditional belief and degrees of belief modalities do not correspond to the plausibility relation, but to sets of world ordered by that relation. This means that while ordinary bisimulation definitions give that bisimilarity implies modal equivalence, modal equivalence does not imply bisimilarity. Defining bisimulations in the radically new way that we do reestablishes the correspondence. We also give a definition of safe belief in terms of the relation derived from the bisimulation, meaning that our version of safe belief corresponds to other definitions in terms of conditional belief, where something is safely believed if it is believed no matter what true information is given. This is the second indication that our bisimulation, though odd, is correct (the first being that modal equivalence means bisimulation).

In the second part we investigate expressivity for the logics of conditional belief, degrees of belief, and safe belief, and combinations of the three. A surprising result (to us as well!) is that the logics are not equally expressive, contradicting what we conjectured in [Andersen et al., 2013].

Chapter 5: This chapter presents work in progress on the generalisation of plausibility planning to multi-agent plausibility models. Though still only with a single acting agent, the work is a step on the road to multi-agent planning in style of [Brenner and Nebel, 2009]. The presented formalism arises from a combination of single-agent plausibility planning of chapter 3, the multi-agent models and bisimulation of chapter 4, and the multi-pointed models of [Bolander and Andersen, 2011].

Also presented is a way for the planning agent to cope with the representational problem of plausibility models, where everything that is considered epistemically possible must retained. The solution is to allow the agent to discard parts of the model that, while not known to be impossible, are believed not to be the case. Being only belief, it can of course happen that the agent was wrong to discard (some of) what was not believed. For such unlucky circumstances, belief revision is required, and a procedure for this is given as well. It bears noting that while chapters 2, 3, and 4 are all joint work with their respective authors, credit and blame for the ideas in chapter 5 should fall solely at my feet.

A note on models and notation: *The reader should beware that the models used in Chapters 2, 3 and 5 are intended for planning. When planning, the indistinguishability relation encodes future indistinguishability, while present indistinguishability is encoded by the entire set of worlds/events in single-agent versions, and sets of worlds/events in the multi-agent version. In Chapter 4 we do not need the distinction, so there the epistemic relation is derived from the plausibility relation as is standard. This point will be further discussed in the chapters diverging from mainstream traditions.*

As the material in these chapters has been produced over quite some time, notation has changed between the writing of the article versions of the different chapters. I have (attempted to) streamline notation throughout, using the same symbols for models, event models, etc. What has not been streamlined is the notation for the plausibility ordering. The meaning is unchanged; we write $w \leq v$ in Chapter 3 and $v \geq w$ in Chapter 4 and 5, so the discrepancy is simply a matter of mirroring. As Chapter 4 deals with new approaches to standard bisimulation, we decided to use $v \geq w$ to match usual notation for accessibility relations, where vRw means that w is R-accessible from v. This convention carried over into chapter 5.

Chapter 2

Conditional Epistemic Planning

This chapter is the extended version of [Andersen et al., 2012] which appears in the proceedings of the 13th European Conference on Logics in Artificial Intelligence (JELIA), 2012, in Toulouse, France. A few typing errors have been corrected.

Conditional Epistemic Planning

Mikkel Birkegaard Andersen Thomas Bolander
Martin Holm Jensen
DTU Compute, Technical University of Denmark

Abstract

Recent work has shown that Dynamic Epistemic Logic (DEL) offers a solid foundation for automated planning under partial observability and non-determinism. Under such circumstances, a plan must branch if it is to guarantee achieving the goal under all contingencies (strong planning). Without branching, plans can offer only the possibility of achieving the goal (weak planning). We show how to formulate planning in uncertain domains using DEL and give a language of conditional plans. Translating this language to standard DEL gives verification of both strong and weak plans via model checking. In addition to plan verification, we provide a tableau-inspired algorithm for synthesising plans, and show this algorithm to be terminating, sound and complete.

2.1 Introduction

Whenever an agent deliberates about the future with the purpose of achieving a goal, she is engaging in the act of planning. When planning, the agent has a view of the environment and knowledge of how her actions affect the environment. Automated Planning is a widely studied area of AI, in which problems are expressed along these lines. Many different variants of planning, with different assumptions

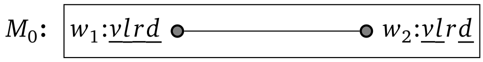

Figure 2.1: The initial situation. The thief is uncertain about whether r holds.

and restrictions, have been studied. In this paper we consider planning under uncertainty (nondeterminism and partial observability), where exact states of affairs and outcomes of actions need not be known by the agent. We formulate such scenarios in an epistemic setting, where states, actions and goals are infused with the notions of knowledge from Dynamic Epistemic Logic (DEL). Throughout this exposition, our running example, starting with Example 2.1, follows the schemings of a thief wanting to steal a precious diamond.

Example 2.1. After following carefully laid plans, a thief has almost made it to her target: The vault containing the invaluable Pink Panther diamond. Standing outside the vault ($\neg v$), she now deliberates on how to get her hands on the diamond (d). She knows the light inside the vault is off ($\neg l$), and that the Pink Panther is on either the right (r) or left ($\neg r$) pedestal inside. Obviously, the diamond cannot be on both the right *and* left pedestal, but nonetheless the agent may be uncertain about its location. This scenario is represented by the epistemic model in Figure 2.1. The edge between w_1 and w_2 signifies that these worlds are indistinguishable to the agent. For visual clarity we omit reflexive edges (each world is always reachable from itself). We indicate with a string the valuation at world w, where an underlined proposition p signifies that p does *not* hold at w.

The agent's goal is to obtain the jewel and to be outside the vault. She can enter and leave the vault, flick the light switch and snatch the contents of either the right or left pedestal. Her aim is to come up with a, possibly conditional, plan, such that she achieves her goal. ■

By applying DEL to scenarios such as the above, we can construct a procedure for the line of reasoning that is of interest to the thief. In the following section we recap the version of DEL relevant to our purposes. Section 2.3 formalises notions from planning in DEL, allowing verification of plans (using model checking) as either weak or strong solutions. In Section 2.4 we introduce an algorithm for plan synthesis (i.e. generation of plans). Further we show that the algorithm is terminating, sound and complete.

2.2 Dynamic Epistemic Logic

Dynamic epistemic logics describe knowledge and how actions change it. These changes may be epistemic (changing knowledge), ontic (changing facts) or both. The work in this paper deals only with the single-agent setting, though we briefly discuss the multi-agent setting in Section 2.5. As in Example 2.1, agent knowledge is captured by epistemic models. Changes are encoded using event models (defined below). The following concise summary of DEL is meant as a reference for the already familiar reader. The unfamiliar reader may consult [van Ditmarsch and Kooi, 2008, Ditmarsch et al., 2007] for a thorough treatment.

Definition 2.2 (Epistemic Language)**.** Let a set of propositional symbols P be given. The language $\mathcal{L}_{\text{DEL}}(P)$ is given by the following BNF:

$$\phi ::= \top \mid p \mid \neg\phi \mid \phi \wedge \phi \mid K\phi \mid [\mathcal{E}, e]\,\phi$$

where $p \in P$, $\mathcal{E}$ denotes an *event model* on $\mathcal{L}_{\text{DEL}}(P)$ as (simultaneously) defined below, and $e \in D(\mathcal{E})$. K is the epistemic modality and $[\mathcal{E}, e]$ the dynamic modality. We use the usual abbreviations for the other boolean connectives, as well as for the dual dynamic modality $\langle \mathcal{E}, e\rangle\,\phi := \neg\,[\mathcal{E}, e]\,\neg\phi$. The dual of K is denoted $\widehat{K}$. $K\phi$ reads as "the (planning) agent knows ϕ" and $[\mathcal{E}, e]\,\phi$ as "after all possible executions of $(\mathcal{E}, e)$, ϕ holds".

Definition 2.3 (Epistemic Models)**.** An *epistemic model* on $\mathcal{L}_{\text{DEL}}(P)$ is a tuple $M = (W, \sim, V)$, where W is a set of *worlds*, $\sim$ is an equivalence relation (the *epistemic relation*) on W, and $V : P \to 2^W$ is a *valuation*. $D(M) = W$ denotes the *domain* of M. For $w \in W$ we name (M, w) a *pointed epistemic model*, and refer to w as the *actual world* of (M, w).

To reason about the dynamics of a changing system, we make use of *event models*. The formulation of event models we use in this paper is due to van Ditmarsch and Kooi [van Ditmarsch and Kooi, 2008]. It adds ontic change to the original formulation of [Baltag et al., 1998] by adding postconditions to events.

Definition 2.4 (Event Models)**.** An *event model* on $\mathcal{L}_{\text{DEL}}(P)$ is a tuple $\mathcal{E} = (E, \sim, pre, post)$, where

- E is a set of *(basic) events*,
- $\sim \subseteq E \times E$ is an equivalence relation called the *epistemic relation*,
- $pre : E \to \mathcal{L}_{\text{DEL}}(P)$ assigns to each event a *precondition*,
- $post : E \to (P \to \mathcal{L}_{\text{DEL}}(P))$ assigns to each event a *postcondition*.

$D(\mathcal{E}) = E$ denotes the *domain* of $\mathcal{E}$. For $e \in E$ we name $(\mathcal{E}, e)$ a *pointed event model*, and refer to e as the *actual event* of $(\mathcal{E}, e)$.

Definition 2.5 (Product Update). Let $M = (W, \sim, V)$ and $\mathcal{E} = (E, \sim', pre, post)$ be an epistemic model resp. event model on $\mathcal{L}_{\mathrm{DEL}}(P)$. The *product update* of M with $\mathcal{E}$ is the epistemic model denoted $M \otimes \mathcal{E} = (W', \sim'', V')$, where

- $W' = \{(w, e) \in W \times E \mid M, w \models pre(e)\}$,
- $\sim'' = \{((w, e), (v, f)) \in W' \times W' \mid w \sim v \text{ and } e \sim' f\}$,
- $V'(p) = \{(w, e) \in W' \mid M, w \models post(e)(p)\}$ for each $p \in P$.

Definition 2.6 (Satisfaction Relation). Let a pointed epistemic model (M, w) on $\mathcal{L}_{\mathrm{DEL}}(P)$ be given. The satisfaction relation is given by the usual semantics, where we only recall the definition of the dynamic modality:

$$M, w \models [\mathcal{E}, e]\, \phi \qquad \text{iff } M, w \models pre(e) \text{ implies } M \otimes \mathcal{E}, (w, e) \models \phi$$

where $\phi \in \mathcal{L}_{\mathrm{DEL}}(P)$ and $(\mathcal{E}, e)$ is a pointed event model. We write $M \models \phi$ to mean $M, w \models \phi$ for all $w \in D(M)$. Satisfaction of the dynamic modality for non-pointed event models $\mathcal{E}$ is introduced by abbreviation, viz. $[\mathcal{E}]\, \phi := \bigwedge_{e \in D(\mathcal{E})} [\mathcal{E}, e]\, \phi$. Furthermore, $\langle \mathcal{E} \rangle\, \phi := \neg\, [\mathcal{E}]\, \neg \phi$.[1]

Throughout the rest of this paper, all languages (sets of propositional symbols) and all models (sets of possible worlds) considered are implicitly assumed to be finite.

2.3 Conditional Plans in DEL

One way to sum up automated planning is that it deals with the *reasoning side of acting* [Ghallab et al., 2004]. When planning under uncertainty, actions can be nondeterministic and the states of affairs partially observable. In the following, we present a formalism expressing planning under uncertainty in DEL, while staying true to the notions of automated planning. We consider a system similar to that of [Ghallab et al., 2004, sect. 17.4], which motivates the following exposition. The type of planning detailed here is *offline*, where planning is done before acting. All reasoning must therefore be based on the agent's initial knowledge.

[1] Hence, $M, w \models \langle \mathcal{E} \rangle\, \phi \Leftrightarrow M, w \models \neg\, [\mathcal{E}]\, \neg \phi \Leftrightarrow M, w \models \neg(\bigwedge_{e \in D(\mathcal{E})} [\mathcal{E}, e]\, \neg \phi) \Leftrightarrow M, w \models \bigvee_{e \in D(\mathcal{E})} \neg\, [\mathcal{E}, e]\, \neg \phi \Leftrightarrow M, w \models \bigvee_{e \in D(\mathcal{E})} \langle \mathcal{E}, e \rangle\, \phi$.

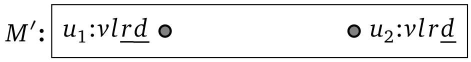

Figure 2.2: A model consisting of two information cells

2.3.1 States and Actions: The Internal Perspective

Automated planning is concerned with achieving a certain goal state from a given initial state through some combination of available actions. In our case, states are epistemic models. These models represent situations from the perspective of the planning agent. We call this the *internal perspective*—the modeller is modelling itself. The internal perspective is discussed thoroughly in [Aucher, 2010, Bolander and Andersen, 2011].

Generally, an agent using epistemic models to model its own knowledge and ignorance, will not be able to point out the actual world. Consider the epistemic model M_0 in Figure 2.1, containing two indistinguishable worlds w_1 and w_2. Regarding this model to be the planning agent's own representation of the initial state of affairs, the agent is of course not able to point out the actual world. It is thus natural to represent this situation as a non-pointed epistemic model. In general, when the planning agent wants to model a future (imagined) state of affairs, she does so by a non-pointed model.

The equivalence classes (wrt. $\sim$) of a non-pointed epistemic model are called the *information cells* of that model (in line with the corresponding concept in [Baltag and Smets, 2008b]. We generally identify any equivalence class $[w]_\sim$ of a model M with the submodel it induces, that is, we identify $[w]_\sim$ with $M \restriction [w]_\sim$. We also use the expression *information cell* on $\mathcal{L}_{\text{DEL}}(P)$ to denote any connected epistemic model on $\mathcal{L}_{\text{DEL}}(P)$, that is, any epistemic model consisting of a single information cell. All worlds in an information cell satisfy the same K-formulas (formulas of the form $K\phi$), thus representing the same situation as seen from the agent's internal perspective. Each information cell of a (non-pointed) epistemic model represents a possible state of knowledge of the agent.

Example 2.7. Recall that our jewel thief is at the planning stage, with her initial information cell M_0. She realises that entering the vault and turning on the light will reveal the location of the Pink Panther. Before actually performing these actions, she can rightly reason that they will lead her to know the location of the diamond, though whether that location is left or right cannot be determined (yet).

Her representation of the possible outcomes of going into the vault and turning on the light is the model M' in Figure 2.2. The information cells $M' \restriction \{u_1\}$ and $M' \restriction \{u_2\}$ of M' are exactly the two distinguishable states of knowledge the jewel

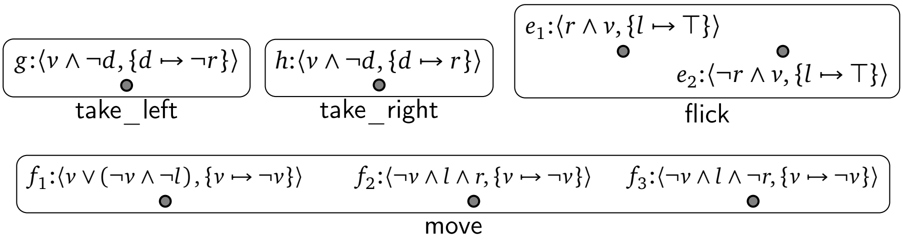
$g{:}\langle v \land \neg d, \{d \mapsto \neg r\}\rangle$
take_left
$h{:}\langle v \land \neg d, \{d \mapsto r\}\rangle$
take_right
$e_1{:}\langle r \land v, \{l \mapsto \top\}\rangle$
$e_2{:}\langle \neg r \land v, \{l \mapsto \top\}\rangle$
flick
$f_1{:}\langle v \lor (\neg v \land \neg l), \{v \mapsto \neg v\}\rangle$
$f_2{:}\langle \neg v \land l \land r, \{v \mapsto \neg v\}\rangle$
$f_3{:}\langle \neg v \land l \land \neg r, \{v \mapsto \neg v\}\rangle$
move

Figure 2.3: Event models representing the actions of the thief

thief considers possible prior turning the light on in the vault. ■

In the DEL framework, actions are naturally represented as event models. Due to the internal perspective, these are also taken to be non-pointed. For instance, in a coin toss action, the agent cannot beforehand point out which side will land face up.

Example 2.8. Continuing Example 2.7 we now formalize the actions available to our thieving agent as the event models in Figure 2.3. We use the same conventions for edges as we did for epistemic models. For a basic event e we label it $\langle pre(e), post(e)\rangle$.[2]

The agent is endowed with four actions: take_left, resp. take_right, represent trying to take the diamond from the left, resp. right, pedestal; the diamond is obtained only if it is on the chosen pedestal. Both actions require the agent to be inside the vault and not holding the diamond. flick requires the agent to be inside the vault and turns the light on. Further, it reveals which pedestal the diamond is on. move represents the agent moving in or out of the vault, revealing the location of the diamond provided the light is on.

It can be seen that the epistemic model M' in Example 2.7 is the result of two successive product updates, namely $M_0 \otimes$ move $\otimes$ flick. ■

2.3.2 Applicability, Plans and Solutions

Reasoning about actions from the initial state as in Example 2.8 is exactly what planning is all about. We have however omitted an important component in the reasoning process, one which is crucial. The notion of *applicability* in automated

[2] For a proposition p whose truth value does not change in e we assume the identity mapping $post(e)(p) = p$, as is also the convention in automated planning.

planning dictates when the outcomes of an action are defined. The idea translates to DEL by insisting that no world the planning agent considers possible is eliminated by the product update of an epistemic model with an event model.

Definition 2.9 (Applicability). An event model $\mathcal{E}$ is said to be *applicable* in an epistemic model M if $M \models \langle \mathcal{E} \rangle \top$.

This concept of applicability is easily shown to be equivalent with the one defined in [Bolander and Andersen, 2011] when restricting the latter to the single-agent case. However, for our purposes of describing plans as formulas, we need to express applicability as formulas as well. The discussion in [de Lima, 2007, sect. 6.6] also notes this aspect, insisting that actions must be meaningful. The same sentiment is expressed by our notion of applicability.

The situation in Example 2.7 calls for a way to express conditional plans. Clearly, our agent can only snatch the jewel from the correct pedestal conditioned on how events unfold when she acts. To this end we introduce a language for conditional plans allowing us to handle such contingencies.

Definition 2.10 (Plan Language). Given a finite set A of event models on $\mathcal{L}_{\mathrm{DEL}}(P)$, the *plan language* $\mathcal{L}_{\mathrm{P}}(P, \mathsf{A})$ is given by:

$$\pi ::= \mathcal{E} \mid \mathsf{skip} \mid \mathsf{if}\ K\phi\ \mathsf{then}\ \pi\ \mathsf{else}\ \pi \mid \pi ; \pi$$

where $\mathcal{E} \in \mathsf{A}$ and $\phi \in \mathcal{L}_{\mathrm{DEL}}(P)$. We name members π of this language *plans*, and use $\mathsf{if}\ K\phi\ \mathsf{then}\ \pi$ as shorthand for $\mathsf{if}\ K\phi\ \mathsf{then}\ \pi\ \mathsf{else}\ \mathsf{skip}$.

The reading of the plan constructs are "do $\mathcal{E}$", "do nothing", "if $K\phi$ then π, else π'", and "first π then π'" respectively. Note that the condition of the if-then-else construct is required to be a K-formula. This is to ensure that the planning agent can only make her choices of actions depend on worlds that are distinguishable to her (cf. the discussion of the internal perspective in Section 2.3.1). The idea is similar to the *meaningful plans* of [de Lima, 2007], where branching is only allowed on *epistemically interpretable formulas*.

An alternative way of specifying conditional plans is *policies*, where (in our terminology) each information cell maps to an event model [Ghallab et al., 2004, Sect. 16.2]. There are slight differences between the expressiveness of conditional plans and policies (e.g. policies can finitely represent repetitions); our main motivation for not using policies is that it would require an enumeration of each information cell of the planning domain.

Definition 2.11 (Translation). We define a *strong translation* $[\![\cdot]\!]_s \cdot$ and a *weak translation* $[\![\cdot]\!]_w \cdot$ as functions from $\mathcal{L}_\mathrm{P}(P, \mathsf{A}) \times \mathcal{L}_\mathrm{DEL}(P)$ into $\mathcal{L}_\mathrm{DEL}(P)$ by:

$$\begin{aligned}
&[\![\mathcal{E}]\!]_s\, \phi := \langle \mathcal{E} \rangle \top \wedge [\mathcal{E}] K\phi \\
&[\![\mathcal{E}]\!]_w\, \phi := \langle \mathcal{E} \rangle \top \wedge \widehat{K} \langle \mathcal{E} \rangle K\phi \\
&[\![\mathsf{skip}]\!]_\cdot\, \phi := \phi \\
&[\![\mathsf{if}\ \phi'\ \mathsf{then}\ \pi\ \mathsf{else}\ \pi']\!]_\cdot\, \phi := (\phi' \rightarrow [\![\pi]\!]_\cdot\, \phi) \wedge (\neg\phi' \rightarrow [\![\pi']\!]_\cdot\, \phi) \\
&[\![\pi; \pi']\!]_\cdot\, \phi := [\![\pi]\!]_\cdot\, ([\![\pi']\!]_\cdot\, \phi)
\end{aligned}$$

Plans describe the manner in which actions are carried out. We interpret plans π relative to a formula ϕ and want to answer the question of whether or not π achieves ϕ. Using Definition 2.11 we can answer this question by verifying truth of the DEL formula provided by the translations. This is supported by the results of Section 2.4. We concisely read $[\![\pi]\!]_s\, \phi$ as "π achieves ϕ", and $[\![\pi]\!]_w\, \phi$ as "π may achieve ϕ" (elaborated below). By not specifying separate semantics for plans our framework is kept as simple as possible. Note that applicability (Definition 2.9) is built into the translations through the occurrence of the conjunct $\langle \mathcal{E} \rangle \top$ in both the strong translation $[\![\mathcal{E}]\!]_s\, \phi$ and the weak translation $[\![\mathcal{E}]\!]_w\, \phi$.

The difference between the two translations relate to the *robustness* of plans: $[\![\pi]\!]_s\, \phi$, resp. $[\![\pi]\!]_w\, \phi$, means that every step of π is applicable and that following π always leads, resp. may lead, to a situation where ϕ is known.

Definition 2.12 (Planning Problems and Solutions). Let P be a finite set of propositional symbols. A planning problem on P is a triple $\mathcal{P} = (M_0, \mathsf{A}, \phi_g)$ where

- M_0 is an information cell on $\mathcal{L}_\mathrm{DEL}(P)$ called the *initial state*.
- A is a finite set of event models on $\mathcal{L}_\mathrm{DEL}(P)$ called the *action library*.
- $\phi_g \in \mathcal{L}_\mathrm{DEL}(P)$ is the *goal (formula)*.

We say that a plan $\pi \in \mathcal{L}_\mathrm{P}(P, \mathsf{A})$ is a *strong solution* to $\mathcal{P}$ if $M_0 \models [\![\pi]\!]_s\, \phi_g$, a *weak solution* if $M_0 \models [\![\pi]\!]_w\, \phi_g$ and not a solution otherwise.

Planning problems are defined with the sentiment we've propagated in our examples up until now. The agent is presently in M_0 and wishes ϕ_g to be the case. To this end, she reasons about the actions (event models) in her action library A, creating a conditional plan. Using model checking, she can verify whether this plan is either a weak or strong solution, since plans translate into formulas of $\mathcal{L}_\mathrm{DEL}(P)$. Further, [van Ditmarsch and Kooi, 2008] gives reduction axioms for DEL-formulas, showing that any formula containing the dynamic modality can be expressed as a formula in (basic) epistemic logic. Consequently, plan verification can be seen simply as epistemic reasoning about M_0.

Example 2.13. We continue our running example by discussing it formally as a planning problem and considering the solutions it allows. The initial state is still M_0, and the action library $\mathsf{A} = \{\mathsf{flick}, \mathsf{move}, \mathsf{take\_left}, \mathsf{take\_right}\}$. We discuss the plans below and their merit for our thief.

- $\pi_1 = \mathsf{flick}; \mathsf{move}; \mathsf{if}\ Kr\ \mathsf{then}\ \mathsf{take\_right}\ \mathsf{else}\ \mathsf{take\_left}; \mathsf{move}$
- $\pi_2 = \mathsf{move}; \mathsf{take\_right}; \mathsf{move}$
- $\pi_3 = \mathsf{move}; \mathsf{flick}; \mathsf{take\_right}; \mathsf{move}$
- $\pi_4 = \mathsf{move}; \mathsf{flick}; \mathsf{if}\ Kr\ \mathsf{then}\ \mathsf{take\_right}\ \mathsf{else}\ \mathsf{take\_left}; \mathsf{move}$

We consider two planning problems varying only on the goal formula, $\mathcal{P}_1 = (M_0, \mathsf{A}, d \land \neg v)$ and $\mathcal{P}_2 = (M_0, \mathsf{A}, \widehat{K}d \land \neg v)$. In $\mathcal{P}_1$ her goal is to obtain the diamond and be outside the vault, whereas in $\mathcal{P}_2$ she wishes to be outside the vault *possibly* having obtained the diamond.

Let $\pi_1' = \mathsf{move}; \mathsf{if}\ Kr\ \mathsf{then}\ \mathsf{take\_right}\ \mathsf{else}\ \mathsf{take\_left}; \mathsf{move}$ and note that $\pi_1 = \mathsf{flick}; \pi_1'$. Using the strong translation of π_1, we get $M_0 \models [\![\pi_1]\!]_s \phi_g$ iff $M_0 \models \langle \mathsf{flick} \rangle \top \land [\mathsf{flick}] \left[\!\left[\pi_1'\right]\!\right]_s \phi_g$. As $M_0 \models \langle \mathsf{flick} \rangle \top$ does not hold, π_1 is not a solution. This is expected, since flicking the switch in the initial state is not an applicable action. Verifying that π_2 is a strong solution to $\mathcal{P}_2$ amounts to checking if $M_0 \models [\![\pi_2]\!]_s \widehat{K}d \land \neg v$ which translates to

$$\begin{aligned} &M_0 \models \langle \mathsf{move} \rangle \top \land \\ &\qquad [\mathsf{move}] \Big(\langle \mathsf{take\_right} \rangle \top \land [\mathsf{take\_right}] \Big(\langle \mathsf{move} \rangle \top \land [\mathsf{move}] \Big(\widehat{K}d \land \neg v \Big) \Big) \Big) \end{aligned}$$

With the same approach we can conclude that π_2 is not a solution to $\mathcal{P}_1$, π_3 is a weak solution to $\mathcal{P}_1$ and $\mathcal{P}_2$, and π_4 is a strong solution to $\mathcal{P}_1$ and $\mathcal{P}_2$. ■

2.4 Plan Synthesis

We now show how to synthesise conditional plans for solving planning problems. To synthesise plans, we need a mechanism for coming up with formulas characterising information cells for if-then-else constructs to branch on. Inspired by [Barwise and Moss, 1996, van Benthem, 1998], these are developed in the following. Proofs are omitted, as they are straightforward and similar to proofs in the aforementioned references.

Definition 2.14 (Characterising Formulas). Let $M = (W, \sim, V)$ denote an information cell on $\mathcal{L}_{\mathrm{DEL}}(P)$. We define for all $w \in W$ a formula ϕ_w by: $\phi_w = \bigwedge_{p \in V(w)} p \land \bigwedge_{p \in P - V(w)} \neg p$. We define the *characterising formula* for M, δ_M, as follows: $\delta_M = K(\bigwedge_{w \in W} \widehat{K}\phi_w \land K \bigvee_{w \in W} \phi_w)$.

Lemma 2.15. *Let M be an information cell on $\mathcal{L}_{\text{DEL}}(P)$. Then for all epistemic models $M' = (W', \sim', V')$ and all $w' \in W'$ we have that $(M', w') \models \delta_M$ if and only if there exists a $w \in D(M)$ such that $(M, w) \leftrightarrows (M', w')$.*[3]

For purposes of synthesis, we use the product update solely on non-pointed epistemic and event models. Lemma 2.16 shows that satisfaction of the dynamic modality for non-pointed event models in non-pointed epistemic models relates to the product update in the obvious way.

Lemma 2.16. *Let M be an epistemic model and $\mathcal{E}$ an event model. Then $M \models [\mathcal{E}]\phi$ iff $M \otimes \mathcal{E} \models \phi$.*

Proof. $M \models [\mathcal{E}]\phi \iff$ for all $w \in D(M) : M, w \models [\mathcal{E}]\phi \iff$
for all $w \in D(M) : M, w \models \bigwedge_{e \in D(\mathcal{E})} [\mathcal{E}, e]\phi \iff$
for all $(w, e) \in D(M) \times D(\mathcal{E}) : M, w \models [\mathcal{E}, e]\phi \iff$
for all $(w, e) \in D(M) \times D(\mathcal{E}) : M, w \models pre(e)$ implies $M \otimes \mathcal{E}, (w, e) \models \phi \iff$
for all $(w, e) \in D(M \otimes \mathcal{E}) : M \otimes \mathcal{E}, (w, e) \models \phi \iff M \otimes \mathcal{E} \models \phi$. □

2.4.1 Planning Trees

When synthesising plans, we explicitly construct the search space of the problem as a labelled AND-OR tree, a familiar model for planning under uncertainty [Ghallab et al., 2004]. Our AND-OR trees are called *planning trees*.

Definition 2.17. A *planning tree* is a finite, labelled AND-OR tree in which each node n is labelled by an epistemic model $M(n)$, and each edge (n, m) leaving an OR-node is labelled by an event model $\mathcal{E}(n, m)$.

Planning trees for planning problems $\mathcal{P} = (M_0, \mathsf{A}, \phi_g)$ are constructed as follows. Let the initial planning tree T_0 consist of just one OR-node $root(T_0)$ with $M(root(T_0)) = M_0$ (the root labels the initial state). A planning tree for $\mathcal{P}$ is then any tree that can be constructed from T_0 by repeated applications of the following non-deterministic tree expansion rule.

Definition 2.18 (Tree Expansion Rule). Let T be a planning tree for a planning problem $\mathcal{P} = (M_0, \mathsf{A}, \phi_g)$. The tree expansion rule is defined as follows. Pick an OR-node n in T and an event model $\mathcal{E} \in \mathsf{A}$ applicable in $M(n)$ with the proviso that $\mathcal{E}$ does not label any existing outgoing edges from n. Then:

[3] Here $(M, w) \leftrightarrows (M', w)$ denotes that (M, w) and (M', w) are bisimilar according to the standard notion of bisimulation on pointed epistemic models.

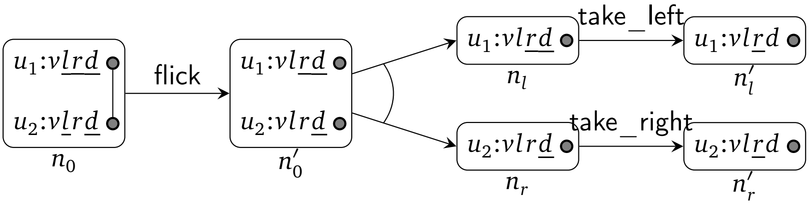

Figure 2.4: Planning tree for a variant of the Pink Panther problem.

1. Add a new node m to T with $M(m) = M(n) \otimes \mathcal{E}$, and add an edge (n, m) with $\mathcal{E}(n, m) = \mathcal{E}$.
2. For each information cell M' in $M(m)$, add an OR-node m' with $M(m') = M'$ and add the edge (m, m').

The tree expansion rule is similar in structure to—and inspired by—the expansion rules used in tableau calculi, e.g. for modal and description logics [Horrocks et al., 2006]. Note that the expansion rule applies only to OR-nodes, and that an applicable event model can only be used once at each node.

Considering single-agent planning a two-player game, a useful analogy for planning trees are game trees. At an OR-node n, the agent gets to pick any applicable action $\mathcal{E}$ it pleases, winning if it ever reaches an epistemic model in which the goal formula holds (see the definition of solved nodes further below). At an AND-node m, the environment responds by picking one of the information cells of $M(m)$—which of the distinguishable outcomes is realised when performing the action.

Example 2.19. In Fig. 2.4 is a planning tree for a variant of the Pink Panther planning problem, this one where the thief is already inside the vault. The root is n_0. Three applications of the tree expansion rule have been made, the labels on edges indicating the chosen action. n_0, n_l and n_r are OR-nodes. n_0', n_l' and n_r' are AND-nodes. The child nodes of the latter two AND-nodes have been omitted, as their information cell is the same as that of their parent nodes. Pay particular attention to how flick reveals the location of the diamond. In the initial state, $M(n_0) \models \neg Kr \wedge \neg K\neg r$, while $M(n_0') \models Kr \vee K\neg r$, $M(n_l) \models K\neg r$ and $M(n_r) \models Kr$. ■

Without restrictions on the tree expansion rule, even very simple planning problems might be infinitely expanded. Finiteness of trees (and therefore termination) is ensured by the following blocking condition.

$\mathcal{B}_1$ The tree expansion rule may not be applied to a node n for which there exists an ancestor node m with $M(m) \leftrightarrows M(n)$.[4]

A planning tree for a planning problem $\mathcal{P}$ is called $\mathcal{B}_1$*-saturated* if no more expansions are possible satisfying condition $\mathcal{B}_1$.

Lemma 2.20 (Termination). *Any procedure that builds a $\mathcal{B}_1$-saturated planning tree for a planning problem $\mathcal{P}$ by repeated application of the tree expansion rule terminates.*

Proof. Planning trees built by repeated application of the tree expansion rule are finitely branching: the action library is finite, and every epistemic model has only finitely many information cells. Furthermore, condition $\mathcal{B}_1$ ensures that no branch has infinite length: there only exists finitely many mutually non-bisimilar epistemic models over any given finite set of propositional symbols [Bolander and Andersen, 2011]. König's Lemma now implies finiteness of the planning tree.

□

Definition 2.21 (Solved Nodes). Let T be any (not necessarily saturated) planning tree for a planning problem $\mathcal{P} = (M_0, \mathsf{A}, \phi_g)$. By recursive definition, a node n in T is called *solved* if one of the following holds:

- $M(n) \models \phi_g$ (the node satisfies the goal formula).
- n is an OR-node having at least one solved child.
- n is an AND-node having all its children solved.

Continuing the game tree analogy, we see that a solved node corresponds is one for which there exists a winning strategy. Regardless of the environment's choice, the agent can achieve its goal. Let T and $\mathcal{P}$ be as above. Below we show that when a node n is solved, it is possible to construct a (strong) solution to the planning problem $(M(n), \mathsf{A}, \phi_g)$. In particular, if the root node is solved, a strong solution to $\mathcal{P}$ can be constructed. As it is never necessary to expand a solved node, nor any of its descendants, we can augment the blocking condition $\mathcal{B}_1$ in the following way.

$\mathcal{B}_2$ The tree expansion rule may not be applied to a node n if one of the following holds: 1) n is solved; 2) n has a solved ancestor; 3) n has an ancestor node m with $M(m) \leftrightarrows M(n)$.

[4] Here $M(m) \leftrightarrows M(n)$ denotes that $M(m)$ and $M(n)$ are bisimilar according to the standard notion of bisimulation between non-pointed epistemic models.

In the following, we will assume that all planning trees have been built according to $\mathcal{B}_2$. One consequence is that a solved OR-node has exactly one solved child. We make use of this in the following definition.

Definition 2.22 (Plans for Solved Nodes). Let T be any planning tree for $\mathcal{P} = (M_0, \mathsf{A}, \phi_g)$. For each solved node n in T, a plan $\pi(n)$ is defined recursively by:

- if $M(n) \models \phi_g$, then $\pi(n) = \mathsf{skip}$.
- if n is an OR-node and m its solved child, then $\pi(n) = \mathcal{E}(n,m); \pi(m)$.
- if n is an AND-node with children $m_1, \ldots, m_k$, then $\pi(n) =$ $\mathsf{if}\ \delta_{M(m_1)}\ \mathsf{then}\ \pi(m_1)\ \mathsf{else\ if}\ \ \delta_{M(m_2)}\ \mathsf{then}\ \pi(m_2)\ \mathsf{else}\ \cdots\ \mathsf{if}\ \delta_{M(m_k)}\ \mathsf{then}\ \pi(m_k)$

Example 2.23. For the goal of achieving the diamond, $\phi_g = d$, we have that the root n_0 of the planning tree of Figure 2.4 is solved, as both n_l' and n_r' satisfy the goal formula. Definition 2.22 gives us

$$\pi(n_0) = \mathsf{flick};\ \mathsf{if}\ \delta_{M(n_l)}\ \mathsf{then\ take\_left; skip\ else\ if}\ \delta_{M(n_r)}\mathsf{then\ take\_right; skip}$$

This plan can easily be shown to be a strong solution to the planning problem of achieving d from the initial state $M(n_0)$. In our soundness result below, we show that plans of solved roots are always strong solutions to their corresponding planing problems. ■

Theorem 2.24 (Soundness). *Let T be a planning tree for a problem $\mathcal{P}$ such that $root(T)$ is solved. Then $\pi(root(T))$ is a strong solution to $\mathcal{P}$.*

Proof. We need to prove that $\pi(root(T))$ is a strong solution to $\mathcal{P}$, that is, $M_0 \models [\![\pi(root(T))]\!]_s\, \phi_g$. Since M_0 is the label of the root, this can be restated as $M(root(T)) \models [\![\pi(root(T))]\!]_s\, \phi_g$. To prove this fact, we will prove the following stronger claim:

- For each solved node n in T, $M(n) \models [\![\pi(n)]\!]_s\, \phi_g$.

We prove this by induction on the height of n. The base case is when n is a leaf. Since n is solved, we must have $M(n) \models \phi_g$. In this case $\pi(n) = \mathsf{skip}$. From $M(n) \models \phi_g$ we can conclude $M(n) \models [\![\mathsf{skip}]\!]_s\, \phi_g$, that is, $M(n) \models [\![\pi(n)]\!]_s\, \phi_g$. This covers the base case. For the induction step, assume that for all solved nodes m of height $< h$, $M(m) \models [\![\pi(m)]\!]_s\, \phi_g$. Let n be an arbitrary solved node n of height h. We then need to show $M(n) \models [\![\pi(n)]\!]_s\, \phi_g$. We have two cases to consider, depending on whether n is an AND- or an OR-node.

Case 1: n is an AND-node. Let $m_1, \ldots, m_k$ be the children of n. By definition, all of these are solved. We have $\pi(n) = \mathsf{if}\ \delta_{M(m_1)}\ \mathsf{then}\ \pi(m_1)\ \mathsf{else\ if}\ \delta_{M(m_2)}\ \mathsf{then}$

$\pi(m_2)$ else $\cdots$ if $\delta_{M(m_k)}$ then $\pi(m_k)$ else skip. The induction hypothesis gives us $M(m_i) \models [\![\pi(m_i)]\!]_s\,\phi_g$ for all $i = 1, \dots, k$.

Claim 1. $M(n) \models \delta_{M(m_i)} \to [\![\pi(m_i)]\!]_s\,\phi_g$, for all $i = 1, \dots, k$.

Proof of claim. Let $w \in D(M(n))$ be chosen arbitrarily. We then need to prove that if $M(n), w \models \delta_{M(m_i)}$ then $M(n), w \models [\![\pi(m_i)]\!]_s\,\phi_g$. Assuming $M(n), w \models \delta_{M(m_i)}$, we get from Lemma 2.15 that there must be a $w' \in D(M(m_i))$ such that $M(m_i), w' \leftrightarrows M(n), w$. Since $M(m_i) \models [\![\pi(m_i)]\!]_s\,\phi_g$, in particular we get $M(m_i), w' \models [\![\pi(m_i)]\!]_s\,\phi_g$, and thus $M(n), w \models [\![\pi(m_i)]\!]_s\,\phi_g$.

Claim 2. $M(n) \models \displaystyle\bigvee_{i=1,\dots,k} \delta_{M(m_i)}$.

Proof of claim. Let $w \in D(M(n))$ be chosen arbitrarily. We then need to prove that $M(n), w \models \vee_{i=1,\dots,k}\delta_{M(m_i)}$. Since $w \in D(M(n))$ it must belong to one of the information cells of $M(n)$, that is, $w \in D(M(m_j))$ for some j. Thus $M(n), w \leftrightarrows M(m_j), w$. From Lemma 2.15 we then get $M(n), w \models \delta_{M(m_j)}$, and thus $M(n), w \models \vee_{i=1,\dots,k}\delta_{M(m_i)}$.

From (1) and (2), we now get:

$$
\begin{aligned}
&M(n) \models \bigwedge_{i=1,\dots,k} (\delta_{M(m_i)} \to [\![\pi(m_i)]\!]_s\,\phi_g) \wedge \bigvee_{i=1,\dots,k} \delta_{M(m_i)} \Rightarrow \\
&M(n) \models \bigwedge_{i=1,\dots,k} (\delta_{M(m_i)} \wedge \bigwedge_{j=1,\dots,i-1} \neg\delta_{M(m_j)} \to [\![\pi(m_i)]\!]_s\,\phi_g) \wedge (\bigwedge_{i=1,\dots,k}\neg\delta_{M(m_i)} \to [\![\mathsf{skip}]\!]_s\,\phi_g) \Rightarrow \\
&M(n) \models (\delta_{M(m_1)} \to [\![\pi(m_1)]\!]_s\,\phi_g) \wedge (\neg\delta_{M(m_1)} \to \\
&\qquad (\delta_{M(m_2)} \to [\![\pi(m_2)]\!]_s\,\phi_g) \wedge (\neg\delta_{M(m_2)} \to \\
&\qquad\quad \cdots \\
&\qquad\qquad (\delta_{M(m_k)} \to [\![\pi(m_k)]\!]_s\,\phi_g) \wedge (\neg\delta_{M(m_k)} \to \\
&\qquad\qquad\qquad [\![\mathsf{skip}]\!]_s\,\phi_g) \cdots) \Rightarrow \\
&M(n) \models [\![\mathsf{if}\ \delta_{M(m_1)}\ \mathsf{then}\ \pi(m_1)\ \mathsf{else} \\
&\qquad \mathsf{if}\ \delta_{M(m_2)}\ \mathsf{then}\ \pi(m_2)\ \mathsf{else} \\
&\qquad\quad \cdots \\
&\qquad\qquad \mathsf{if}\ \delta_{M(m_k)}\ \mathsf{then}\ \pi(m_k)\ \mathsf{else} \\
&\qquad\qquad\quad \mathsf{skip}]\!]\ \phi_g \Rightarrow \\
&M(n) \models [\![\pi(n)]\!]_s\,\phi_g.
\end{aligned}
$$

Case 2: n is an or-node. Here we have $\pi(n) = \mathcal{E}(n, m); \pi(m)$ for the solved child m of n. The induction hypothesis gives $M(m) \models [\![\pi(m)]\!]_s\,\phi_g$, and hence $M(m) \models K\,[\![\pi(m)]\!]_s\,\phi_g$. We now show $M(n) \models [\![\pi(n)]\!]_s\,\phi_g$. Since, by definition, $M(m) = M(n) \otimes \mathcal{E}(n, m)$, we get $M(n) \otimes \mathcal{E}(n, m) \models K\,[\![\pi(m)]\!]_s\,\phi_g$. We can now apply Lemma 2.16 to conclude $M(n) \models [\mathcal{E}(n, m)]\,K\,[\![\pi(m)]\!]_s\,\phi_g$. By definition,

$\mathcal{E}(n,m)$ must be applicable in $M(n)$, that is, $M(n) \models \langle \mathcal{E}(n,m) \rangle \top$. Thus we now have $M(n) \models \langle \mathcal{E}(n,m) \rangle \top \wedge [\mathcal{E}(n,m)] K \llbracket \pi(m) \rrbracket_s \phi_g$. Using Definition 2.11, we can rewrite this as $M(n) \models \llbracket \mathcal{E}(n,m) \rrbracket_s \llbracket \pi(m) \rrbracket_s \phi_g$. Using Definition 2.11 again, we get $M(n) \models \llbracket \mathcal{E}(n,m); \pi(m) \rrbracket_s \phi_g$, and thus finally $M(n) \models \llbracket \pi(n) \rrbracket_s \phi_g$, as required. □

Theorem 2.25 (Completeness). *If there is a strong solution to the planning problem $\mathcal{P} = (M_0, \mathsf{A}, \phi_g)$, then a planning tree T for $\mathcal{P}$ can be constructed, such that $root(T)$ is solved.*

Proof. We first prove the following claim.

Claim 1. If (if ϕ then π_1 else π_2) is a strong solution to $\mathcal{P} = (M_0, \mathsf{A}, \phi_g)$, then so is π_1 or π_2.

Proof of claim. Assume (if ϕ then π_1 else π_2) is a strong solution to $(M_0, \mathsf{A}, \phi_g)$, that is, $M_0 \models \llbracket \text{if } \phi \text{ then } \pi_1 \text{ else } \pi_2 \rrbracket_s \phi_g$. Then, by definition, $M_0 \models (\phi \rightarrow \llbracket \pi_1 \rrbracket_s \phi_g) \wedge (\neg\phi \rightarrow \llbracket \pi_2 \rrbracket_s \phi_g)$. Since M_0 is an information cell, and ϕ is a K-formula, we must have either $M_0 \models \phi$ or $M_0 \models \neg\phi$. Thus we get that either $M_0 \models \llbracket \pi_1 \rrbracket_s \phi_g$ or $M_0 \models \llbracket \pi_2 \rrbracket_s \phi_g$, as required.

Note that we have $\llbracket \text{skip}; \pi \rrbracket_s \phi_g = \llbracket \text{skip} \rrbracket_s (\llbracket \pi \rrbracket_s \phi_g) = \llbracket \pi \rrbracket_s \phi_g$. Thus, we can without loss of generality assume that no plan contains a subexpression of the form skip; π. The length of a plan π, denoted $|\pi|$, is defined recursively by: $|\text{skip}| = 1$; $|\mathcal{E}| = 1$; $|\text{if } \phi \text{ then } \pi_1 \text{ else } \pi_2| = |\pi_1| + |\pi_2|$; $|\pi_1; \pi_2| = |\pi_1| + |\pi_2|$.

Claim 2. Let π be a strong solution to $\mathcal{P} = (M_0, \mathsf{A}, \phi_g)$ with $|\pi| \geq 2$. Then there exists a strong solution of the form $\mathcal{E}; \pi'$ with $|\mathcal{E}; \pi'| \leq |\pi|$.

Proof of claim. Proof by induction on $|\pi|$. The base case is $|\pi| = 2$. We have two cases, $\pi = \text{if } \phi \text{ then } \pi_1 \text{ else } \pi_2$ and $\pi = \pi_1; \pi_2$, both with $|\pi_1| = |\pi_2| = 1$. If π is the latter, it already has desired the form. If $\pi = \text{if } \phi \text{ then } \pi_1 \text{ else } \pi_2$ we have by Claim 1 that either π_1 or π_2 is a strong solution to $\mathcal{P}$. Thus also either π_1; skip or π_2; skip is a strong solution to $\mathcal{P}$, and both of these have length $|\pi|$. This completes the base case. For the induction step, we assume that if π', with $|\pi'| < l$, is a strong solution to a planning problem $\mathcal{P}'$, then there exists is a strong solution of the form $(\mathcal{E}; \pi'')$, with $|\mathcal{E}; \pi''| \leq |\pi'|$. Now consider a plan π of length l which is a strong solution to $\mathcal{P}$. We again have two cases to consider, $\pi = \text{if } \phi \text{ then } \pi_1 \text{ else } \pi_2$ and $\pi = \pi_1; \pi_2$. If $\pi = \pi_1; \pi_2$ is a strong solution to $\mathcal{P}$, then π_1 is a strong solution to the planning problem $\mathcal{P}' = (M_0, \mathsf{A}, \llbracket \pi_2 \rrbracket_s \phi_g)$, as $M_0 \models \llbracket \pi_1; \pi_2 \rrbracket_s \phi_g \Leftrightarrow M_0 \models \llbracket \pi_1 \rrbracket_s \llbracket \pi_2 \rrbracket_s \phi_g$. Clearly $|\pi_1| < l$, so the induction hypothesis gives that there is a strong solution $(\mathcal{E}; \pi_1')$ to $\mathcal{P}'$, with $|\mathcal{E}; \pi_1'| \leq |\pi_1|$. Then, $\mathcal{E}; \pi_1'; \pi_2$ is a strong solution to $\mathcal{P}$ and we have $|\mathcal{E}; \pi_1'; \pi_2| = |\mathcal{E}; \pi_1'| + |\pi_2| \leq |\pi_1| + |\pi_2| = |\pi|$. If $\pi = \text{if } \phi \text{ then } \pi_1 \text{ else } \pi_2$ is a strong solution to $\mathcal{P}$, then we have by Claim 1 that either π_1 or π_2 is a strong solution to $\mathcal{P}$. With both $|\pi_1| < l$ and $|\pi_2| < l$, the induction hypothesis gives the existence a strong solution $\mathcal{E};\pi'$, with $|\mathcal{E};\pi'| \leq |\pi|$. This completes the proof of the claim.

We now prove the theorem by induction on $|\pi|$, where π is a strong solution to $\mathcal{P} = (M_0, \mathsf{A}, \phi_g)$. We need to prove that there exists a planning tree T for $\mathcal{P}$ in which the root is solved. Let T_0 denote the planning tree for $\mathcal{P}$ only consisting of its root node with label M_0. The base case is when $|\pi| = 1$. Here, we have two cases, $\pi = \mathsf{skip}$ and $\pi = \mathcal{E}$. In the first case, the planning tree T_0 already has its root solved, since $M_0 \models [\![\mathsf{skip}]\!]_s\, \phi_g \Leftrightarrow M_0 \models \phi_g$. In the second case $\pi = \mathcal{E}$. Since π is a strong solution to $\mathcal{P}$, we have $M_0 \models [\![\mathcal{E}]\!]_s\, \phi_g$, that is, $M_0 \models \langle\mathcal{E}\rangle \top \wedge [\mathcal{E}]\, K\phi_g$. Thus $\mathcal{E}$ is applicable in M_0 meaning that we can apply the tree expansion rule to T_0, which will produce an AND-node m with $\mathcal{E}(root(T_0), m) = \mathcal{E}$ and $M(m) = M_0 \otimes \mathcal{E}$. Call the expanded tree T_1. Since we have $M_0 \models [\mathcal{E}]K\phi_g$, Lemma 2.16 gives us $M_0 \otimes \mathcal{E} \models K\phi_g$, that is, $M(m) \models K\phi_g$, and hence $M(m) \models \phi_g$. This implies that $M(m)$ and thus $root(T_1)$ is solved. The base case is hereby completed.

For the induction step, assume that a planning tree with solved root can be constructed for problems with strong solutions of length $< l$. Let π be a strong solution to $\mathcal{P}$ with $|\pi| = l$. By Claim 2, there exists a strong solution of the form $\mathcal{E};\pi'$ with $|\mathcal{E};\pi'| \leq |\pi|$. As $M_0 \models [\![\mathcal{E};\pi']\!]_s\, \phi_g \Leftrightarrow M_0 \models [\![\mathcal{E}]\!]_s\, [\![\pi']\!]_s\, \phi_g \Leftrightarrow M_0 \models \langle\mathcal{E}\rangle\top \wedge [\mathcal{E}]K([\![\pi']\!]_s\, \phi_g)$, the tree expansion rule can be applied by picking $\mathcal{E}$ and M_0. This produces the AND-node m with $\mathcal{E}(n, m) = \mathcal{E}$ and $M(m) = M_0 \otimes \mathcal{E}$. $m_1, \ldots, m_k$ are the children of m, and $M(m_i) = M_i$ the information cells in $M(m)$. From $M_0 \models [\mathcal{E}]K([\![\pi']\!]_s\, \phi_g)$ we get $M_0 \otimes \mathcal{E} \models K\, [\![\pi']\!]_s\, \phi_g$, using Lemma 2.16. This implies $M_i \models K\, [\![\pi']\!]_s\, \phi_g$, and hence $M_i \models [\![\pi']\!]_s\, \phi_g$, for each information cell M_i of $M(m) = M_0 \otimes \mathcal{E}$. Thus π' must be a strong solution to each of the planning problems $\mathcal{P}_i = (M_i, \mathcal{A}, \phi_g)$. As $|\pi'| < |\mathcal{E};\pi'| \leq l$, the induction hypothesis gives that planning trees T_i with solved roots can be constructed for each $\mathcal{P}_i$. Let T denote T_0 expanded with $m, m_1, \ldots, m_k$, and each T_i be the subtree rooted at m_i. Then each of the nodes m_i are solved in T, and in turn both m and $root(T)$ are solved. □

2.4.2 Strong Planning Algorithm

With all the previous in place, we now have an algorithm for synthesising strong solutions for planning problems $\mathcal{P}$, given as follows.

```
StrongPlan($\mathcal{P}$)
Let $T$ be the plan. tree only consisting of $root(T)$ labelled by the init. state of $\mathcal{P}$.
Repeatedly apply the tree expansion rule of $\mathcal{P}$ to $T$ until it is $\mathcal{B}_2$-saturated.
If $root(T)$ is solved, return $\pi(root(T))$, otherwise return FAIL.
```

Theorem 2.26. StrongPlan($\mathcal{P}$) *is a terminating, sound and complete algorithm for producing strong solutions to planning problems. Soundness means that if* StrongPlan($\mathcal{P}$) *returns a plan, it is a strong solution to $\mathcal{P}$. Completeness means that if $\mathcal{P}$ has a strong solution,* StrongPlan($\mathcal{P}$) *will return one.*

Proof. Termination comes from Lemma 2.20 (with $\mathcal{B}_1$ replaced by the stronger condition $\mathcal{B}_2$), soundness from Theorem 2.24 and completeness from Theorem 2.25 (given any two saturated planning trees T_1 and T_2 for the same planning problem, the root node of T_1 is solved iff the root node of T_2 is). □

2.4.3 Weak Planning Algorithm

With few changes, the machinery already in place gives an algorithm for synthesising weak solutions. Rather than requiring all children of an AND-node be solved, we require only one. This corresponds to the notion of weak, defined in Definition 2.11. Only one possible execution need lead to the goal.

Definition 2.27 (Weakly Solved Nodes)**.** A node n is called *weakly solved* if either $M(n) \models \phi_g$ or n has at least one weakly solved child.

We keep the tree expansion rule, but make use of a new blocking condition $\mathcal{B}_3$ using Definition 2.27 rather than Definition 2.21.

Definition 2.28 (Plans for Weakly Solved Nodes)**.** Let T be any planning tree for $\mathcal{P} = (M_0, \mathsf{A}, \phi_g)$. For each weakly solved node n in T, a plan $\pi_w(n)$ is defined recursively by:

- if $M(n) \models \phi_g$, then $\pi_w(n) = \mathsf{skip}$
- if n is an OR-node and m its weakly solved child, then $\pi_w(n) = \mathcal{E}(n,m); \pi_w(m)$
- if n is an AND-node and m its weakly solved child, then $\pi_w(n) = \pi_w(m)$

The algorithm for weak planning is defined as follows.

```
WeakPlan($\mathcal{P}$)
Let $T$ be the plan. tree only consisting of $root(T)$ labelled by the init. state of $\mathcal{P}$.
Repeatedly apply the tree expansion rule of $\mathcal{P}$ to $T$ until it is $\mathcal{B}_3$-saturated.
If $root(T)$ is weakly solved, return $\pi_w(root(T))$, otherwise return fail.
```

Theorem 2.29. WeakPlan($\mathcal{P}$) *is a terminating, sound and complete algorithm for producing weak solutions to planning problems.*

2.5 Related and Future Work

In this paper, we have presented a syntactic characterisation of weak and strong solutions to epistemic planning problems, that is, we have characterised solutions as formulas. [Bolander and Andersen, 2011] takes a semantic approach to strong solutions for epistemic planning problems. In their work plans are sequences of actions, requiring conditional choice of actions at different states to be encoded in the action structure itself. We represent choice explicitly, using a language of conditional plans. An alternative to our approach of translating conditional plans into formulas of DEL would be to translate plans directly into (complex) event models. This is the approach taken in [Baltag and Moss, 2004], where they have a language of epistemic programs similar to our language of plans (modulo the omission of ontic actions). Using this approach in a planning setting, one could translate each possible plan π into the corresponding event model $\mathcal{E}(\pi)$, check its applicability, and check whether $M_0 \otimes \mathcal{E}(\pi) \models \phi_g$ (the goal is satisfied in the product update of the initial state with the event model). However, even for a finite action library, there are infinitely many distinct plans, and thus infinitely many induced event models to consider when searching for a solution. To construct a terminating planning algorithm with this approach, one would still have to limit the plans considered (e.g. by using characterising formulas), and also develop a more involved loop-checking mechanism working at the level of plans. Furthermore, our approach more obviously generalises to algorithms for replanning, which is current work.

The meaningful plans of [de Lima, 2007, chap. 2] are reminiscent of the work in this paper. Therein, plan verification is cast as validity of an EDL-consequence in a given system description. Like us, they consider single-agent scenarios, conditional plans, applicability and incomplete knowledge in the initial state. Unlike us, they consider only deterministic actions. In the multi-agent treatment [de Lima, 2007, chap. 4], action laws are translated to a fragment of DEL with only public announcements and public assignments, making actions singleton event models. This means foregoing nondeterminism and therefore sensing actions.

Planning problems in [Löwe et al., 2011a] are solved by producing a sequence of pointed event models where an external variant of applicability (called *possible at*) is used. Using such a formulation means outcomes of actions are fully determined, making conditional plans and weak solutions superfluous. As noted by the authors, and unlike our framework, their approach does not consider factual change. We stress that [Bolander and Andersen, 2011, Löwe et al., 2011a, de Lima, 2007] all consider the multi-agent setting which we have not treated here.

In our work so far, we haven't treated the problem of where domain formulations come from, assuming just that they are given. Standardised description languages are vital if modal logic-based planning is to gain wide acceptance in the planning community. Recent work worth noting in this area includes [Baral et al., 2012], which presents a specification language for the multi-agent belief case.

As suggested by our construction of planning trees, there are several connections between our approach and two-player imperfect information games. First, product updates imply perfect recall [van Benthem, 2001]. Second, when the game is at a node belonging to an information set, the agent knows a proposition only if it holds throughout the information set; corresponding to our use of information cells. Finally, the strong solutions we synthesise are very similar to mixed strategies. A strong solution caters to any information cell (contingency) it may bring about, by selecting exactly one sub-plan for each [Aumann and Hart, 1992].

Our work naturally relates to [Ghallab et al., 2004], where the notions of strong and weak solutions are found. Their belief states are sets of states which may be partioned by observation variables. Our partition of epistemic models into information cells follows straight from the definition of product update. A clear advantage in our approach is that actions encode both nondetermism and partial observability. [Rintanen, 2004] shows that for conditional planning (prompted by nondeterministic actions) in partially observable domains the *plan existence problem* is 2-EXP-complete (plans must succeed with probability 1; i.e. be strong solutions). StrongPlan($\mathcal{P}$) implicitly answers the same question for $\mathcal{P}$ (it gives a strong solution if one exists). Reductions between the two decision problem variants would give a complexity measure of our approach, and also formally link conditional epistemic planning with the approaches used in automated planning.

We would like to do plan verification and synthesis in the multi-agent settings. We believe that generalising the notions introduced in this paper to multi-pointed epistemic and event models are key. Plan synthesis in the multi-agent setting is undecidable [Bolander and Andersen, 2011], but considering restricted classes of actions as is done in [Löwe et al., 2011a] seems a viable route for achieving decidable multi-agent planning. Another interesting area is to consider modalities such as plausibility and preferences. This would allow an agent to plan for (perhaps only) the most likely outcomes of its own actions and the preferred actions taken by other agents in the system. This could then be combined with the possibility of doing replanning, as mentioned above.

Chapter 3

Don’t Plan for the Unexpected: Planning Based on Plausibility Models

This chapter is a replicate of [Andersen et al., 2014], accepted for publication in a special issue of Logique et Analyses devoted to the theme “Dynamics in Logic”.

Don’t Plan for the Unexpected: Planning Based on Plausibility Models

Mikkel Birkegaard Andersen Thomas Bolander
Martin Holm Jensen
DTU Compute, Technical University of Denmark

Abstract

We present a framework for automated planning based on plausibility models, as well as algorithms for computing plans in this framework. Our plausibility models include postconditions, as ontic effects are essential for most planning purposes. The framework presented extends a previously developed framework based on dynamic epistemic logic (DEL), without plausibilities/beliefs. In the pure epistemic framework, one can distinguish between strong and weak epistemic plans for achieving some, possibly epistemic, goal. By taking all possible outcomes of actions into account, a strong plan guarantees that the agent achieves this goal. Conversely, a weak plan promises only the possibility of leading to the goal. In real-life planning scenarios where the planning agent is faced with a high degree of uncertainty and an almost endless number of possible exogenous events, strong epistemic planning is not computationally feasible. Weak epistemic planning is not satisfactory either, as there is no way to qualify which of two weak plans is more likely to lead to the goal. This seriously limits the practical uses of weak planning, as the planning agent might for instance always choose a plan that relies on serendipity. In the present paper we introduce a planning framework with the potential of overcoming the problems of both weak and strong epistemic planning. This framework is based on plausibility models, allowing us to define different types of plausibility planning. The simplest type of plausibility plan is one in

which the goal will be achieved when all actions in the plan turn out to have the outcomes found most plausible by the agent. This covers many cases of everyday planning by human agents, where we—to limit our computational efforts—only plan for the most plausible outcomes of our actions.

3.1 Introduction

Whenever an agent deliberates about the future with the purpose of achieving a goal, she is engaging in the act of planning. Automated Planning is a widely studied area of AI dealing with such issues under many different assumptions and restrictions. In this paper we consider *planning under uncertainty* [Ghallab et al., 2004] (nondeterminism and partial observability), where the agent has knowledge and beliefs about the environment and how her actions affect it. We formulate scenarios using plausibility models obtained by merging the frameworks in [Baltag and Smets, 2006, van Ditmarsch and Kooi, 2008].

Example 3.1 (The Basement)**.** An agent is standing at the top of an unlit stairwell leading into her basement. If she walks down the steps in the dark, it’s likely that she will trip. On the other hand, if the lights are on, she is certain to descend unharmed. There is a light switch just next to her, though she doesn’t know whether the bulb is broken.

She wishes to find a plan that gets her safely to the bottom of the stairs. Planning in this scenario is contingent on the situation; e.g. is the bulb broken? Will she trip when attempting her descent? In planning terminology a plan that *might* achieve the goal is a *weak solution*, whereas one that *guarantees* it is a *strong solution*.

In this case, a weak solution is to simply descend the stairs in the dark, risking life and limb for a trip to the basement. On the other hand, there is no strong solution as the bulb might be broken (assuming it cannot be replaced). Intuitively, the *best* plan is to flick the switch (expecting the bulb to work) and then descend unharmed, something neither weak nor strong planning captures. ■

Extending the approach in [Andersen et al., 2012] to a logical framework incorporating beliefs via a plausibility ordering, we formalise plans which an agent considers most likely to achieve her goals. This notion is incorporated into algorithms developed for the framework in [Andersen et al., 2012], allowing us to synthesise plans like the *best* one in Example 3.1.

In the following section we present the logical framework we consider throughout

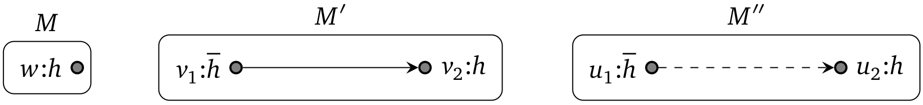

Figure 3.1: Three plausibility models.

the paper. Section 3.3 formalises planning in this framework, and introduces the novel concept of plausibility solutions to planning problems. As planning is concerned with representing possible ways in which the future can unfold, it turns out we need a belief modality corresponding to a globally connected plausibility ordering, raising some technical challenges. Section 3.4 introduces an algorithm for plan synthesis (i.e. generation of plans). Further we show that the algorithm is terminating, sound and complete. To prove termination, we must define bisimulations and bisimulation contractions.

3.2 Dynamic Logic of Doxastic Ontic Actions

The framework we need for planning is based on a dynamic logic of doxastic ontic actions. Actions can be epistemic (changing knowledge), doxastic (changing beliefs), ontic (changing facts) or any combination. The following formalisation builds on the *dynamic logic of doxastic actions* [Baltag and Smets, 2006], adding postconditions to event models as in [van Ditmarsch and Kooi, 2008]. We consider only the single-agent case. Before the formal definitions are given, we present some intuition behind the framework in the following example, which requires some familiarity with epistemic logic.

Example 3.2. Consider an agent and a coin biased towards heads, with the coin lying on a table showing heads (h). She contemplates tossing the coin and realizes that it can land either face up, but (due to nature of the coin) believes it will land heads up. In either case, after the toss she knows exactly which face is showing.

The initial situation is represented by the *plausibility model* (defined later) M and the contemplation by M'' (see Figure 3.1). The two worlds u_1, u_2 are epistemically distinguishable ($u_1 \not\sim u_2$) and represent the observable non-deterministic outcome of the toss. The *dashed* directed edge signifies a *(global) plausibility relation*, where the direction indicates that she finds u_2 more plausible than u_1 (we overline proposition symbols that are false). ■

Example 3.3. Consider again the agent and biased coin. She now reasons about shuffling the coin under a dice cup, leaving the dice cup on top to conceal the coin. She cannot observe which face is up, but due to the bias of the coin believes it to be heads. She then reasons further about lifting the dice cup in this situation, and realises that she will observe which face is showing. Due to her beliefs about the shuffle she finds it most plausible that heads is observed.

The initial situation is again M. Consider the model M', where the *solid* directed edge indicates a *local plausibility relation*, and the direction that v_2 is believed over v_1. By *local* we mean that the two worlds v_1, v_2 are *(epistemically) indistinguishable* ($v_1 \sim v_2$), implying that she is ignorant about whether h or $\neg h$ is the case.[1] Together this represents the concealed, biased coin. Her contemplations on lifting the cup is represented by the model M'' as in the previous example. ■

In Example 3.2 the agent reasons about a non-deterministic action whose outcomes are *distinguishable* but not equally plausible, which is different from the initial contemplation in Example 3.3 where the outcomes are *not* distinguishable (due to the dice cup). In Example 3 she subsequently reasons about the observations made after a sensing action. In both examples she reasons about the future, and in both cases the final result is the model M''. In Example 3.8 we formally elaborate on the actions used here.

It is the nature of the agent's ignorance that make M' and M'' two inherently different situations. Whereas in the former she is ignorant about h due to the coin being concealed, her ignorance in the latter stems from not having lifted the cup yet. In general we can model ignorance either as a consequence of epistemic indistinguishability, or as a result of not yet having acted. Neither type subsumes the other and both are necessary for reasoning about actions. We capture this distinction by defining both local and global plausibility relations. The end result is that local plausibility talks about belief in a particular epistemic equivalence class, and global plausibility talks about belief in the entire model. We now remedy the informality we allowed ourselves so far by introducing the necessary definitions for a more formal treatment.

Definition 3.4 (Dynamic Language)**.** Let a countable set of propositional symbols P be given. The language $L(P)$ is given by the following BNF:

$$\phi ::= p \mid \neg\phi \mid \phi \wedge \phi \mid K\phi \mid B^{\phi}\phi \mid X\phi \mid [\mathcal{E}, e]\,\phi$$

where $p \in P$, $\mathcal{E}$ is an *event model* on $L(P)$ as (simultaneously) defined below, and $e \in D(\mathcal{E})$. K is the *local* knowledge modality, B^{ϕ} the *global* conditional belief modality, X is a (non-standard) *localisation* modality (explained later) and $[\mathcal{E}, e]$ the dynamic modality.

[1] In the remainder, we use (in)distinguishability without qualification to refer to epistemic (in)distinguishability.

We use the usual abbreviations for the other boolean connectives, as well as for the dual dynamic modality $\langle \mathcal{E}, e \rangle \, \phi := \neg \, [\mathcal{E}, e] \, \neg\phi$ and unconditional (or absolute) global belief $B\phi := B^{\top}\phi$. The duals of K and B^{ϕ} are denoted $\widehat{K}$ and $\widehat{B}^{\phi}$.

$K\phi$ reads as "the (planning) agent knows ϕ", $B^{\psi}\phi$ as "conditional on ψ, the (planning) agent believes ϕ", and $[\mathcal{E}, e] \, \phi$ as "after all possible executions of $(\mathcal{E}, e)$, ϕ holds". $X\phi$ reads as "locally ϕ".

Definition 3.5 (Plausibility Models)**.** A *plausibility model* on a set of propositions P is a tuple $M = (W, \sim, \leq, V)$, where

- W is a set of *worlds*,
- $\sim \subseteq W \times W$ is an equivalence relation called the *epistemic relation*,
- $\leq \subseteq W \times W$ is a connected well-preorder called the *plausibility relation*,[2]
- $V : P \rightarrow 2^W$ is a *valuation*.

$D(M) = W$ denotes the *domain* of M. For $w \in W$ we name (M, w) a *pointed plausibility model*, and refer to w as the *actual world* of (M, w). $<$ denotes the *strict plausibility relation*, that is $w < w'$ iff $w \leq w'$ and $w' \not\leq w$. $\simeq$ denotes *equiplausibility*, that is $w \simeq w'$ iff $w \leq w'$ and $w' \leq w$.

In our model illustrations a directed edge from w to w' indicates $w' \leq w$. By extension, strict plausibility is implied by unidirected edges and equiplausibility by bidirected edges. For the models in Figure 3.1, we have $v_1 \sim v_2$, $v_2 < v_1$ in M' and $u_1 \not\sim u_2$, $u_2 < u_1$ in M''. The difference between these two models is in the epistemic relation, and is what gives rise to local (solid edges) and global (dashed edges) plausibility. In [Baltag and Smets, 2006] the local plausibility relation is defined as $\trianglelefteq := \sim \cap \leq$; i.e. $w \trianglelefteq w'$ iff $w \sim w'$ and $w \leq w'$. $\trianglelefteq$ is a *locally well-preordered relation*, meaning that it is a union of *mutually disjoint well-preorders*. Given a plausibility model, the domain of each element in this union corresponds to an $\sim$-equivalence class.

Our distinction between local and global is not unprecedented in the literature, but it can be a source of confusion. In [Baltag and Smets, 2006], $\leq$ was indeed connected (i.e. global), but in later versions of the framework [Baltag and Smets, 2008b] this was no longer required. The iterative development in [van Ditmarsch, 2005] also discuss the distinction between local and global plausibility (named preference by the author). Relating the notions to the wording in [Baltag and

[2] A well-preorder is a reflexive, transitive binary relation s.t. every non-empty subset has minimal elements [Baltag and Smets, 2008b].

Smets, 2006], $\leq$ captures *a priori* beliefs about *virtual* situations, *before* obtaining any direct information about the actual situation. On the other hand, $\trianglelefteq$ captures *a posteriori* beliefs about an *actual* situation, that is, the agent's beliefs *after* she obtains (or assumes) information about the actual world.

M'' represents two distinguishable situations (v_1 and v_2) that are a result of reasoning about the future, with v_2 being considered more plausible than v_1. These situations are identified by restricting M'' to its $\sim$-equivalence classes; i.e. $M'' \restriction \{v_1\}$ and $M'' \restriction \{v_2\}$. Formally, given an epistemic model M, the *information cells* in M are the submodels of the form $M \restriction [w]_\sim$ where $w \in D(M)$. We overload the term and name any $\sim$-connected plausibility model on P an *information cell*. This use is slightly different from the notion in [Baltag and Smets, 2008b], where an information cell is an $\sim$-equivalence class rather than a restricted model. An immediate property of information cells is that $\leq=\trianglelefteq$; i.e. the local and global plausibility relations are identical. A partition of a plausibility model into its information cells corresponds to a *localisation* of the plausibility model, where each information cell represents a local situation. The (later defined) semantics of X enables reasoning about such localisations using formulas in the dynamic language.

Definition 3.6 (Event Models)**.** An *event model* on the language $L(P)$ is a tuple $\mathcal{E} = (E, \sim, \leq, pre, post)$, where

- E is a finite set of *(basic) events*,
- $\sim \subseteq E \times E$ is an equivalence relation called the *epistemic relation*,
- $\leq \subseteq E \times E$ is a connected well-preorder called the *plausibility relation*,
- $pre : E \to L(P)$ assigns to each event a *precondition*,
- $post : E \to (P \to L(P))$ assigns to each event a *postcondition* for each proposition. Each $post(e)$ is required to be only finitely different from the identity.

$D(\mathcal{E}) = E$ denotes the *domain* of $\mathcal{E}$. For $e \in E$ we name $(\mathcal{E}, e)$ a *pointed event model*, and refer to e as the *actual event* of $(\mathcal{E}, e)$. We use the same conventions for accessibility relations as in the case of plausibility models.

Definition 3.7 (Product Update)**.** Let $M = (W, \sim, \leq, V)$ and $\mathcal{E} = (E, \sim', \leq', pre, post)$ be a plausibility model on P resp. event model on $L(P)$. The *product update* of M with $\mathcal{E}$ is the plausibility model denoted $M \otimes \mathcal{E} = (W', \sim'', \leq'', V')$, where

- $W' = \{(w,e) \in W \times E \mid M, w \models pre(e)\}$,
- $\sim'' = \{((w,e),(v,f)) \in W' \times W' \mid w \sim v \text{ and } e \sim' f\}$,
- $\leq'' = \{((w,e),(v,f)) \in W' \times W' \mid e <' f \text{ or } (e \simeq' f \text{ and } w \leq v)\}$,
- $V'(p) = \{(w,e) \in W' \mid M, w \models post(e)(p)\}$ for each $p \in P$.

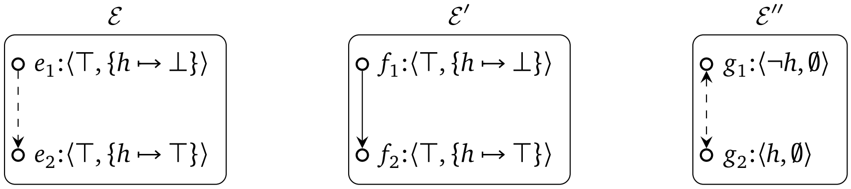

Figure 3.2: Three event models.

The reader may consult [Baltag and Moss, 2004, Baltag and Smets, 2006, Baltag and Smets, 2008b, van Ditmarsch and Kooi, 2008] for thorough motivations and explanations of the product update. Note that the event model's plausibilities take priority over those of the plausibility model (action-priority update).

Example 3.8. Consider Figure 3.2, where the event model $\mathcal{E}$ represents the biased non-deterministic coin toss of Example 3.2, $\mathcal{E}'$ shuffling the coin under a dice cup, and $\mathcal{E}''$ lifting the dice cup of Example 3.3. We indicate $\sim$ and $\leq$ with edges as in our illustrations of plausibility models. Further we use the convention of labelling basic events e by $\langle pre(e), post(e) \rangle$. We write $post(e)$ on the form $\{p_1 \mapsto \phi_1, \ldots, p_n \mapsto \phi_n\}$, meaning that $post(e)(p_i) = \phi_i$ for all i, and $post(e)(q) = q$ for $q \notin \{p_1, \ldots, p_n\}$.

Returning to Example 3.2 we see that $M \otimes \mathcal{E} = M''$ where $u_1 = (w, e_1), u_2 = (w, e_2)$. In $\mathcal{E}$ we have that $e_2 < e_1$, which encodes the bias of the coin, and $e_1 \not\sim e_2$ encoding the observability, which leads to u_1 and u_2 being distinguishable.

Regarding Example 3.3 we have that $M \otimes \mathcal{E}' = M'$ (modulo renaming). In contrast to $\mathcal{E}$, we have that $f_1 \sim f_2$, representing the inability to see the face of the coin due to the dice cup. For the sensing action $\mathcal{E}''$, we have $M \otimes \mathcal{E}' \otimes \mathcal{E}'' = M''$, illustrating how, when events are equiplausible ($g_1 \simeq g_2$), the plausibilities of M' carry over to M''. ■

We've shown examples of how the interplay between plausibility model and event model can encode changes in belief, and further how to model both ontic change and sensing. In [Bolander and Andersen, 2011] there is a more general treatment of action types, but here such a classification is not our objective. Instead we simply encode actions as required for our exposition and leave these considerations as future work.

Among the possible worlds, $\leq$ gives an ordering defining what is believed. Given a plausibility model $M = (W, \sim, \leq, V)$, any non-empty subset of W will have one or more minimal worlds with respect to $\leq$, since $\leq$ is a well-preorder. For $S \subseteq W$, the set of $\leq$-minimal worlds, denoted $Min_{\leq}S$, is defined as:

$$Min_{\leq}S = \{s \in S \mid \forall s' \in S : s \leq s'\}.$$

The worlds in $Min_{\leq}S$ are called the *most plausible* worlds in S. The worlds of $Min_{\leq}D(M)$ are referred to as the *most plausible* of M. With belief defined via minimal worlds (see the definition below), the agent has the same beliefs for any $w \in D(M)$. Analogous to most plausible worlds, an information cell M' of M is called *most plausible* if $D(M') \cap Min_{\leq}D(M) \neq \emptyset$ (M' contains at least one of the most plausible worlds of M).

Definition 3.9 (Satisfaction Relation). Let a plausibility model $M = (W, \sim, \leq, V)$ on P be given. The satisfaction relation is given by, for all $w \in W$:

$$\begin{array}{ll}
M, w \models p & \text{iff } w \in V(p) \\
M, w \models \neg\phi & \text{iff } \mathit{not}\ M, w \models \phi \\
M, w \models \phi \wedge \psi & \text{iff } M, w \models \phi \text{ and } M, w \models \psi \\
M, w \models K\phi & \text{iff } M, v \models \phi \text{ for all } w \sim v \\
M, w \models B^{\psi}\phi & \text{iff } M, v \models \phi \text{ for all } v \in Min_{\leq}\{u \in W \mid M, u \models \psi\} \\
M, w \models X\phi & \text{iff } M \upharpoonright [w]_{\sim}, w \models \phi \\
M, w \models [\mathcal{E}, e]\,\phi & \text{iff } M, w \models pre(e) \text{ implies } M \otimes \mathcal{E}, (w, e) \models \phi
\end{array}$$

where $\phi, \psi \in L(P)$ and $(\mathcal{E}, e)$ is a pointed event model. We write $M \models \phi$ to mean $M, w \models \phi$ for all $w \in D(M)$. Satisfaction of the dynamic modality for non-pointed event models $\mathcal{E}$ is introduced by abbreviation, viz. $[\mathcal{E}]\,\phi := \bigwedge_{e \in D(\mathcal{E})} [\mathcal{E}, e]\,\phi$. Furthermore, $\langle\mathcal{E}\rangle\,\phi := \neg\,[\mathcal{E}]\,\neg\phi$.[3]

The reader may notice that the semantic clause for $M, w \models X\phi$ is equivalent to the clause for $M, w \models [\mathcal{E}, e]\phi$ when $[\mathcal{E}, e]$ is a public announcement of a *characteristic formula* [van Benthem, 1998] being true exactly at the worlds in $[w]_{\sim}$ (and any other world modally equivalent to one of these). In this sense, the X operator can be thought of as a public announcement operator, but a special one that always announces the current information cell. In the special case where M is an information cell, we have for all $w \in D(M)$ that $M, w \models X\phi$ iff $M, w \models \phi$.

3.3 Plausibility Planning

The previous covered a framework for dealing with knowledge and belief in a dynamic setting. In the following, we will detail how a rational agent would adapt

[3] Hence, $M, w \models \langle\mathcal{E}\rangle\,\phi \Leftrightarrow M, w \models \neg\,[\mathcal{E}]\,\neg\phi \Leftrightarrow M, w \models \neg(\bigwedge_{e \in D(\mathcal{E})} [\mathcal{E}, e]\,\neg\phi) \Leftrightarrow M, w \models \bigvee_{e \in D(\mathcal{E})} \neg\,[\mathcal{E}, e]\,\neg\phi \Leftrightarrow M, w \models \bigvee_{e \in D(\mathcal{E})} \langle\mathcal{E}, e\rangle\,\phi$.

these concepts to model her own reasoning about how her actions affect the future. Specifically, we will show how an agent can predict whether or not a particular plan leads to a desired goal. This requires reasoning about the *conceivable* consequences of actions without *actually* performing them.

Two main concepts are required for our formulation of planning, both of which build on notions from the logic introduced in the previous section. One is that of states, a representation of the planning agent's view of the world at a particular time. Our states are plausibility models. The other concept is that of actions. These represent the agent's view of everything that can happen when she does something. Actions are event models, changing states into other states via product update.

In our case, the agent has knowledge and beliefs about the initial situation, knowledge and beliefs about actions, and therefore also knowledge and beliefs about the result of actions.

All of what follows regards planning in the internal perspective. Section 3.3.1 shows how plausibility models represent states, Section 3.3.2 how event models represent actions and Section 3.3.3 how these ideas can formalise planning problems with various kinds of solutions.

3.3.1 The Internal Perspective On States

In the internal perspective, an agent using plausibility models to represent her own view will, generally, not be able to point out the actual world. Consider again the model M in Figure 3.1, that has two indistinguishable worlds w_1 and w_2. If M is the agent's view of the situation, she will of course not be able to say which is the actual world. If she was, then the model *could not represent the situation where the two worlds are indistinguishable*. By requiring the agent to reason from non-pointed plausibility models only (a similar argument makes the case for non-pointed event models), we enforce the internal perspective.

3.3.2 Reasoning About Actions

Example 3.10 (Friday Beer). Nearing the end of the month, an agent is going to have an end-of-week beer with her coworkers. Wanting to save the cash she has on hand for the bus fare, she would like to buy the beer using her debit card. Though she isn't certain, she believes that there's no money ($\overline{m}$) on the associated account. Figure 3.3 shows this initial situation as M, where $\overline{t}$ signifies that the transaction hasn't been completed. In this small example her goal is to make t true.

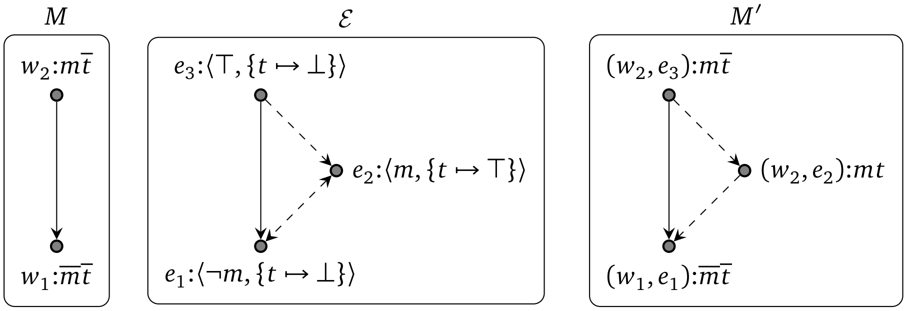
M
$w_2{:}m\bar{t}$
$w_1{:}\bar{m}\bar{t}$
$\mathcal{E}$
$e_3{:}\langle\top,\{t\mapsto\bot\}\rangle$
$e_2{:}\langle m,\{t\mapsto\top\}\rangle$
$e_1{:}\langle\neg m,\{t\mapsto\bot\}\rangle$
M'
$(w_2,e_3){:}m\bar{t}$
$(w_2,e_2){:}mt$
$(w_1,e_1){:}\bar{m}\bar{t}$

Figure 3.3: The situation before and after attempting to pay with a debit card, plus the event model depicting the attempt. This illustrates that the most plausible information cell can contain the least plausible world.

When attempting to complete the transaction (using a normal debit card reader), a number of different things can happen, captured by $\mathcal{E}$ in Figure 3.3. If there is money on the account, the transaction will go through (e_2), and if there isn't, it won't (e_1). This is how the card reader operates most of the time and why e_1 and e_2 are the most plausible events. Less plausible, but still possible, is that the reader malfunctions for some other reason (e_3). The only feedback the agent will receive is whether the transaction was completed, not the reasons why it did or didn't ($e_1 \sim e_3 \not\sim e_2$). That the agent finds out whether the transaction was successful is why we do not collapse e_1 and e_2 to one event e' with $pre(e') = \top$ and $post(e')(t) = m$.

$M \otimes \mathcal{E}$ expresses the agent's view on the possible outcomes of attempting the transaction. The model M' is the bisimulation contraction of $M \otimes \mathcal{E}$, according to the definition in Section 3.4.1 (the world (w_1, e_3) having been removed, as it is bisimilar to (w_1, e_1)).

M' consists of two information cells, corresponding to whether or not the transaction was successful. What she believes will happen is given by the global plausibility relation. When actually attempting the transaction the result will be one of the information cells of M', namely $M_{\bar{t}} = M' \upharpoonright \{(w_1, e_1), (w_2, e_3)\}$ or $M_t = M' \upharpoonright \{(w_2, e_2)\}$, in which she will know $\neg t$ and t respectively. As (w_1, e_1) is the most plausible, we can say that she *expects* to end up in (w_1, e_1), and, by extension, in the information cell $M_{\bar{t}}$: She expects to end up in a situation where she knows $\neg t$, but is ignorant concerning m. If, unexpectedly, the transaction *is* successful, she will know that the balance is sufficient (m). The most plausible information cell(s) in a model are those the agent expects. That (w_2, e_3) is in the expected information cell, when the globally more plausible world (w_2, e_2) is not, might seem odd. It isn't. The partitioning of M into the information cells $M_{\bar{t}}$ and M_t suggests that she will sense the value of t ($\neg t$ holds everywhere in the former, t everywhere in the latter). As she expects to find out that t does not to hold, she expects to be able to rule out all the worlds in which t *does* hold. Therefore, she expects to be able to rule out (w_2, e_2) and *not* (w_2, e_3) (or w_1, e_1). This gives $M' \models BX(K\neg t \wedge B\neg m \wedge \widehat{K}m)$: She expects to come to know that the transaction has failed and that she will believe there's no money on the account (though she does consider it possible that there is). ■

Under the definition of planning that is to follow in Section 3.3.3, an agent has a number of actions available to construct plans. She needs a notion of which actions can be considered at different stages of the planning process. As in the planning literature, we call this notion *applicability*.

Definition 3.11 (Applicability). An event model $\mathcal{E}$ is said to be *applicable* in a plausibility model M if $M \models \langle \mathcal{E} \rangle \top$.

Unfolding the definition of $\langle \mathcal{E} \rangle$, we see what applicability means:

$$\begin{aligned}
&M \models \langle \mathcal{E} \rangle \top \Leftrightarrow \forall w \in D(M) : M, w \models \langle \mathcal{E} \rangle \top \Leftrightarrow \\
&\forall w \in D(M) : M, w \models \vee_{e \in D(\mathcal{E})} \langle \mathcal{E}, e \rangle \top \Leftrightarrow \\
&\forall w \in D(M), \exists e \in D(\mathcal{E}) : M, w \models \langle \mathcal{E}, e \rangle \top \Leftrightarrow \\
&\forall w \in D(M), \exists e \in D(\mathcal{E}) : M, w \models pre(e) \text{ and } M \otimes \mathcal{E}, (w, e) \models \top \Leftrightarrow \\
&\forall w \in D(M), \exists e \in D(\mathcal{E}) : M, w \models pre(e).
\end{aligned}$$

This says that no matter which is the actual world (it must be one of those considered possible), the action defines an outcome. This concept of applicability is equivalent to the one in [Bolander and Andersen, 2011]. The discussion in [de Lima, 2007, sect. 6.6] also notes this aspect, insisting that actions must be *meaningful*. The same sentiment is expressed by our notion of applicability.

Proposition 3.12. *Given a plausibility model M and an applicable event model $\mathcal{E}$, we have $D(M \otimes \mathcal{E}) \neq \emptyset$.*

The product update $M \otimes \mathcal{E}$ expresses the outcome(s) of doing $\mathcal{E}$ in the situation M, in the planning literature called *applying* $\mathcal{E}$ in M. The dynamic modality $[\mathcal{E}]$ expresses reasoning about what holds after applying $\mathcal{E}$.

Lemma 3.13. *Let M be a plausibility model and $\mathcal{E}$ an event model. Then $M \models [\mathcal{E}]\phi$ iff $M \otimes \mathcal{E} \models \phi$.*

Proof.

$$
\begin{aligned}
&M \models [\mathcal{E}]\phi \Leftrightarrow \forall w \in D(M) : M,w \models [\mathcal{E}]\phi \Leftrightarrow \\
&\forall w \in D(M) : M,w \models \bigwedge_{e \in D(\mathcal{E})} [\mathcal{E},e]\phi \Leftrightarrow \\
&\forall (w,e) \in D(M) \times D(\mathcal{E}) : M,w \models [\mathcal{E},e]\phi \Leftrightarrow \\
&\forall (w,e) \in D(M) \times D(\mathcal{E}) : M,w \models pre(e) \text{ implies } M \otimes \mathcal{E},(w,e) \models \phi \Leftrightarrow \\
&\forall (w,e) \in D(M \otimes \mathcal{E}) : M \otimes \mathcal{E},(w,e) \models \phi \Leftrightarrow \\
&M \otimes \mathcal{E} \models \phi.
\end{aligned}
$$

□

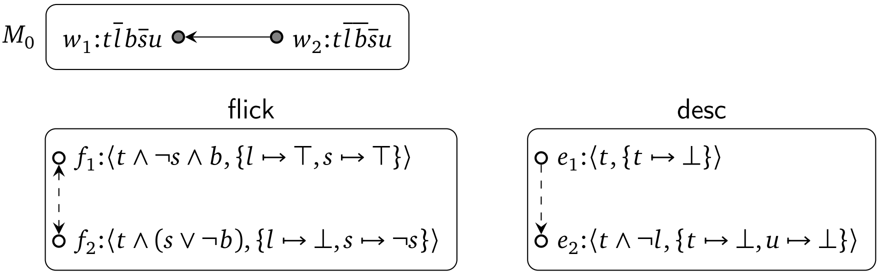
M_0
$w_1{:}t\bar{l}b\bar{s}u$
$w_2{:}t\bar{l}\bar{b}\bar{s}u$
flick
$f_1{:}\langle t \wedge \neg s \wedge b, \{l \mapsto \top, s \mapsto \top\}\rangle$
$f_2{:}\langle t \wedge (s \vee \neg b), \{l \mapsto \bot, s \mapsto \neg s\}\rangle$
desc
$e_1{:}\langle t, \{t \mapsto \bot\}\rangle$
$e_2{:}\langle t \wedge \neg l, \{t \mapsto \bot, u \mapsto \bot\}\rangle$

Figure 3.4: An information cell, M_0, and two event models, flick and desc.

Here we are looking at *global satisfaction*, by evaluating $[\mathcal{E}]\phi$ in all of M, rather than a specific world. The reason is that evaluation in planning must happen from the perspective of the planning agent and its "information state". Though one of the worlds of M is the actual world, the planning agent is ignorant about which it is. Whatever plan it comes up with, it must work in all of the worlds which are indistinguishable to the agent, that is, in the entire model. A similar point, and a similar solution, is found in [Jamroga and Ågotnes, 2007].

Example 3.14. We now return to the agent from Example 3.1. Her view of the initial situation (M_0) and her available actions (flick and desc) are seen in Figure 3.4. The propositional letters mean t: "top of stairs", l: "light on", b: "bulb working", s : "switch on" and u: "unharmed". Initially, in M_0, she believes that the bulb is working, and knows that she is at the top of the stairs, unharmed and that the switch and light is off: $M_0 \models Bb \wedge K(t \wedge u \wedge \neg l \wedge \neg s)$.

flick and desc represent flicking the light switch and trying to descend the stairs, respectively. Both require being at the top of the stairs (t). f_1 of flick expresses that if the bulb is working, turning on the switch will turn on the light, and f_2 that if the bulb is broken or the switch is currently on, the light will be off. The events are epistemically distinguishable, as the agent will be able to tell whether

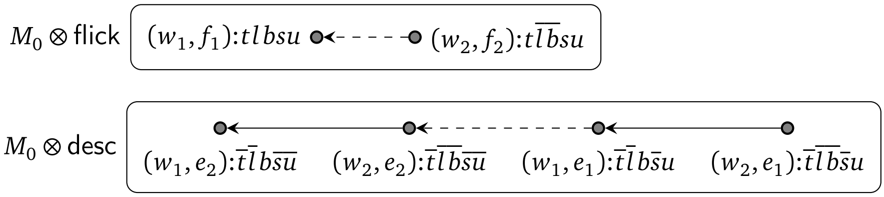
$M_0 \otimes$ flick
$(w_1, f_1){:}tlbsu$
$(w_2, f_2){:}t\overline{lb}su$
$M_0 \otimes$ desc
$(w_1, e_2){:}\overline{t}\overline{l}b\overline{su}$
$(w_2, e_2){:}\overline{t}\overline{l}\overline{b}\overline{su}$
$(w_1, e_1){:}\overline{t}\overline{l}b\overline{s}u$
$(w_2, e_1){:}\overline{t}\overline{l}\overline{b}\overline{s}u$

Figure 3.5: The models resulting from applying the actions flick and desc in M_0. Reflexive edges are not shown and the transitive closure is left implicit.

the light is on or off. desc describes descending the stairs, with or without the light on. e_1 covers the agent descending the stairs unharmed, and can happen regardless of there being light or not. The more plausible event e_2 represents the agent stumbling, though this can only happen in the dark. If the light is on, she will descend safely. Definition 3.11 and Lemma 3.13 let us express the action sequences possible in this scenario.

- $M_0 \models \langle \text{flick} \rangle \top \wedge \langle \text{desc} \rangle \top$. The agent can initially do either flick or desc.
- $M_0 \models [\text{flick}] \langle \text{desc} \rangle \top$. After doing flick, she can do desc.
- $M_0 \models [\text{desc}] (\neg \langle \text{flick} \rangle \top \wedge \neg \langle \text{desc} \rangle \top)$. Nothing can be done after desc.

Figure 3.5 shows the plausibility models arising from doing flick and desc in M_0. Via Lemma 3.13 she can now conclude:

- $M_0 \models [\text{flick}] (Kb \vee K\neg b)$: Flicking the light switch gives knowledge of whether the bulb works or not.
- $M_0 \models [\text{flick}] BKb$. She expects to come to know that it works.
- $M_0 \models [\text{desc}] (K\neg t \wedge B\neg u)$. Descending the stairs in the dark will definitely get her to the bottom, though she believes she will end up hurting herself.

■

3.3.3 Planning

We now turn to formalising planning and then proceed to answer two questions of particular interest: How do we verify that a given plan achieves a goal? And can we compute such plans? This section deals with the first question, plan verification, while the second, plan synthesis, is detailed in Section 3.4.

Definition 3.15 (Plan Language)**.** Given a finite set A of event models on $L(P)$, the *plan language* $\mathcal{L}(P, \mathsf{A})$ is given by:

$$\pi ::= \mathcal{E} \mid \mathsf{skip} \mid \mathsf{if}\ \phi\ \mathsf{then}\ \pi\ \mathsf{else}\ \pi \mid \pi; \pi$$

where $\mathcal{E} \in \mathsf{A}$ and $\phi \in L(P)$. We name members π of this language *plans*, and use $\mathsf{if}\ \phi\ \mathsf{then}\ \pi$ as shorthand for $\mathsf{if}\ \phi\ \mathsf{then}\ \pi\ \mathsf{else}\ \mathsf{skip}$.

The reading of the plan constructs are "do $\mathcal{E}$", "do nothing", "if ϕ then π, else π'", and "first π then π'" respectively. In the translations provided in Definition 3.16, the condition of the if-then-else construct becomes a K-formula, ensuring that branching depends only on worlds which are distinguishable to the agent. The idea is similar to the *meaningful plans* of [de Lima, 2007], where branching is allowed on *epistemically interpretable formulas* only.

Definition 3.16 (Translation)**.** Let α be one of s, w, sp or wp. We define an α-*translation* as a function $[\cdot]_\alpha : \mathcal{L}(P, \mathsf{A}) \to (L(P) \to L(P))$:

$$[\mathcal{E}]_\alpha\, \phi := \langle \mathcal{E} \rangle \top \wedge \begin{cases} [\mathcal{E}]\, XK\phi & \text{if } \alpha = s \\ \widehat{K}\, \langle \mathcal{E} \rangle XK\phi & \text{if } \alpha = w \\ [\mathcal{E}]\, BXK\phi & \text{if } \alpha = sp \\ [\mathcal{E}]\, \widehat{B}XK\phi & \text{if } \alpha = wp \end{cases}$$

$$[\mathsf{skip}]_\alpha\, \phi := \phi$$

$$\left[\mathsf{if}\ \phi'\ \mathsf{then}\ \pi\ \mathsf{else}\ \pi'\right]_\alpha \phi := (K\phi' \to [\pi]_\alpha\phi) \wedge (\neg K\phi' \to [\pi']_\alpha\phi)$$

$$[\pi; \pi']_\alpha\phi := [\pi]_\alpha([\pi']_\alpha\phi)$$

We call $[\cdot]_s$ the *strong translation*, $[\cdot]_w$ the *weak translation*, $[\cdot]_{sp}$ the *strong plausibility* translation and $[\cdot]_{wp}$ the *weak plausibility* translation.

The translations are constructed specifically to make the following lemma hold, providing a semantic interpretation of plans (leaving out skip and $\pi_1; \pi_2$).

Lemma 3.17. *Let M be an information cell, $\mathcal{E}$ an event model and ϕ a formula of $L(P)$. Then:*

1. $M \models [\mathcal{E}]_s\phi$ iff $M \models \langle \mathcal{E} \rangle \top$ *and for each information cell M' of $M \otimes \mathcal{E} : M' \models \phi$.*
2. $M \models [\mathcal{E}]_w\phi$ iff $M \models \langle \mathcal{E} \rangle \top$ *and for some information cell M' of $M \otimes \mathcal{E} : M' \models \phi$.*
3. $M \models [\mathcal{E}]_{sp}\phi$ iff $M \models \langle \mathcal{E} \rangle \top$ *and for each most plausible information cell M' of $M \otimes \mathcal{E} : M' \models \phi$.*

4. $M \models [\mathcal{E}]_{wp}\phi$ iff $M \models \langle\mathcal{E}\rangle\top$ *and for some most plausible information cell* M' *of* $M \otimes \mathcal{E} : M' \models \phi$.
5. $M \models [\text{if } \phi' \text{ then } \pi \text{ else } \pi']_\alpha\phi$ iff
($M \models \phi'$ *implies* $M \models [\pi]_\alpha\phi$) *and* ($M \not\models \phi'$ *implies* $M \models [\pi']_\alpha\phi$).

Proof. We only prove 4 and 5, as 1–4 are very similar. For 4 we have:

$M \models [\mathcal{E}]_{wp}\,\phi \iff M \models \langle\mathcal{E}\rangle\top \wedge [\mathcal{E}]\,\widehat{B}XK\phi \iff^{\text{Lemma 3.13}}$
$M \models \langle\mathcal{E}\rangle\top$ and $M \otimes \mathcal{E} \models \widehat{B}XK\phi \iff$
$M \models \langle\mathcal{E}\rangle\top$ and $\forall(w,e) \in D(M \otimes \mathcal{E}) : M \otimes \mathcal{E},(w,e) \models \widehat{B}XK\phi \iff^{\text{Prop. 3.12}}$
$M \models \langle\mathcal{E}\rangle\top$ and $\exists(w,e) \in Min_{\leq}D(M \otimes \mathcal{E}) : M \otimes \mathcal{E},(w,e) \models XK\phi \iff$
$M \models \langle\mathcal{E}\rangle\top$ and $\exists(w,e) \in Min_{\leq}D(M \otimes \mathcal{E}) : M \otimes \mathcal{E} \restriction [(w,e)]_\sim,(w,e) \models K\phi \iff$
$M \models \langle\mathcal{E}\rangle\top$ and $\exists(w,e) \in Min_{\leq}D(M \otimes \mathcal{E}) : M \otimes \mathcal{E} \restriction [(w,e)]_\sim \models \phi \iff$
$M \models \langle\mathcal{E}\rangle\top$ and in some most plausible information cell M' of $M \otimes \mathcal{E}$, $M' \models \phi$.

For if-then-else, first note that:

$M \models \neg K\phi' \to [\pi]_\alpha\phi \iff \forall w \in D(M): M,w \models \neg K\phi' \to [\pi]_\alpha\phi \iff$
$\forall w \in D(M): M,w \models \neg K\phi'$ implies $M,w \models [\pi]_\alpha\phi \iff^{M \text{ is an info. cell}}$
$\forall w \in D(M)$: if $M,v \models \neg\phi'$ for some $v \in D(M)$ then $M,w \models [\pi]_\alpha\phi \iff$
if $M,v \models \neg\phi'$ for some $v \in D(M)$ then $\forall w \in D(M): M,w \models [\pi]_\alpha\phi \iff$
$M \not\models \phi'$ implies $M \models [\pi']_\alpha\phi$.

Similarly, we can prove:

$$M \models K\phi' \to [\pi]_\alpha\phi \iff M \models K\phi' \text{ implies } M \models [\pi']_\alpha\phi.$$

Using these facts, we get:

$M \models [\text{if } \phi' \text{ then } \pi \text{ else } \pi']_\alpha\phi \iff M \models (K\phi' \to [\pi]_\alpha\phi) \wedge (\neg K\phi' \to [\pi']_\alpha\phi) \iff$
$M \models K\phi' \to [\pi]_\alpha\phi$ and $M \models \neg K\phi' \to [\pi']_\alpha\phi \iff$
($M \models \phi'$ implies $M \models [\pi]_\alpha\phi$) and ($M \not\models \phi'$ implies $M \models [\pi']_\alpha\phi$).

□

Using XK (as is done in all translations) means that reasoning after an action is relative to a particular information cell (as $M,w \models XK\phi \iff M \restriction [w]_\sim,w \models K\phi \iff M \restriction [w]_\sim \models \phi$).

Definition 3.18 (Planning Problems and Solutions). Let P be a finite set of propositional symbols. A planning problem on P is a triple $\mathcal{P} = (M_0, \mathsf{A}, \phi_g)$ where

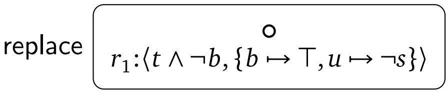
replace
$r_1 {:} \langle t \wedge \neg b, \{b \mapsto \top, u \mapsto \neg s\}\rangle$

Figure 3.6: Event model for replacing a broken bulb.

- M_0 is a finite information cell on P called the *initial state.*
- A is a finite set of event models on $L(P)$ called the *action library.*
- $\phi_g \in L(P)$ is the *goal (formula).*

A plan $\pi \in \mathcal{L}(P, \mathsf{A})$ is an *α-solution* to $\mathcal{P}$ if $M_0 \models [\pi]_\alpha \phi_g$. For a specific choice of $\alpha = s/w/sp/wp$, we will call π a *strong/weak/strong plausibility/weak plausibility*-solution respectively.

Given a π, we wish to check whether π is an α-solution (for some particular α) to $\mathcal{P}$. This can be done via model checking the dynamic formula given by the translation $[\pi]_\alpha\, \phi_g$ in the initial state of $\mathcal{P}$.

A strong solution π is one that guarantees that ϕ_g will hold after executing it ("π achieves ϕ_g"). If π is a weak solution, it achieves ϕ_g for at least one particular sequence of outcomes. Strong and weak plausibility-solutions are as strong- and weak-solutions, except that they need only achieve ϕ_g for *all of/some of* the most plausible outcomes.

Example 3.19. The basement scenario (Example 3.1) can be formalised as the planning problem $\mathcal{P}_B = (M_0, \{\mathsf{flick}, \mathsf{desc}\}, \phi_g)$ with M_0, flick and desc being defined in Figure 3.4 and $\phi_g = \neg t \wedge u$. Let $\pi_1 = \mathsf{desc}$. We then have that:

$$M_0 \models [\mathsf{desc}]_w (\neg t \wedge u) \Leftrightarrow M_0 \models \langle \mathsf{desc} \rangle \top \wedge \widehat{K}\, \langle \mathsf{desc} \rangle XK(\neg t \wedge u) \Leftrightarrow^{\text{desc is applic.}}$$
$$M_0 \models \widehat{K}\, \langle \mathsf{desc} \rangle XK(\neg t \wedge u) \Leftrightarrow \exists w \in D(M_0) : M_0, w \models \langle \mathsf{desc} \rangle XK(\neg t \wedge u).$$

Picking w_1, we have

$$M_0, w_1 \models \langle \mathsf{desc} \rangle XK(\neg t \wedge u) \Leftrightarrow M_0 \otimes \mathsf{desc}, (w_1, e_1) \models XK(\neg t \wedge u) \Leftrightarrow$$
$$M_0 \otimes \mathsf{desc} \restriction [(w_1, e_1)]_\sim \models (\neg t \wedge u)$$

which holds as seen in Figure 3.5. Thus, π_1 is a weak solution. Further, Lemma 3.17 tells us that π_1 is not a $s/wp/sp$ solution, as u does not hold in the (most plausible) information cell $M \otimes \mathsf{desc} \restriction \{(w_1, e_2), (w_2, e_2)\}$.

The plan $\pi_2 = \mathsf{flick}; \mathsf{desc}$ is a strong plausibility solution, as can be verified by $M_0 \models [\pi_2]_{sp} (\neg t \wedge u)$. Without an action for replacing the lightbulb, there are no strong solutions. Let replace be the action in Figure 3.6, where $post(r_1)(u) = \neg s$ signifies that if the power is on, the agent will hurt herself, and define a new problem $\mathcal{P}'_B = \{M_0, \{\mathsf{flick}, \mathsf{desc}, \mathsf{replace}\}, \phi_g)$. Then

$$\pi_3 = \mathsf{flick}; (\mathsf{if}\ \neg l\ \mathsf{then}\ \mathsf{flick}; \mathsf{replace}; \mathsf{flick}); \mathsf{desc}$$

is a strong solution (we leave verification to the reader): If the light comes on after flicking the switch (as expected) she can safely walk down the stairs. If it does not, she turns off the power, replaces the broken bulb, turns the power on again (this time knowing that the light will come on), and then proceeds as before. ■

Besides being an sp-solution, π_2 is also a w- and a wp-solution, indicating a hierarchy of strengths of solutions. This should come as no surprise, given both the formal and intuitive meaning of planning and actions presented so far. In fact, this hierarchy exists for any planning problem, as shown by the following result which is a consequence of Lemma 3.17 (stated without proof).

Lemma 3.20. *Let $\mathcal{P} = (M_0, \mathsf{A}, \phi_g)$ be a planning problem. Then:*

- *Any strong solution to $\mathcal{P}$ is also a strong plausibility solution:* $M_0 \models [\pi]_s\, \phi_g \Rightarrow M_0 \models [\pi]_{sp}\, \phi_g$.
- *Any strong plausibility solution to $\mathcal{P}$ is also a weak plausibility solution:* $M_0 \models [\pi]_{sp}\, \phi_g \Rightarrow M_0 \models [\pi]_{wp}\, \phi_g$.
- *Any weak plausibility solution to $\mathcal{P}$ is also a weak solution:* $M_0 \models [\pi]_{wp}\, \phi_g \Rightarrow M_0 \models [\pi]_w\, \phi_g$.

3.4 Plan Synthesis

In this section we show how to synthesise conditional plans for solving planning problems. Before we can give the concrete algorithms, we establish some technical results which are stepping stones to proving termination of our planning algorithm, and hence decidability of plan existence in our framework.

3.4.1 Bisimulations, contractions and modal equivalence

We now define bisimulations on plausibility models. For our purpose it is sufficient to define bisimulations on $\sim$-connected models, that is, on information cells. First we define a *normal plausibility relation* which will form the basis of our bisimulation definition.

Definition 3.21 (Normality)**.** Given is an information cell $M = (W, \sim, \leq, V)$ on P. By slight abuse of language, two worlds $w, w' \in W$ are said to *have the same valuation* if for all $p \in P$: $w \in V(p) \Leftrightarrow w' \in V(p)$. Define an equivalence relation on W: $w \approx w'$ iff w and w' has the same valuation. Now define $w \preceq w'$ iff $Min_{\leq}([w]_{\approx}) \leq Min_{\leq}([w']_{\approx})$. This defines the *normal plausibility relation*. M is called *normal* if $\preceq = \leq$. The *normalisation* of $M = (W, \sim, \leq, V)$ is $M' = (W, \sim, \preceq, V)$.

Definition 3.22 (Bisimulation)**.** Let $M = (W, \sim, \leq, V)$ and $M' = (W', \sim', \leq', V')$ be information cells on P. A non-empty relation $\mathcal{R} \subseteq W \times W'$ is a *bisimulation* between M and M' (and M, M' are called *bisimilar*) if for all $(w, w') \in \mathcal{R}$:

[atom] For all $p \in P$: $w \in V(p)$ iff $w' \in V'(p)$.

[forth] If $v \in W$ and $v \preceq w$ then there is a $v' \in W'$ s.t. $v' \preceq' w'$ and $(v, v') \in \mathcal{R}$.

[back] If $v' \in W'$ and $v' \preceq w'$ then there is a $v \in W$ s.t. $v \preceq w$ and $(v, v') \in \mathcal{R}$.

If $\mathcal{R}$ has domain W and codomain W', it is called *total*. If $M = M'$, it is called an *autobisimulation* (on M). Two worlds w and w' of an information cell $M = (W, \sim, \leq, V)$ are called *bisimilar* if there exists an autobisimulation $\mathcal{R}$ on M with $(w, w') \in \mathcal{R}$.

We are here only interested in total bisimulations, so, unless otherwise stated, we assume this in the following. Note that our definition of bisimulation immediately implies that there exists a (total) bisimulation between any information cell and its normalisation. Note also that for normal models, the bisimulation definition becomes the standard modal logic one.[4]

Lemma 3.23. *If two worlds of an information cell have the same valuation they are bisimilar.*

Proof. Assume worlds w and w' of an information cell $M = (W, \sim, \leq, V)$ have the same valuation. Let $\mathcal{R}$ be the relation that relates each world of M to itself and additionally relates w to w'. We want to show that $\mathcal{R}$ is a bisimulation. This amounts to showing [atom], [forth] and [back] for the pair $(w, w') \in \mathcal{R}$. [atom] holds trivially since $w \approx w'$. For [forth], assume $v \in W$ and $v \preceq w$. We need to find a $v' \in W$ s.t. $v' \preceq w'$ and $(v, v') \in \mathcal{R}$. Letting $v' = v$, it suffices to prove $v \preceq w'$. Since $w \approx w'$ this is immediate: $v \preceq w \Leftrightarrow Min_{\leq}([v]_{\approx}) \leq Min_{\leq}([w]_{\approx}) \overset{w \approx w'}{\Leftrightarrow} Min_{\leq}([v]_{\approx}) \leq Min_{\leq}([w']_{\approx}) \Leftrightarrow v \preceq w'$. [back] is proved similarly. □

[4] We didn't include a condition for the epistemic relation, $\sim$, in [back] and [forth], simply because we are here only concerned with $\sim$-connected models.

Unions of autobisimulations are autobisimulations. We can then in the standard way define the *(bisimulation) contraction* of a normal information cell as its quotient with respect to the union of all autobisimulations [Blackburn and van Benthem, 2006].[5] The contraction of a non-normal model is taken to be the contraction of its normalisation. In a contracted model, no two worlds are bisimilar, by construction. Hence, by Lemma 3.23, no two worlds have the same valuation. Thus, the contraction of an information cell on a finite set of proposition symbols P contains at most $2^{|P|}$ worlds. Since any information cell is bisimilar to its contraction [Blackburn and van Benthem, 2006], this shows that there can only exist finitely many non-bisimilar information cells on any given finite set P.

Two information cells M and M' are called *modally equivalent*, written $M \equiv M'$, if for all formulas ϕ in $L(P)$: $M \models \phi \Leftrightarrow M' \models \phi$. Otherwise, they are called *modally inequivalent*. We now have the following standard result (the result is standard for standard modal languages and bisimulations, but it is not trivial that it also holds here).

Theorem 3.24. *If two information cells are (totally) bisimilar they are modally equivalent.*

Proof. We need to show that if $\mathcal{R}$ is a total bisimulation between information cells M and M', then for all formulas ϕ of $L(P)$: $M \models \phi \Leftrightarrow M' \models \phi$. First we show that we only have to consider formulas ϕ of the static sublanguage of $L(P)$, that is, the language without the $[\mathcal{E}, e]$ modalities. In [Baltag and Smets, 2006], reduction axioms from the dynamic to the static language are given for a language similar to $L(P)$. The differences in language are our addition of postconditions and the fact that our belief modality is defined from the global plausibility relation rather than being localised to epistemic equivalence classes. The latter difference is irrelevant when only considering information cells as we do here. The former difference of course means that the reduction axioms presented in [Baltag and Smets, 2006] will not suffice for our purpose. [van Ditmarsch and Kooi, 2008] shows that adding postconditions to the language without the doxastic modalities only requires changing the reduction axiom for $[\mathcal{E}, e]\, p$, where p is a propositional symbol. Thus, if we take the reduction axioms of [Baltag and Smets, 2006] and replace the reduction axiom for $[\mathcal{E}, e]\, p$ by the one in [van Ditmarsch and Kooi, 2008], we get reduction axioms for our framework. We leave out the details.

We now need to show that if $\mathcal{R}$ is a total bisimulation between information cells M and M', then for all $[\mathcal{E}, e]$-free formulas ϕ of $L(P)$: $M \models \phi \Leftrightarrow M' \models \phi$. Since $\mathcal{R}$ is total, it is sufficient to prove that for all $[\mathcal{E}, e]$-free formulas ϕ of $L(P)$ and all $(w, w') \in \mathcal{R}$: $M, w \models \phi \Leftrightarrow M', w' \models \phi$. The proof is by induction on ϕ. In the

[5] More precisely, let M be a normal information cell and let $\mathcal{R}$ be the union of all autobisimulations on M. Then the contraction $M' = (W', \sim', \leq', V')$ of M has as worlds the equivalence classes $[w]_{\mathcal{R}} = \{w' \mid (w, w') \in \mathcal{R}\}$ and has $[w]_{\mathcal{R}} \leq' [w']_{\mathcal{R}}$ iff $v \leq v'$ for some $v \in [w]_{\mathcal{R}}$ and $v' \in [w']_{\mathcal{R}}$.

induction step we are going to need the induction hypothesis for several different choices of $\mathcal{R}, w$ and w', so what we will actually prove by induction on ϕ is this: For all formulas ϕ of $L(P)$, if $\mathcal{R}$ is a total bisimulation between information cells M and M' on P and $(w,w') \in \mathcal{R}$, then $M,w \models \phi \Leftrightarrow M',w' \models \phi$.

The base case is when ϕ is propositional. Then the required follows immediately from [atom], using that $(w,w') \in \mathcal{R}$. For the induction step, we have the following cases of ϕ: $\neg\psi, \psi \wedge \gamma, X\psi, K\psi, B^\gamma\psi$. The first two cases are trivial. So is $X\psi$, as $X\psi \longleftrightarrow \psi$ holds on any information cell. For $K\psi$ we reason as follows. Let $\mathcal{R}$ be a total bisimulation between information cells M and M' with $(w,w') \in \mathcal{R}$. Using that $\mathcal{R}$ is total and that M and M' are both $\sim$-connected we get: $M,w \models K\psi \Leftrightarrow \forall v \in W\colon M,v \models \psi \overset{\text{i.h.}}{\Leftrightarrow} \forall v' \in W'\colon M',v \models \psi \Leftrightarrow M',w' \models K\psi$.

The case of $B^\gamma\psi$ is more involved. Let $M, M', \mathcal{R}, w$ and w' be as above. By symmetry, it suffices to prove $M,w \models B^\gamma\psi \Rightarrow M',w' \models B^\gamma\psi$. So assume $M,w \models B^\gamma\psi$, that is, $M,v \models \psi$ for all $v \in Min_{\leq}\{u \in W \mid M,u \models \gamma\}$. We need to prove $M',v' \models \psi$ for all $v' \in Min_{\leq'}\{u' \in W' \mid M',u' \models \gamma\}$. So let $v' \in Min_{\leq'}\{u' \in W' \mid M',u' \models \gamma\}$. By definition of $Min_{\leq'}$ this means that:

$$\text{for all } u' \in W', \text{ if } M',u' \models \gamma \text{ then } v' \leq' u'. \tag{3.1}$$

Choose an $x \in Min_{\leq}\{u \in W \mid u \approx u' \text{ and } (u',v') \in \mathcal{R}\}$. We want to use (3.1) to show that the following holds:

$$\text{for all } u \in W, \text{ if } M,u \models \gamma \text{ then } x \leq u. \tag{3.2}$$

To prove (3.2), let $u \in W$ with $M,u \models \gamma$. Choose u' with $(u,u') \in \mathcal{R}$. The induction hypothesis implies $M',u' \models \gamma$. We now prove that $v' \leq' Min_{\leq'}([u']_\approx)$. To this end, let $u'' \in [u']_\approx$. We need to prove $v' \leq' u''$. Since $u'' \approx u'$, Lemma 3.23 implies that u' and u'' are bisimilar. By induction hypothesis we then get $M',u'' \models \gamma$.[6] Using (3.1) we now get $v' \leq' u''$, as required. This show $v' \leq' Min_{\leq'}([u']_\approx)$. We now have $Min_{\leq'}([v']_\approx) \leq' v' \leq' Min_{\leq'}([u']_\approx)$, and hence $v' \preceq u'$. By [back] there is then a v s.t. $(v,v') \in \mathcal{R}$ and $v \preceq u$. By choice of x, $x \leq Min_{\leq}([v]_\approx)$. Using $v \preceq u$, we now finally get: $x \leq Min_{\leq}([v]_\approx) \leq Min_{\leq}([u]_\approx) \leq u$. This shows that (3.2) holds.

From (3.2) we can now conclude $x \in Min_{\leq}\{u \in W \mid M,u \models \gamma\}$ and hence, by original assumption, $M,x \models \psi$. By choice of x there is an $x' \approx x$ with $(x',v') \in \mathcal{R}$. Since $M,x \models \psi$ and $x' \approx x$, we can again use Lemma 3.23 and the induction hypothesis to conclude $M,x' \models \psi$. Since $(x',v') \in \mathcal{R}$, another instance of the induction hypothesis gives us $M',v' \models \psi$, and we are done. □

[6] Note that we here use the induction hypothesis for the autobisimulation on M' linking u' and u'', not the bisimulation $\mathcal{R}$ between M and M'.

Previously we proved that there can only be finitely many non-bisimilar information cells on any finite set P. Since we have now shown that bisimilarity implies modal equivalence, we immediately get the following result, which will be essential to our proof of termination of our planning algorithms.

Corollary 3.25. *Given any finite set P, there are only finitely many modally inequivalent information cells on P.*

3.4.2 Planning Trees

When synthesising plans, we explicitly construct the search space of the problem as a labelled AND-OR tree, a familiar model for planning under uncertainty [Ghallab et al., 2004]. Our AND-OR trees are called *planning trees*.

Definition 3.26 (Planning Tree). A *planning tree* is a finite, labelled AND-OR tree in which each node n is labelled by a plausibility model $M(n)$, and each edge (n,m) leaving an OR-node is labelled by an event model $\mathcal{E}(n,m)$.

Planning trees for planning problems $\mathcal{P} = (M_0, \mathsf{A}, \phi_g)$ are constructed as follows: Let the initial planning tree T_0 consist of just one OR-node $root(T_0)$ with $M(root(T_0)) = M_0$ (the root labels the initial state). A planning tree for $\mathcal{P}$ is then any tree that can be constructed from T_0 by repeated applications of the following non-deterministic tree expansion rule.

Definition 3.27 (Tree Expansion Rule). Let T be a planning tree for a planning problem $\mathcal{P} = (M_0, \mathsf{A}, \phi_g)$. The tree expansion rule is defined as follows. Pick an OR-node n in T and an event model $\mathcal{E} \in \mathsf{A}$ applicable in $M(n)$ with the proviso that $\mathcal{E}$ does not label any existing outgoing edges from n. Then:

1. Add a new AND-node m to T with $M(m) = M(n) \otimes \mathcal{E}$, and add an edge (n,m) with $\mathcal{E}(n,m) = \mathcal{E}$.
2. For each information cell M' in $M(m)$, add an OR-node m' with $M(m') = M'$ and add the edge (m,m').

The tree expansion rule is similar in structure to—and inspired by—the expansion rules used in tableau calculi, e.g. for modal and description logics [Horrocks et al., 2006]. Note that the expansion rule applies only to OR-nodes, and that an applicable event model can only be used once at each node.

Considering single-agent planning a two-player game, a useful analogy for planning trees are game trees. At an OR-node n, the agent gets to pick any applicable action

$\mathcal{E}$ it pleases, winning if it ever reaches an information model in which the goal formula holds (see the definition of solved nodes further below). At an AND-node m, the environment responds by picking one of the information cells of $M(m)$—which of the distinguishable outcomes is realised when performing the action.

Without restrictions on the tree expansion rule, even very simple planning problems might be infinitely expanded (e.g. by repeatedly choosing a no-op action). Finiteness of trees (and therefore termination) is ensured by the following blocking condition.

$\mathcal{B}$ The tree expansion rule may not be applied to an OR-node n for which there exists an ancestor OR-node m with $M(m) \equiv M(n)$.[7]

Lemma 3.28 (Termination)**.** *Any planning tree built by repeated application of the tree expansion rule under condition $\mathcal{B}$ is finite.*

Proof. Planning trees built by repeated application of the tree expansion rule are finitely branching: the action library is finite, and every plausibility model has only finitely many information cells (the initial state and all event models in the action library are assumed to be finite, and taking the product update of a finite information cell with a finite event model always produces a finite result). Furthermore, condition $\mathcal{B}$ ensures that no branch has infinite length: there only exists finitely many modally inequivalent information cells over any language $L(P)$ with finite P (Corollary 3.25). König's Lemma now implies finiteness of the planning tree. □

Example 3.29. Let's consider a planning tree in relation to our basement scenario (cf. Example 3.19). Here the planning problem is $\mathcal{P}_B = (M_0, \{\mathsf{flick}, \mathsf{desc}\}, \phi_g)$ with M_0, flick and desc being defined in Figure 3.4 and $\phi_g = \neg t \wedge u$. We have illustrated the planning tree T in Figure 3.7. The root n_0 is an OR-node (representing the initial state M_0), to which the tree expansion rule of Definition 3.27 has been applied twice, once with action $\mathcal{E} = \mathsf{flick}$ and once with $\mathcal{E} = \mathsf{desc}$.

The result of the two tree expansions on n_0 is two AND-nodes (children of n_0) and four OR-nodes (grandchildren of n_0). We end our exposition of the tree expansion rule here, and note that the tree has been fully expanded under the blocking condition $\mathcal{B}$, the dotted edge indicating a leaf having a modally equivalent ancestor. Without the blocking condition, this branch could have been expanded ad infinitum. ■

Let T denote a planning tree containing an AND-node n with a child m. The node m is called a *most plausible child* of n if $M(m)$ is among the most plausible information cells of $M(n)$.

[7] Modal equivalence between information cells can be decided by taking their respective bisimulation contractions and then compare for isomorphism, cf. Section 3.4.1.

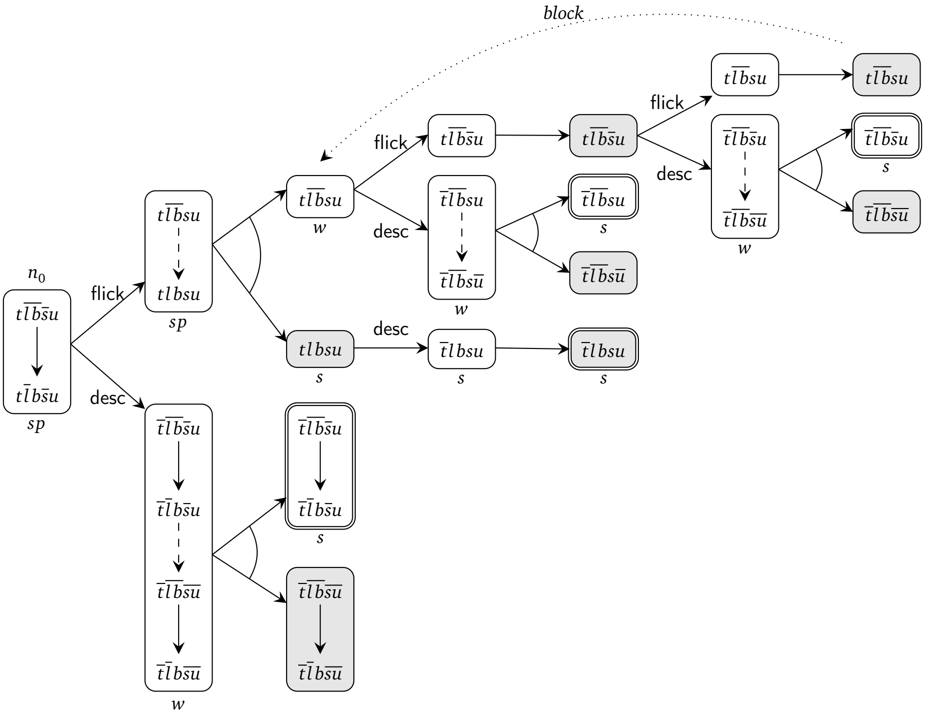

Figure 3.7: A planning tree T for $\mathcal{P}_B$. Each node contains a (visually compacted) plausibility model. Most plausible children of AND-nodes are gray, doubly drawn OR-nodes satisfy the goal formula, and below solved nodes we've indicated their strength.

Definition 3.30 (Solved Nodes)**.** Let T be any planning tree for a planning problem $\mathcal{P} = (M_0, \mathsf{A}, \phi_g)$. Let α be one of s, w, sp or wp. By recursive definition, a node n in T is called *α-solved* if one of the following holds:

- $M(n) \models \phi_g$ (the node satisfies the goal formula).
- n is an OR-node having at least one α-solved child.
- n is an AND-node and:
  - If $\alpha = s$ then all children of n are α-solved.
  - If $\alpha = w$ then at least one child of n is α-solved.
  - If $\alpha = sp$ then all most plausible children of n are α-solved.
  - If $\alpha = wp$ then at least one of the most plausible children of n is α-solved.

Let T denote any planning tree for a planning problem $\mathcal{P} = (M_0, \mathsf{A}, \phi_g)$. Below we show that when an OR-node n of T is α-solved, it is possible to construct an α-solution to the planning problem $(M(n), \mathsf{A}, \phi_g)$. In particular, if the root node is α-solved, an α-solution to $\mathcal{P}$ can be constructed. As it is never necessary to expand an α-solved node, nor any of its descendants, we can augment the blocking condition $\mathcal{B}$ in the following way (parameterised by α where α is one of s, w, sp or wp).

$\mathcal{B}_\alpha$ The tree expansion rule may not be applied to an OR-node n if one of the following holds: 1) n is α-solved; 2) n has an α-solved ancestor; 3) n has an ancestor OR-node m with $M(m) \equiv M(n)$.

A planning tree that has been built according to $\mathcal{B}_\alpha$ is called an *α-planning tree*. Since $\mathcal{B}_\alpha$ is more strict than $\mathcal{B}$, Lemma 3.28 immediately gives finiteness of α-planning trees—and hence termination of any algorithm building such trees by repeated application of the tree expansion rule. Note that a consequence of $\mathcal{B}_\alpha$ is that in any α-planning tree an α-solved OR-node is either a leaf or has exactly one α-solved child. We make use of this in the following definition.

Definition 3.31 (Plans for Solved Nodes). Let T be any α-planning tree for $\mathcal{P} = (M_0, \mathsf{A}, \phi_g)$. For each α-solved node n in T, a plan $\pi(n)$ is defined recursively by:

- if $M(n) \models \phi_g$, then $\pi(n) = \mathsf{skip}$.
- if n is an OR-node and m its α-solved child, then $\pi(n) = \mathcal{E}(n,m); \pi(m)$.
- if n is an AND-node and $m_1, \dots, m_k$ its α-solved children, then
  - If $k = 1$ then $\pi(n) = \pi(m_1)$.
  - If $k > 1$ then for all $i = 1, \dots, k$ let δ_{m_i} denote a formula true in $M(m_i)$ but not in any of the $M(m_j) \not\equiv M(m_i)$ and let $\pi(n) =$ $\mathsf{if}\ \delta_{m_1}\ \mathsf{then}\ \pi(m_1)\ \mathsf{else\ if}\ \delta_{m_2}\ \mathsf{then}\ \pi(m_2)\ \mathsf{else}\ \cdots\ \mathsf{if}\ \delta_{m_k}\ \mathsf{then}\ \pi(m_k)$.

Note that the plan $\pi(n)$ of a α-solved node n is only uniquely defined up to the choice of δ-formulas in the if-then-else construct. This ambiguity in the definition of $\pi(n)$ will not cause any troubles in what follows, as it only depends on formulas satisfying the stated property. We need, however, to be sure that such formulas always exist and can be computed. To prove this, assume n is an AND-node and $m_1, \dots, m_k$ its α-solved children. Choose $i \in \{1, \dots, k\}$, and let $m_{n_1}, \dots, m_{n_l}$ denote the subsequence of $m_1, \dots, m_k$ for which $M(m_{n_j}) \not\equiv M(m_i)$. We need to prove the existence of a formula δ_{m_i} such that $M(m_i) \models \delta_{m_i}$ but $M(m_{n_j}) \not\models \delta_{m_i}$ for all $j = 1, \dots, l$. Since $M(m_{n_j}) \not\equiv M(m_i)$ for all $j = 1, \dots, l$, there exists formulas δ_j such that $M(m_i) \models \delta_j$ but $M(m_{n_j}) \not\models \delta_j$. We then get that $\delta_1 \wedge \delta_2 \wedge \cdots \wedge \delta_l$ is true in

$M(m_i)$ but none of the $M(m_{n_j})$. Such formulas can definitely be computed, either by brute force search through all formulas ordered by length or more efficiently and systematically by using characterising formulas as in [Andersen et al., 2012] (however, characterising formulas for the present formalism are considerably more complex than in the purely epistemic framework of the cited paper).

Let n be a node of a planning tree T. We say that n is *solved* if it is α-solved for some α. If n is s-solved then it is also sp-solved, if sp-solved then wp-solved, and if wp-solved then w-solved. This gives a natural ordering $s > sp > wp > w$. Note the relation to Lemma 3.20. We say that a solved node n has *strength* α, if it is α-solved but not β-solved for any $\beta > \alpha$, using the aforementioned ordering.

Example 3.32. Consider again the planning tree T in Figure 3.7 for the planning problem $\mathcal{P}_B = (M_0, \{\mathsf{flick}, \mathsf{desc}\}, \phi_g)$ with $\phi_g = \neg t \wedge u$. Each solved node has been labelled by its strength. The reader is encouraged to check that each node has been labelled correctly according to Definition 3.30. The leafs satisfying the goal formula ϕ_g have strength s, by definition. The strength of the root node is sp, as its uppermost child has strength sp. The reason this child has strength sp is that *its* most plausible child has strength s.

We see that T is an sp-planning tree, as it is possible to achieve T from n_0 by applying tree expansions in an order that respects $\mathcal{B}_{sp}$. However, it is not the smallest sp-planning tree for the problem, as e.g. the lower subtree is not required for n_0 to be sp-solved. Moreover, T is *not* a w-planning tree, as $\mathcal{B}_w$ would have blocked further expansion once either of the three solved leafs were expanded.

In our soundness result below, we show that plans of α-solved roots are always α-solutions to their corresponding planning problems. Applying Definition 3.31 to the sp-planning tree T gives an sp-solution to the basement planning problem, viz. $\pi(n_0) = \mathsf{flick};\ \mathsf{desc};\ \mathsf{skip}$. This is the solution we referred to as the *best* in Example 3.1: Assuming all actions result in their most plausible outcomes, the best plan is to flick the switch and then descend. After having executed the first action of the plan, flick, the agent will know whether the bulb is broken or not. This is signified by the two distinct information cells resulting from the flick action, see Figure 3.7. An agent capable of replanning could thus choose to revise her plan and/or goal if the bulb turns out to be broken. ■

Theorem 3.33 (Soundness). *Let α be one of s, w, sp or wp. Let T be an α-planning tree for a problem $\mathcal{P} = (M_0, \mathsf{A}, \phi_g)$ such that $root(T)$ is α-solved. Then $\pi(root(T))$ is an α-solution to $\mathcal{P}$.*

Proof. We need to prove that $\pi(root(T))$ is an α-solution to $\mathcal{P}$, that is, $M_0 \models [\pi(root(T))]_\alpha\, \phi_g$. With M_0 labelling the the label of the root, this can be restated as $M(root(T)) \models [\pi(root(T))]_\alpha\, \phi_g$. To prove this fact, we will prove the following stronger claim:

$$\text{For each } \alpha\text{-solved } \textsc{or}\text{-node } n \text{ in } T,\ M(n) \models [\pi(n)]_\alpha\, \phi_g.$$

We prove this by induction on the height of n. The base case is when n is a leaf (height 0). Since n is α-solved, we must have $M(n) \models \phi_g$. In this case $\pi(n) = \mathsf{skip}$. From $M(n) \models \phi_g$ we can conclude $M(n) \models [\mathsf{skip}]_\alpha\, \phi_g$, that is, $M(n) \models [\pi(n)]_\alpha\, \phi_g$. This covers the base case. For the induction step, let n be an arbitrary α-solved OR-node n of height $h > 0$. Let m denote the α-solved child of n, and $m_1, \dots, m_l$ denote the children of m. Let $m_{n_1}, \dots, m_{n_k}$ denote the subsequence of $m_1, \dots, m_l$ consisting of the α-solved children of m. Then, by Definition 3.31,

- If $k = 1$ then $\pi(n) = \mathcal{E}(n, m); \pi(m_{n_1})$.
- If $k > 1$ then $\pi(n) = \mathcal{E}(n, m); \pi(m)$ where $\pi(m) =$ $\mathsf{if}\ \delta_{m_{n_1}}\ \mathsf{then}\ \pi(m_{n_1})\ \mathsf{else\ if}\ \delta_{m_{n_2}}\ \mathsf{then}\ \pi(m_{n_2})\ \mathsf{else}\ \cdots\ \mathsf{if}\ \delta_{m_{n_k}}\ \mathsf{then}\ \pi(m_{n_k})$.

We here consider only the (more complex) case $k > 1$. Our goal is to prove $M(n) \models [\pi(n)]_\alpha\, \phi_g$, that is, $M(n) \models [\mathcal{E}(n, m); \pi(m)]_\alpha\, \phi_g$. By the induction hypothesis we have $M(m_{n_i}) \models \left[\pi(m_{n_i})\right]_\alpha \phi_g$ for all $i = 1, \dots, k$ (the m_{n_i} are of lower height than n).

Claim 1. $M(m_{n_i}) \models [\pi(m)]_\alpha\, \phi_g$ for all $i = 1, \dots, k$.

Proof of claim. Let i be given. We need to prove

$$M(m_{n_i}) \models \left[\mathsf{if}\ \delta_{m_{n_1}}\ \mathsf{then}\ \pi(m_{n_1})\ \mathsf{else}\ \cdots\ \mathsf{if}\ \delta_{m_{n_k}}\ \mathsf{then}\ \pi(m_{n_k})\right]_\alpha \phi_g.$$

Note that by using item 5 of Lemma 3.17 it suffices to prove that for all $j = 1, \dots, k$,

$$M(m_{n_i}) \models \delta_{m_{n_j}} \text{ implies } M(m_{n_i}) \models \left[\pi(m_{n_j})\right]_\alpha \phi_g. \tag{3.3}$$

Let $j \in \{1, \dots, k\}$ be chosen arbitrarily. Assume first $j = i$. By induction hypothesis we have $M(m_{n_j}) \models \left[\pi(m_{n_j})\right]_\alpha \phi_g$, and hence $M(m_{n_i}) \models \left[\pi(m_{n_j})\right]_\alpha \phi_g$. From this (3.3) immediately follows. Assume now $j \neq i$. By the construction of the δ-formulas, either $M(m_{n_j}) \equiv M(m_{n_i})$ or $M(m_{n_i}) \not\models \delta_{m_{n_j}}$. In the latter case, (3.3) holds trivially. In case of $M(m_{n_j}) \equiv M(m_{n_i})$ we immediately get $M(m_{n_i}) \models \left[\pi(m_{n_j})\right]_\alpha \phi_g$, since by induction hypothesis we have $M(m_{n_j}) \models \left[\pi(m_{n_j})\right]_\alpha \phi_g$. This concludes the proof of the claim.

Note that by definition of the tree expansion rule (Definition 3.27), $M(m_1), \dots, M(m_l)$ are the information cells in $M(m)$.

Claim 2. The following holds:

- If $\alpha = s$ (w), then for every (some) information cell M' in $M(m)$: $M' \models [\pi(m)]_\alpha \phi_g$.
- If $\alpha = sp$ (wp), then for every (some) most plausible information cell M' in $M(m)$: $M' \models [\pi(m)]_\alpha \phi_g$.

Proof of claim. We only consider the most complex cases, $\alpha = sp$ and $\alpha = wp$. First consider $\alpha = sp$. Let M' be a most plausible information cell in $M(m)$. We need to prove $M' \models [\pi(m)]_\alpha \phi_g$. Since, as noted above, $M(m_1), \dots, M(m_l)$ are the information cells in $M(m)$, we must have $M' = M(m_i)$ for some $i \in \{1, \dots, l\}$. Furthermore, as M' is among the most plausible information cells in $M(m)$, m_i must by definition be a most plausible child of m. Definition 3.30 then gives us that m_i is α-solved. Thus $m_i = m_{n_j}$ for some $j \in \{1, \dots, k\}$. By Claim 1 we have $M(m_{n_j}) \models [\pi(m)]_\alpha \phi_g$, and since $M' = M(m_i) = M(m_{n_j})$ this gives the desired conclusion. Now consider the case $\alpha = wp$. Definition 3.30 gives us that at least one of the most plausible children of m are α-solved. By definition, this must be one of the m_{n_i}, $i \in \{1, \dots, k\}$. Claim 1 gives $M(m_{n_i}) \models [\pi(m)]_\alpha \phi_g$. Since m_{n_i} is a most plausible child of m, we must have that $M(m_{n_i})$ is among the most plausible information cells in $M(m)$. Hence we have proven that $[\pi(m)]_\alpha \phi_g$ holds in a most plausible information cell of $M(m)$.

By definition of the tree expansion rule (Definition 3.27), $M(m) = M(n) \otimes \mathcal{E}(n,m)$. Thus we can replace $M(m)$ by $M(n) \otimes \mathcal{E}(n,m)$ in Claim 2 above. Using items 1–4 of Lemma 3.17, we immediately get from Claim 2 that independently of α the following holds: $M(n) \models [\mathcal{E}(n,m)]_\alpha [\pi(m)]_\alpha \phi_g$ (the condition $M(n) \models \langle \mathcal{E}(n,m) \rangle \top$ holds trivially by the tree expansion rule). From this we can then finally conclude $M(n) \models [\mathcal{E}(n,m); \pi(m)]_\alpha \phi_g$, as required. □

Theorem 3.34 (Completeness). *Let α be one of s, w, sp or wp. If there is an α-solution to the planning problem $\mathcal{P} = (M_0, \mathsf{A}, \phi_g)$, then an α-planning tree T for $\mathcal{P}$ can be constructed, such that $root(T)$ is α-solved.*

Proof. First note that we have $[\mathsf{skip}; \pi]_\alpha \phi_g = [\mathsf{skip}]_\alpha ([\pi]_\alpha \phi_g) = [\pi]_\alpha \phi_g$. Thus, we can without loss of generality assume that no plan contains a subexpression of the form $\mathsf{skip}; \pi$. The length of a plan π, denoted $|\pi|$, is defined recursively by: $|\mathsf{skip}| = 1$; $|\mathcal{E}| = 1$; $|\mathsf{if}\ \phi\ \mathsf{then}\ \pi_1\ \mathsf{else}\ \pi_2| = |\pi_1| + |\pi_2|$; $|\pi_1; \pi_2| = |\pi_1| + |\pi_2|$.

Claim 1. Let π be an α-solution to $\mathcal{P} = (M_0, \mathsf{A}, \phi_g)$ with $|\pi| \geq 2$. Then there exists an α-solution of the form $\mathcal{E}; \pi'$ with $|\mathcal{E}; \pi'| \leq |\pi|$.

Proof of claim. Proof by induction on $|\pi|$. The base case is $|\pi| = 2$. We have two cases, $\pi = \mathsf{if}\ \phi\ \mathsf{then}\ \pi_1\ \mathsf{else}\ \pi_2$ and $\pi = \pi_1; \pi_2$, both with $|\pi_1| = |\pi_2| = 1$. If π is the latter, it already has desired the form. If $\pi = \mathsf{if}\ \phi\ \mathsf{then}\ \pi_1\ \mathsf{else}\ \pi_2$ then, by assumption on π, $M_0 \models \left[\mathsf{if}\ \phi\ \mathsf{then}\ \pi_1\ \mathsf{else}\ \pi_2\right]_\alpha \phi_g$. Item 5 of Lemma 3.17 now gives that $M_0 \models \phi$ implies $M_0 \models [\pi_1]_\alpha \phi_g$ and $M_0 \not\models \phi$ implies $M_0 \models [\pi_2]_\alpha \phi_g$. Thus we must either have $M_0 \models [\pi_1]_\alpha \phi_g$ or $M_0 \models [\pi_2]_\alpha \phi_g$, that is, either π_1 or π_2 is an α-solution to $\mathcal{P}$. Thus either $\pi_1;\mathsf{skip}$ or $\pi_2;\mathsf{skip}$ is an α-solution to $\mathcal{P}$, and both of these have length $|\pi|$. This completes the base case. For the induction step, consider a plan π of length $l > 2$ which is an α-solution to $\mathcal{P}$. We again have two cases to consider, $\pi = \mathsf{if}\ \phi\ \mathsf{then}\ \pi_1\ \mathsf{else}\ \pi_2$ and $\pi = \pi_1;\pi_2$. If $\pi = \pi_1;\pi_2$ is an α-solution to $\mathcal{P}$, then π_1 is an α-solution to the planning problem $\mathcal{P}' = (M_0, \mathsf{A}, [\pi_2]_\alpha \phi_g)$, as $M_0 \models [\pi_1;\pi_2]_\alpha \phi_g \Leftrightarrow M_0 \models [\pi_1]_\alpha [\pi_2]_\alpha \phi_g$. Clearly $|\pi_1| < l$, so the induction hypothesis gives that there is an α-solution $(\mathcal{E};\pi_1')$ to $\mathcal{P}'$, with $|\mathcal{E};\pi_1'| \leq |\pi_1|$. Then, $\mathcal{E};\pi_1';\pi_2$ is an α-solution to $\mathcal{P}$ and we have $|\mathcal{E};\pi_1';\pi_2| = |\mathcal{E};\pi_1'| + |\pi_2| \leq |\pi_1| + |\pi_2| = |\pi|$. If $\pi = \mathsf{if}\ \phi\ \mathsf{then}\ \pi_1\ \mathsf{else}\ \pi_2$ is an α-solution to $\mathcal{P}$, then we can as above conclude that either π_1 or π_2 is an α-solution to $\mathcal{P}$. With both $|\pi_1| < l$ and $|\pi_2| < l$, the induction hypothesis gives the existence an α-solution $\mathcal{E};\pi'$, with $|\mathcal{E};\pi'| \leq |\pi|$. This completes the proof of the claim.

We now prove the theorem by induction on $|\pi|$, where π is an α-solution to $\mathcal{P} = (M_0, \mathsf{A}, \phi_g)$. We need to prove that there exists an α-planning tree for $\mathcal{P}$ in which the root is α-solved. Let T_0 denote the planning tree for $\mathcal{P}$ only consisting of its root node with label M_0. The base case is when $|\pi| = 1$. Here, we have two cases, $\pi = \mathsf{skip}$ and $\pi = \mathcal{E}$. In the first case, the planning tree T_0 already has its root α-solved, since $M_0 \models [\mathsf{skip}]_\alpha \phi_g \Leftrightarrow M_0 \models \phi_g$. In the second case, $\pi = \mathcal{E}$, we have $M_0 \models [\mathcal{E}]_\alpha \phi_g$ as $\pi = \mathcal{E}$ is an α-solution to $\mathcal{P}$. By definition, this means that $\mathcal{E}$ is applicable in M_0, and we can apply the tree expansion rule to T_0, which will produce:

1. A child m of the root node with $M(m) = M_0 \otimes \mathcal{E}$.
2. Children $m_1, \ldots, m_l$ of m, where $M(m_1), \ldots, M(m_l)$ are the information cells of $M(m)$.

Call the expanded tree T_1. Since $M_0 \models [\mathcal{E}]_\alpha \phi_g$, Lemma 3.17 implies that for every/some/every most plausible/some most plausible information cell M' in $M_0 \otimes \mathcal{E}$, $M' \models \phi_g$ (where $\alpha = s/w/sp/wp$). Since $M(m_1), \ldots, M(m_l)$ are the information cells of $M_0 \otimes \mathcal{E}$, we can conclude that every/some/every most plausible/some most plausible child of m is α-solved. Hence also m and thus n are α-solved. The base is hereby completed.

For the induction step, let π be an α-solution to $\mathcal{P}$ with length $l > 1$. Let T_0 denote the planning tree for $\mathcal{P}$ consisting only of its root node with label M_0. By Claim 1, there exists an α-solution to $\mathcal{P}$ of the form $\mathcal{E};\pi'$ with $|\mathcal{E};\pi'| \leq |\pi|$. As $M_0 \models [\mathcal{E};\pi']_\alpha \phi_g \Leftrightarrow M_0 \models [\mathcal{E}]_\alpha [\pi']_\alpha \phi_g$, $\mathcal{E}$ is applicable in M_0. Thus, as in the base case, we can apply the tree expansion rule to T_0 which will produce nodes as in 1 and 2 above. Call the expanded tree T_1. Since $M_0 \models [\mathcal{E}]_\alpha [\pi']_\alpha \phi_g$, items

1–4 of Lemma 3.17 implies that for every/some/every most plausible/some most plausible information cell in $M_0 \otimes \mathcal{E}$, $[\pi']_\alpha \phi_g$ holds. Hence, for every/some/every most plausible/some most plausible child m_i of m, $M(m_i) \models [\pi']_\alpha \phi_g$. Let $m_{n_1}, \ldots, m_{n_k}$ denote the subsequence of $m_1, \ldots, m_l$ consisting of the children of m for which $M(m_{n_i}) \models [\pi']_\alpha \phi_g$. Then, by definition, π' is an α-solution to each of the planning problem $\mathcal{P}_i = (M(m_{n_i}), \mathsf{A}, \phi_g)$, $i = 1, \ldots, k$. As $|\pi'| < |\mathcal{E}; \pi'| \leq l$, the induction hypothesis gives that α-planning trees T_i' with α-solved roots can be constructed for each $\mathcal{P}_i$. Let T_2 denote T_1 expanded by adding each planning tree T_i' as the subtree rooted at M_{n_i}. Then each of the nodes m_{n_i} are α-solved in T, and in turn both m and $root(T_2)$ are α-solved. The final thing we need to check is that T_2 has been correctly constructed according to the tree expansion rule, more precisely, that condition $\mathcal{B}_\alpha$ has not been violated. Since each T_i' has in itself been correctly constructed in accordance with $\mathcal{B}_\alpha$, the condition can only have been violated if for one of the non-leaf OR-nodes m' in one of the T_i's, $M(m') \equiv M(root(T_2))$. We can then replace the entire planning tree T_2 by a (node-wise modally equivalent) copy of the subtree rooted at m', and we would again have an α-planning tree with an α-solved root. □

3.4.3 Planning Algorithm

In the following, let $\mathcal{P}$ denote any planning problem, and α be one of s, w, sp or wp. With all the previous in place, we now have an algorithm for synthesising an α-solution to $\mathcal{P}$, given as follows.

```
PLAN($\alpha, \mathcal{P}$)
  1 Let $T$ be the $\alpha$-planning tree only consisting of $root(T)$ labelled by the initial
    state of $\mathcal{P}$.
  2 Repeatedly apply the tree expansion rule of $\mathcal{P}$ to $T$ until no more rules apply
    satisfying condition $\mathcal{B}_\alpha$.
  3 If $root(T)$ is $\alpha$-solved, return $\pi(root(T))$, otherwise return FAIL.
```

Theorem 3.35. PLAN$(\alpha, \mathcal{P})$ *is a terminating, sound and complete algorithm for producing α-solutions to planning problems $\mathcal{P}$. Soundness means that if* PLAN$(\alpha, \mathcal{P})$ *returns a plan, it is an α-solution to $\mathcal{P}$. Completeness means that if $\mathcal{P}$ has an α-solution,* PLAN$(\alpha, \mathcal{P})$ *will return one.*

Proof. Termination comes from Lemma 3.28 (with $\mathcal{B}$ replaced by the stronger condition $\mathcal{B}_\alpha$), soundness from Theorem 3.33 and completeness from Theorem 3.34 (given any two $\mathcal{B}_\alpha$-saturated α-planning trees T_1 and T_2 for the same planning problem, the root node of T_1 is α-solved iff the root node of T_2 is). □

With PLAN($\alpha, \mathcal{P}$) we have given an algorithm for solving α-parametrised planning problems. The α parameter determines the strength of the synthesised plan π, cf. Lemma 3.20. Whereas the cases of weak ($\alpha = w$) and strong ($\alpha = s$) plans have been the subject of much research, the generation of weak plausibility ($\alpha = wp$) and strong plausibility ($\alpha = sp$) plans based on pre-encoded beliefs is a novelty of this paper. Plans taking plausibility into consideration have several advantages. Conceptually, the basement scenario as formalised by $\mathcal{P}_B$ (cf. Example 3.19) allowed for several weak solutions (with the shortest one being hazardous to the agent) and *no* strong solutions. In this case, the synthesised strong plausibility solution corresponds to the course of action a rational agent (mindful of her beliefs) should take. There are also computational advantages. An invocation of PLAN($sp, \mathcal{P}$) will expand at most as many nodes as an invocation of PLAN($s, \mathcal{P}$) before returning a result (assuming the same order of tree expansions). As plausibility plans only consider the most plausible information cells, we can prune non-minimal information cells during plan search.

We also envision using this technique in the context of an agent framework where planning, acting and execution monitoring are interleaved.[8] Let us consider the case of strong plausibility planning ($\alpha = sp$). From some initial situation an sp-plan is synthesised which the agent starts executing. If reaching a situation that is not covered by the plan, she restarts the process from this point; i.e. she *replans*. Note that the information cell to replan from is present in the tree as a sibling of the most plausible information cell(s) expected from executing the last action. Such replanning mechanisms allow for the *repetition* of actions necessary in some planning problems with cyclic solutions.

We return one last time to the basement problem and consider a modified replace action such that the replacement light bulb might, though it is unlikely, be broken. This means that there is no strong solution. Executing the sp-solution flick; desc, she would replan after flick if that action didn't have the effect of turning on the light. A strong plausibility solution from this point would then be flick; replace; flick; desc.

3.5 Related and Future Work

In this paper we have presented α-solutions to planning problems incorporating ontic, epistemic and doxastic notions. The cases of $\alpha = sp/sw$ are, insofar as we are aware, novel concepts not found elsewhere in the literature. Our previous paper [Andersen et al., 2012] concerns the cases $\alpha = s/w$, so that framework deals

[8] Covering even more mechanisms of agency is *situated planning* [Ghallab et al., 2004].

only with *epistemic* planning problems without a doxastic component. Whereas we characterise solutions as formulas, [Bolander and Andersen, 2011] takes a semantic approach to strong solutions for epistemic planning problems. In their work plans are sequences of actions, requiring conditional choice of actions at different states to be encoded in the action structure itself. By using the $\mathcal{L}(P, \mathsf{A})$ we represent this choice explicitly.

The meaningful plans of [de Lima, 2007, chap. 2] are reminiscent of the work in this paper. Therein, plan verification is cast as validity of an EDL-consequence in a given system description. Like us, they consider single-agent scenarios, conditional plans, applicability and incomplete knowledge in the initial state. Unlike us, they consider only deterministic epistemic actions (without plausibility). In the multi-agent treatment [de Lima, 2007, chap. 4], action laws are translated to a fragment of DEL with only public announcements and public assignments, making actions singleton event models. This means foregoing nondeterminism and therefore sensing actions.

Epistemic planning problems in [Löwe et al., 2011a] are solved by producing a sequence of pointed epistemic event models where an external variant of applicability (called *possible at*) is used. Using such a formulation means outcomes of actions are fully determined, making conditional plans and weak solutions superfluous. As noted by the authors, and unlike our framework, their approach does not consider factual change. We stress that [Bolander and Andersen, 2011, Löwe et al., 2011a, de Lima, 2007] all consider the multi-agent setting which we have not treated here.

In our work so far, we haven't treated the problem of where domain formulations come from, assuming just that they are given. Standardised description languages are vital if modal logic-based planning is to gain wide acceptance in the planning community. Recent work worth noting in this area includes [Baral et al., 2012], which presents a specification language for the multi-agent belief case.

As suggested by our construction of planning trees, there are several connections between our approach for $\alpha = s$ and two-player imperfect information games. First, product updates imply perfect recall [van Benthem, 2001]. Second, when the game is at a node belonging to an information set, the agent knows a proposition only if it holds throughout the information set. Finally, the strong solutions we synthesise are very similar to mixed strategies. A strong solution caters to any information cell (contingency) it may bring about, by selecting exactly one sub-plan for each [Aumann and Hart, 1992].

Our work relates to [Ghallab et al., 2004], where the notions of strong and weak solutions are found, but without plausibilites. Their belief states are sets of states which may be partioned by observation variables. The framework in [Rintanen, 2004] describes strong conditional planning (prompted by nondeterministic actions) with partial observability modelled using a fixed set of observable state variables. Our partition of plausibility models into information cells follows straight from the definition of product update. A clear advantage in our approach is that *actions* readily encode both nondetermism and partial observability. [Jensen, 2013a] shows that the *strong plan existence problem* for the framework in [Andersen et al., 2012] is 2-EXP-complete. In our formulation, $\textsc{Plan}(s, \mathcal{P})$ answers the same question for $\mathcal{P}$ (it gives a strong solution if one exists), though with a richer modal language.

We would like to do plan verification and synthesis in the multi-agent setting. We believe that generalising the notions introduced in this paper to multi-pointed plausibility and event models are key. Plan synthesis in the multi-agent setting is undecidable [Bolander and Andersen, 2011], but considering restricted classes of actions as is done in [Löwe et al., 2011a] seems a viable route for achieving decidable multi-agent planning. Other ideas for future work include replanning algorithms and learning algorithms where plausibilities of actions can be updated when these turn out to have different outcomes than expected.

Acknowledgements

For valuable comments at various stages of the work presented in this article, we would like to extend our gratitude to the following persons: Patrick Allo, Alexandru Baltag, Johan van Benthem, Hans van Ditmarsch, Jens Ulrik Hansen, Sonja Smets and the anonymous reviewers.

Chapter 4

Bisimulation and expressivity for conditional belief, degrees of belief, and safe belief

This chapter is the first printing of joint work with Thomas Bolander, Martin Holm Jensen and Hans van Ditmarsch. It contains the multi-agent version of the bisimulation published in [Andersen et al., 2013] plus expressivity results for the logics under scrutiny.

Bisimulation and expressivity for conditional belief, degrees of belief, and safe belief

Mikkel Birkegaard Andersen Thomas Bolander
Hans van Ditmarsch Martin Holm Jensen

Abstract

Plausibility models are Kripke models that agents use to reason about knowledge and belief, both of themselves and of each other. Such models are used to interpret the notions of conditional belief, degrees of belief, and safe belief. The logic of conditional belief contains that modality and also the knowledge modality, and similarly for the logic of degrees of belief and the logic of safe belief. With respect to these logics, plausibility models may contain too much information. A proper notion of bisimulation is required that characterizes them. We define that notion of bisimulation and prove the required characterizations: on the class of image-finite and preimage-finite models (with respect to the plausibility relation), two pointed Kripke models are modally equivalent in either of the three logics, if and only if they are bisimilar. As a result, the information content of such a model can be similarly expressed in the logic of conditional belief, or the logic of degrees of belief, or that of safe belief. This, we found a surprising result. Still, that does not mean that the logics are equally expressive: the logic of conditional belief and the logic of degrees of belief are incomparable and both are less expressive than the logic of safe belief. In view of the result on bisimulation characterization, this is an equally surprising result. We hope our insights may contribute to the growing community of formal epistemology and on the relation between qualitative and quantitative modelling.

4.1 Introduction

A typical approach in belief revision involves preferential orders to express degrees of belief and knowledge [Kraus et al., 1990, Meyer et al., 2000]. This goes back to the 'systems of spheres' in [Lewis, 1973, Grove, 1988]. Dynamic doxastic logic was proposed and investigated in [Segerberg, 1998] in order to provide a link between the (non-modal logical) belief revision and modal logics with explicit knowledge and belief operators. A similar approach was pursued in belief revision in dynamic epistemic logic [Aucher, 2005, van Ditmarsch, 2005, van Benthem, 2007, Baltag and Smets, 2008b, van Ditmarsch and Labuschagne, 2007], that continues to develop strongly [Britz and Varzinczak, 2013, van Benthem, 2011]. We focus on the proper notion of structural equivalence on models encoding knowledge and belief simultaneously. A prior investigation into that is [Demey, 2011], which we relate our results to at the end of the paper. Our motivation is to find suitable structural notions to reduce the complexity of planning problems. Such plans are sequences of actions, such as iterated belief revision. It is the dynamics of knowledge and belief that, after all, motivates our research.

The semantics of belief depends on the structural properties of models. To relate the structural properties of models to a logical language we need a notion of structural similarity, known as bisimulation. A bisimulation relation relates a modal operator to an accessibility relation. Plausibility models do not have an accessibility relation as such but a plausibility relation. This induces a set of accessibility relations: the *most plausible* worlds are the *accessible* worlds for the modal belief operator; and the *plausible* worlds are the *accessible* worlds for the modal knowledge operator. But it contains much more information: to each modal operator of conditional belief (or of degree of belief) one can associate a possibly distinct accessibility relation.This raises the question of how to represent the bisimulation conditions succinctly. Can this be done by reference to the plausibility relation directly, instead of by reference to these, possibly many, induced accessibility relations? It is now rather interesting to observe that relative to the modalities of knowledge and belief, the plausibility relation is already in some way too rich.

The plausibility model M_L on the left in Figure 4.1 consists of five worlds. The proposition p is true in the top ones and false in the bottom ones. The reverse holds for q: true at the bottom and false at the top. The a relations in the model correspond to the plausibility order $w_3 >_a w_2 >_a w_1$, interpreted such that the smaller of two elements in the order is the most plausible of the two. Further, everything that is comparable with the plausibility order is considered epistemically possible. We can then view the model as a standard multi-agent $S5$ plus an ordering on the epistemic possibilities. As w_1 is the most plausible world for a, she believes p and that b believes $\neg p \wedge q$. This works differently from the usual doxastic modal logic, where belief corresponds to the accessibility relation. In the logics of belief

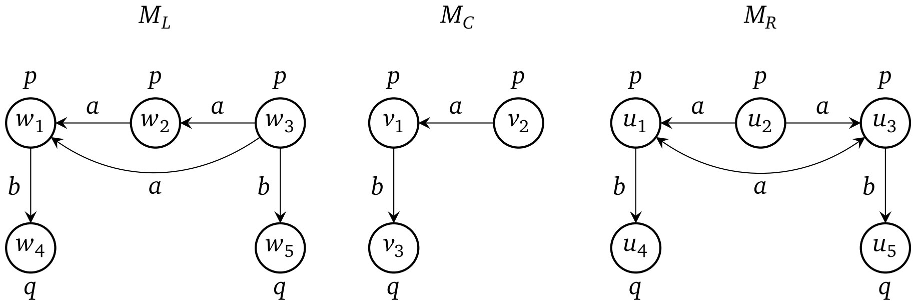

Figure 4.1: An arrow $x \to y$ labelled by a means $x \geq_a y$; agent a considers y at least as plausible as x. We use $x >_a y$ to mean $x \geq_a y$ and $y \not\geq_a x$. Here $w_2 >_a w_1$, so w_1 is strictly more plausible than w_2. Reflexive edges are omitted. Unlisted propositions are false.

that we study, belief is what holds in the most plausible world(s) in an epistemic equivalence class. For a the most plausible world in all three p-worlds is w_1, so a believes the same formulas in all of them.

In w_2 agent b knows p. If a is given the information that b does not consider q possible (that is, the information that neither w_1 nor w_3 is the actual world), then a believes that b knows p – or conditional on $K_b\neg q$, a believes $K_b p$. Such a statement is an example of the logic of conditional belief L^C defined in Section 4.3. In L^C we write this statement as $B_a^{K_b\neg q} K_b p$.

Now examine w_3. You might not know it yet, but w_1 and w_3 are modally equivalent for L^C: they agree on all formulas of that language – no information expressible in L^C distinguishes the two worlds. This leads to the observation that no matter where we move w_3 in the plausibility ordering for a, modal equivalence is preserved. Similarly, we can move w_2 anywhere we like *except* making it more plausible than w_1. If we did, then a would believe $K_b p$ unconditionally, and the the formulas true in the model would have been changed.

It turns out that moving worlds about in the plausibility order can be done for all models, as long as we obey one (conceptually) simple rule: Grouping worlds into “modal equivalence classes” of worlds modally equivalent to each other, we are only required to preserve the ordering between the *most* plausible worlds in each modal equivalence class. *Only the most plausible world in each class matters.*

Another crucial observation is that standard bisimulation in terms of $\geq_a$ does not give correspondence between bisimulation and modal equivalence. For instance, while w_1 and w_3 are modally equivalent, they are not “standardly” bisimilar with respect to $\geq_a$: w_3 has a $\geq_a$-edge to a $K_b p$ world (w_2), whereas w_1 does not. Thus, the straight-forward, standard definition of bisimulation does not work, because there is no modality corresponding to the plausibility relation itself. Instead we have an infinite set of modalities corresponding to relations derived from the plausibility relation. One of the major contributions of this paper is a solution to exactly this problem.

Making w_3 as plausible as w_1 and appropriately renaming worlds gets us M_R of Figure 4.1. Here the modally equivalent worlds u_1 and u_3 are equally plausible, modally equivalent *and* standardly bisimilar. This third observation gives a sense of how we solve the problem generally. Rather than using $\geq_a$ directly, our definition of bisimulation checks accessibility with respect to a relation $\geq_a^R$ derived from $\geq_a$ and the bisimulation relation R itself. Postponing details for later we just note that in the present example the derived relation for M_L is exactly the plausibility relation for M_R. This indicates what we later prove: This new derived relation reestablishes the correspondence between bisimilarity and modal equivalence.

The model M_C of Figure 4.1 is the bisimulation contraction of the right model using standard bisimilarity. It is the bisimulation contraction of both models with the bisimulation notion informally defined in the previous paragraph. In previous work on planning with single-agent plausibility models [Andersen et al., 2014], finding contractions of plausibility models is needed for decidability and complexity results. In this paper we do this for the first time for multi-agent plausibility models, opening new vistas in applications of modal logic to automated planning.

Overview of content In Section 4.2 we introduce plausibility models and the proper and novel notion of bisimulation on these models, and prove various properties of bisimulation. In Section 4.3 we define the three logics of conditional belief, degrees of belief, and safe belief, and provide some further historical background on these logics. In Section 4.4 we demonstrate that bisimilarity corresponds to logical equivalence (on image-finite and preimage-finite models) for all three core logics, so that, somewhat remarkably, one could say that the content of a given model can equally well be described in any of these logics. Then, in Section 4.5 we determine the relative expressivity of the three logics, including more expressive combinations of their primitive modalities. The main result here is that the logics of conditional and degrees of belief are incomparable, and that the logics of degrees of belief and safe belief are incomparable, but that the logic of safe belief is (strictly) more expressive than the logic of conditional belief. In Section 4.6, we put our result in the perspective of other recent investigations, mainly the study by Lorenz Demey [Demey, 2011], and in the perspective of possible applications: decidable planning.

4.2 Plausibility models and bisimulation

A *well-preorder* on a set X is a reflexive and transitive binary relation $\trianglerighteq$ on X such that every non-empty subset has $\trianglerighteq$-minimal elements. The set of *minimal elements* (for $\trianglerighteq$) of some $Y \subseteq X$ is the set $Min_{\trianglerighteq} Y$ defined as $\{y \in Y \mid y' \trianglerighteq y \text{ for all } y' \in Y\}$.[1] As any two-element subset $Y = \{x, y\}$ of X also has minimal elements, we have that $z \trianglerighteq y$ or $y \trianglerighteq z$. Thus all elements in X are $\trianglerighteq$-comparable. Given subsets $Y, Z \subseteq X$, we define $Y \trianglerighteq Z$ if and only if $y \trianglerighteq z$ for all $y \in Y$ and all $z \in Z$.

Given any binary relation R on X, we use $R^{=}$ to denote the reflexive, symmetric, and transitive closure of R (the equivalence closure of R). For any equivalence relation R on X, we write $[x]_R$ for $\{x' \in X \mid (x, x') \in R\}$. A binary relation R on X is *image-finite* if and only if for every $x \in X$, $\{x' \in X \mid (x, x') \in R\}$ is finite. A relation is *preimage-finite* if and only if for every $x \in X$, $\{x' \in X \mid (x', x) \in R\}$ is finite. We say R is *(pre)image-finite* if it is both image-finite and preimage-finite.

Definition 4.1 (Plausibility model). A *plausibility model* for a countably infinite set of propositional symbols P and a finite set of agents A is a tuple $M = (W, \geq, V)$, where

- W is a set of *worlds* called the *domain*, denoted $D(W)$;
- $\geq: A \to \mathcal{P}(W \times W)$ is a plausibility function, such that for each $a \in A$, $\geq(a)$ is a set of mutually disjoint well-preorders covering W, called the *plausibility relation* (for agent a);
- $V : W \to 2^P$ is a *valuation*.

For $w \in W$, (M, w) is a *pointed plausibility model*.

For $\geq(a)$ we write $\geq_a$. If $w \geq_a v$ then v is *at least as plausible* as w (for agent a), and the $\geq_a$-minimal elements are the *most plausible* worlds. For the symmetric closure of $\geq_a$ we write $\sim_a$: this is an equivalence relation on W called the *epistemic relation* (for agent a). If $w \geq_a v$ but $v \not\geq_a w$ we write $w >_a v$ (v is *more plausible* than w), and for $w \geq_a v$ and $v \geq_a w$ we write $w \simeq_a v$ (w and v are *equiplausible*). Instead of $w \geq_a v$ ($w >_a v$) we may write $v \leq_a w$ ($v <_a w$).

We now proceed to define a notion of autobisimulation on a plausibility model. This notion is non-standard, because there is no one-to-one relation between the plausibility relation for an agent and a modality for that agent in the logics defined

[1] This notion of minimality is non-standard and taken from [Baltag and Smets, 2008b]. Usually a minimal element of a set is an element that is not greater than any other element.

later. In the definition below (and from now on), we allow ourselves some further notational abbreviations. Let $M = (W, \geq, V)$ denote a plausibility model. Let $a \in A$ and $w \in W$, then we write $[w]_a$ instead of $[w]_{\sim_a}$. Now let $Z \subseteq [w]_a$, then we write $Min_a Z$ instead of $Min_{\geq_a} Z$. For any binary relation R on W, we write $w \geq_a^R v$ for $Min_a([w]_{R^=} \cap [w]_a) \geq_a Min_a([v]_{R^=} \cap [v]_a)$. When $w \geq_a^R v$ and $v \geq_a^R w$, we write $w \simeq_a^R v$.

Definition 4.2 (Autobisimulation). Let $M = (W, \geq, V)$ be a plausibility model. An *autobisimulation* on M is a non-empty relation $R \subseteq W \times W$ such that for all $(w, w') \in R$ and for all $a \in A$:

[atoms] $V(w) = V(w')$;

[forth$_\geq$] If $v \in W$ and $w \geq_a^R v$, there is a $v' \in W$ such that $w' \geq_a^R v'$ and $(v, v') \in R$;

[back$_\geq$] If $v' \in W$ and $w' \geq_a^R v'$, there is a $v \in W$ such that $w \geq_a^R v$ and $(v, v') \in R$;

[forth$_\leq$] If $v \in W$ and $w \leq_a^R v$, there is a $v' \in W$ such that $w' \leq_a^R v'$ and $(v, v') \in R$;

[back$_\leq$] If $v' \in W$ and $w' \leq_a^R v'$, there is a $v \in W$ such that $w \leq_a^R v$ and $(v, v') \in R$.

A *total autobisimulation* on M is an autobisimulation with W as both domain and codomain.

Our bisimulation relation is non-standard in the [back] and [forth] clauses. A standard [forth] condition based on an accessibility relation $\geq_a$ would be: If $v \in W$ and $w \geq_a v$, there is a $v' \in W'$ such that $w' \geq_a v'$ and $(v, v') \in R$. Here, R only appears in the part '$(v, v') \in R$'. But in the definition of autobisimulation for plausibility models, in [forth$_\geq$], the relation R also features in the condition for applying [forth$_\geq$] and in its consequent, namely as the upper index in $w \geq_a^R v$ and $w' \geq_a^R v'$. This means that R also determines which v are accessible from w, and which v' are accessible from w'. This explains why we define an autobisimulation on a single model before a bisimulation between distinct models: We need the bisimulation relation R to determine the plausibility relation $\geq_a^R$ from the plausibility relation $\geq_a$ on any given model first, before structurally comparing distinct models.

Example 4.3. The models M_L and M_R of Figure 4.1 are reproduced in Figure 4.2. Consider the relation $R = R_{id} \cup \{(w_1, w_3), (w_3, w_1), (w_4, w_5), (w_5, w_4)\}$, where R_{id} is the identity relation on W. With this R, we get that w_1 and w_3 are equiplausible for $\geq_a^R$:

$$\begin{gathered} w_1 \simeq_a^R w_3 \text{ iff} \\ Min_a([w_1]_{R^=} \cap [w_1]_a) \simeq_a Min_a([w_3]_{R^=} \cap [w_3]_a) \text{ iff} \\ Min_a\{w_1, w_3\} \simeq_a Min_a\{w_1, w_3\} \text{ iff} \\ w_1 \simeq_a w_1 \end{gathered}$$

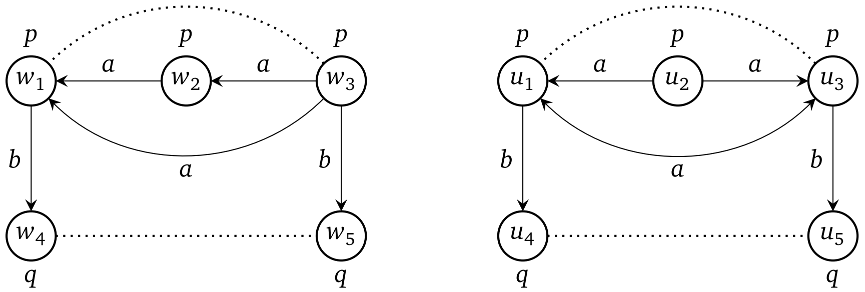

Figure 4.2: The left and right models of Figure 4.1, with the dotted lines showing the maximal autobisimulations (modulo reflexivity).

We also get that $w_2 \geq_a^R w_3$:

$$\begin{aligned} w_2 &\geq_a^R w_3 \text{ iff} \\ Min_a([w_2]_{R^=} \cap [w_2]_a) &\geq_a Min_a([w_3]_{R^=} \cap [w_3]_a) \text{ iff} \\ Min_a\{w_2\} &\geq_a Min_a\{w_1, w_3\} \text{ iff} \\ w_2 &\geq_a w_1 \end{aligned}$$

This gives $\geq_a^R = \{(w_1, w_3), (w_3, w_1), (w_2, w_3), (w_2, w_1)\} \cup R_{id}$. For b, we get $\geq_b^R = \geq_b$. The autobisimulation R on M_L is shown in Figure 4.2. It should be easy to check that R is indeed an autobisimulation. To help, we'll justify why (w_4, w_5) is in R: For $\geq_b^R$, we have that, as $(w_1, w_3) \in R$ and $w_1 \geq_b^R w_4$, there must be a world v such that $w_3 \geq_b^R v$ and $(w_4, v) \in R$. This v is w_5.

Note that R is the maximal autobisimulation. Based on [atoms] there are only two possible candidate pairs that could potentially be added to R (modulo symmetry), namely (w_1, w_2) and (w_2, w_3). But w_2 does not have a b-edge to a q world, whereas both w_1 and w_3 do. There is therefore nothing more to add.

The maximal autobisimulation for M_R is completely analogous, as shown in Figure 4.2. ■

Lemma 4.4. *Let $M = (W, \geq, V)$ and R be a binary relation on W. If $(w, w') \in R^=$ and $w \sim_a w'$ then $w \simeq_a^R w'$.*

Proof. From $(w, w') \in R^=$ and $w \sim_a w'$ we get $[w]_{R^=} = [w']_{R^=}$ and $[w]_a = [w']_a$ and hence $[w]_{R^=} \cap [w]_a = [w']_{R^=} \cap [w']_a$. Thus also $Min_a([w]_{R^=} \cap [w]_a) = Min_a([w']_{R^=} \cap [w']_a)$, immediately implying $w \simeq_a^R v$. □

Lemma 4.5. *Let $\geq_a$ be a set of mutually disjoint well-preorders covering a plausibility model $M = (W, \geq, V)$ and let R be a binary relation on M. Then $\geq_a^R$ is a set of well-*

preorders inducing the same partition, that is, $\geq_a^R$ partitions W into a well-preorder on each $\sim_a$-equivalence class.

Proof. The relation $\geq_a$ partitions W into a well-preorder on each $\sim_a$-equivalence class. We need to show that $\geq_a^R$ does the same. Hence we need to prove: 1) $\geq_a^R$ is reflexive; 2) $\geq_a^R$ is transitive; 3) any $\sim_a$-equivalence class has $\geq_a^R$-minimal elements; 4) if two worlds are related by $\geq_a^R$ they are also related by $\sim_a$.

Reflexivity of is trivial. *Transitivity*: Let $(w,v),(v,u) \in \geq_a^R$. Then $Min_a([w]_{R^=} \cap [w]_a) \geq_a Min_a([v]_{R^=} \cap [v]_a)$, and $Min_a([v]_{R^=} \cap [v]_a) \geq_a Min_a([u]_{R^=} \cap [u]_a)$. Using that for any sets X,Y,Z, if $X \geq_a Y$ and $Y \geq_a Z$ then $X \geq_a Z$ (transitivity of $\geq_a$ for sets is easy to check), we obtain that $Min_a([w]_{R^=} \cap [w]_a) \geq_a Min_a([u]_{R^=} \cap [u]_a)$ and therefore $(w,u) \in \geq_a^R$. *Minimal elements*: Consider a $\sim_a$-equivalence class $W'' \subseteq W$, and let $W' \subseteq W''$ be a non-empty subset. Suppose W' does not have $\geq_a^R$ minimal elements. Then for all $w' \in W'$ there is a $w'' \in W'$ such that $w'' <_a^R w'$, i.e. $Min_a([w'']_{R^=} \cap [w'']_a) <_a Min_a([w']_{R^=} \cap [w']_a)$. As $w' \in [w']_{R^=} \cap [w']_a$, we get $\{w'\} \geq_a Min_a([w']_{R^=} \cap [w']_a)$ and then $Min_a([w'']_{R^=} \cap [w'']_a) <_a \{w'\}$. In other words, for all $w' \in W'$ there is a $u \in W$, namely any $u \in Min_a([w'']_{R^=} \cap [w'']_a)$, such that $u <_a w'$. This contradicts $\geq_a$ being a well-preorder on W''. We have now shown 1), 2) and 3). Finally we show 4): Assume $w \geq_a^R v$, that is, $Min_a([w]_{R^=} \cap [w]_a) \geq_a Min_a([v]_{R^=} \cap [v]_a)$. This implies the existence of an $x \in Min_a([w]_{R^=} \cap [w]_a)$ and a $y \in Min_a([w]_{R^=} \cap [v]_a)$ with $x \geq_a y$. By choice of x and y we have $x \sim_a w$ and $y \sim_a v$. From $x \geq_a y$ we get $x \sim_a y$. Hence we have $w \sim_a x \sim_a y \sim_a v$, as required. □

While showing that the union of autobisimulations is an autobisimulation is trivial for standard bisimulations, it is decidedly non-trivial in our setting (Proposition 4.9). For two autobisimulations, say R and S, we have to relate conditions on $w \geq_a^{R \cup S} v$ to conditions on $w \geq_a^R v$ and $w \geq_a^S v$. Because of chaining, the set $[w]_{(R \cup S)^=}$ can be much larger than either of the sets $[w]_{R^=}$, $[w]_{S^=}$, and even their union. The crucial lemma below proceeds by induction on the distance from a world to one of the minimal ones.

Definition 4.6. Given two autobisimulations R and S on a model $(W, \geq, V)$, define the depth of a world as a function $d : A \to W \to \mathbb{N}$ (writing d_a for $d(a)$), as follows: Let $a \in A$, and $u, v \in W$ with $u \sim_a v$, then

$$d_a(u) = \min\{n \mid \exists v \in Min_a([u]_{(R \cup S)^=} \cap [u]_a) : (u,v) \in (R^= \cup S^=)^n \cap [u]_a\}$$

If $d_a(u) = n$ there is an alternating $R^=/S^=$ chain of length n connecting u to a $\geq_a$-minimal member of class $[u]_a$. We cannot assume that this is a $\geq_a$-descending chain (or even, for example, that an $R^=$ class closer to that minimal member is more plausible than one further away from it). As $(R^= \cup S^=)^n$ is the ith power of $(R^= \cup S^=)$, $(R^= \cup S^=)^0$ is just the identity, so if $d(u) = 0$, u is already a $\geq_a$-minimal world.

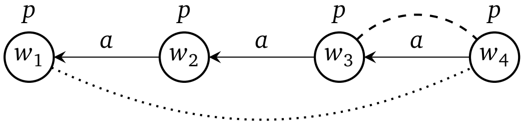

Figure 4.3: Two autobisimulations (modulo reflexivity) on the same model. From left to right, the depths are $0, 0, 2$, and 1 respectively.

Example 4.7. Figure 4.3 shows two autobisimulations on a plausibility model. In addition to the reflexive and symmetric pairs, R (the dashed edge) contains the pairs (w_3, w_4) while S (the dotted edge) contains (w_1, w_4). First, note that $([w_1]_{(R\cup S)^=} \cap [w_1]_a) = ([w_3]_{(R\cup S)^=} \cap [w_3]_a) = ([w_4]_{(R\cup S)^=} \cap [w_4]_a) = \{w_1, w_3, w_4\}$, and $Min_a(\{w_1, w_3, w_4\}) = \{w_1\}$. From $(w_1, w_1) \in (R^= \cup S^=)^0$, $(w_3, w_1) \in (R^= \cup S^=)^2$, and $(w_4, w_1) \in (R^= \cup S^=)^1$, we get $d(w_1) = 0$, $d(w_3) = 2$, and $d(w_4) = 1$. For w_2, we have $Min_a([w_2]_{(R\cup S)^=} \cap [w_2]_a) = Min_a\{w_2\} = \{w_2\}$, and therefore $d(w_2) = 0$.

This example illustrates why autobisimulations are tricky in our setting. Getting from w_3 via autobisimulations to a most plausible world in $[w_3]_{(R\cup S)^=} \cap [w_3]_a$ requires going via $(w_3, w_4) \in R^=$ to the less plausible world w_4, and then going to w_1 via $(w_4, w_1) \in S^=$. Also, taking the union of autobisimulations may change plausibilities in non-obvious ways. In this example we have, for instance, $w_3 >_a^R w_2$, $w_3 >_a^S w_2$, but $w_2 >_a^{R\cup S} w_3$. ■

To prove that $R \cup S$ is an autobisimulation, we first prove the following lemma (which takes us *almost* all the way). The proof proceeds by induction on the measure $\leq_3$ on triples of natural numbers defined as follows: Let $k, l, m, n, o, p \in \mathbb{N}$. Then $(k, l, m) \leq_3 (n, o, p)$ iff $[k \leq n, l \leq o$, and $m \leq p]$. Further, $(k, l, m) <_3 (n, o, p)$ iff $[(k, l, m) \leq_3 (n, o, p)$ and $(n, o, p) \not\leq_3 (k, l, m)]$.

Lemma 4.8. *For all $k, l, m \in \mathbb{N}$, for all $a \in A$, and for all $(w, w') \in R \cup S$ with $d_a(w) \leq k$ and $d_a(w') \leq l$:*

[atoms] *$V(w) = V(w')$;*

[forth$_\geq$] *If $v \in W$ and $w \geq_a^{R\cup S} v$, and $d_a(v) \leq m$, there is a $v' \in W$ such that $w' \geq_a^{R\cup S} v'$ and $(v, v') \in R \cup S$;*

[back$_\geq$] *If $v' \in W$ and $w' \geq_a^{R\cup S} v'$, and $d_a(v') \leq m$, there is a $v \in W$ such that $w \geq_a^{R\cup S} v$ and $(v, v') \in R \cup S$;*

[forth$_\leq$] *If $v \in W$ and $w \leq_a^{R\cup S} v$, and $d_a(v) \leq m$, there is a $v' \in W$ such that $w' \leq_a^{R\cup S} v'$ and $(v, v') \in R \cup S$;*

[back$_\leq$] *If $v' \in W$ and $w' \leq_a^{R\cup S} v'$, and $d_a(v') \leq m$,
there is a $v \in W$ such that $w \leq_a^{R\cup S} v$ and $(v,v') \in R \cup S$.*

Proof. The proof is by induction on k, l, and m (according to $<_3$ above), and by distinguishing various cases. (Although only one agent a plays a role in the proof below, the other agents are implicitly present, as the induction hypothesis assumes that R and S satisfy the bisimulation clauses for *all* agents.)

For the base case, let $k = l = m = 0$ and consider $a \in A$. Let $(w,w') \in R \cup S$ and assume without loss of generality that $(w,w') \in R$. Clause [atoms] is satisfied. We now prove [forth$_\geq$]. Let $v \in W$ and $w \geq_a^{R\cup S} v$, i.e. $Min_a([w]_{(R\cup S)^=} \cap [w]_a) \geq_a Min_a([v]_{(R\cup S)^=} \cap [v]_a)$. We want to find a $v' \in W$, s.t. $w' \geq_a^{R\cup S} v'$, i.e. $Min_a([w']_{(R\cup S)^=} \cap [w']_a) \geq_a Min_a([v']_{(R\cup S)^=} \cap [v']_a)$. From $d_a(w) = d_a(v) = 0$, we get $w \in Min_a([w]_{(R\cup S)^=} \cap [w]_a)$ and $v \in Min_a([v]_{(R\cup S)^=} \cap [v]_a)$, thereby allowing us to conclude $w \geq_a v$ from $w \geq_a^{R\cup S} v$. Further, as any element that is minimal in a set is also minimal in a subset containing that element, we have $w \in Min_a([w]_{R^=} \cap [w]_a)$ and $v \in Min_a([v]_{R^=} \cap [v]_a)$. As $w \geq_a v$, this gives $w \geq_a^R v$.

With $(w,w') \in R$, [forth$_\geq$] for R now gets us a $v' \in W$ such that $w' \geq_a^R v'$ and $(v,v') \in R$. As $d_a(w') = 0$, we have $w' \in Min_a([w']_{(R\cup S)^=} \cap [w']_a)$. This gives $Min_a([w']_{(R\cup S)^=} \cap [w']_a) \simeq_a Min_a([w']_{R^=} \cap [w']_a)$ and because $w' \geq_a^R v'$ that $Min_a([w']_{(R\cup S)^=} \cap [w']_a) \geq_a Min_a([v']_{R^=} \cap [v']_a)$. Finally, we then get $Min_a([w']_{(R\cup S)^=} \cap [w']_a) \geq_a Min_a([v']_{(R\cup S)^=} \cap [v']_a)$, i.e. $w' \geq_a^{R\cup S} v'$. From $(v,v') \in R$ we have that $(v,v') \in R \cup S$, thereby satisfying [forth$_\geq$]. The case when $(w,w') \in S$ is similar, as are the cases for [back$_\geq$], [forth$_\leq$], and [back$_\leq$].

We now consider the case where $k > 0$ and $l,m = 0$. As other cases where one or more of k,l,m are 0 can be treated similarly, we only do this one, before proceeding to the case where $k,l,m > 0$.

Let $k > 0$ and $l,m = 0$. As before, consider $a \in A$, let $(w,w') \in R \cup S$ and without loss of generality consider $(w,w') \in R$. Clause [atoms] is satisfied. We now prove [forth$_\geq$]. Let $v \in W$ and $w \geq_a^{R\cup S} v$ (with $d_a(v) = 0$).

As $d_a(w) = k > 0$, there is a world x in either $[w]_{R^=} \cap [w]_a$ or $[w]_{S^=} \cap [w]_a$ closer to $Min_a([w]_{(R\cup S)^=} \cap [w]_a)$ than w (with $d_a(x) = k - 1$). This world is the first step in the $(R \cup S)^=$ chain from w to $\geq_a$-minimal worlds in $[w]_{(R\cup S)^=} \cap [w]_a$, and it can be reached from w via either $R^=$ or $S^=$ (or both).

Consider the **first case (R)**. (See also Fig. 4.4, although this better illustrates the general case.) Let $x \in Min_a([w]_{R^=} \cap [w]_a)$. Then $x \geq_a^{R\cup S} v$ (as $Min_a([w]_{(R\cup S)^=} \cap [w]_a) = Min_a([x]_{(R\cup S)^=} \cap [x]_a)$). From $w \simeq_a^R x$ follows $w \geq_a^R x$. From $w \geq_a^R x$ and

$(w, w') \in R$ and the fact that R is a bisimulation follows from [forth$_\geq$] (for R) that there is a x' such that $w' \geq^R_a x'$ and $(x, x') \in R$. As $d_a(w') = l = 0$ we may take $x' = w'$. As $d_a(x) = k - 1$, then from $(k - 1, 0, 0) <_3 (k, 0, 0)$ and the inductive hypothesis it follows that there is (clause [forth$_\geq$] for $R \cup S$) a v' with $w' \geq^{R\cup S}_a v'$ and $(v, v') \in R \cup S$. This v' also satisfies the requirements of clause [forth$_\geq$] for $R \cup S$ for pair (w, w') (so not merely for (x, w')).

Now consider the **second case (S)**. First, consider $w \geq^R_a Min_a([w]_{R^=} \cap [w]_a)$ and again let $x \in Min_a([w]_{R^=} \cap [w]_a)$. As above, from $w \geq^R_a x$ and $(w, w') \in R$ follows from [forth$_\geq$] for R that there is a x' such that $w' \geq^R_a x'$ and $(x, x') \in R$, and similar to above we find a v' satisfying the requirements of clause [forth$_\geq$] for $R \cup S$ for pair (w, w'). Now, consider $w \leq^R_a Min_a([w]_{R^=} \cap [w]_a)$ and let $y \in Min_a([w]_{R^=} \cap [w]_a)$. From $w \leq^R_a y$ and $(w, w') \in R$ follows from [forth$_\leq$] for R (note that we now use a different clause, not [forth$_\geq$] but [forth$_\leq$]) that there is a y' such that $w' \leq^R_a y'$ and $(y, y') \in R$. As $d_a(y) = k - 1$, then from $(k - 1, 0, 0) <_3 (k, 0, 0)$ and the inductive hypothesis it follows that there is (clause [forth$_\geq$] for $R \cup S$) a v' with $y' \geq^{R\cup S}_a v'$ and $(v, v') \in R \cup S$. Now although $w' \leq^R_a y'$, we still have that $w' \geq^{R\cup S}_a y'$. This is because we have $w' \simeq^S_a y'$ (from $w \simeq^S_a y$) and from $w' \simeq^S_a y'$ and $y' \geq^{R\cup S}_a v'$ follows $w' \geq^{R\cup S}_a v'$. So this v' also satisfies the requirements of clause [forth$_\geq$] for $R \cup S$ for pair (w, w').

We now proceed with the general case where $k, l, m > 0$. Consider $a \in A$, let $(w, w') \in R \cup S$ and without loss of generality consider $(w, w') \in R$; clause [atoms] is satisfied.

[forth$_\geq$]: Let $v \in W$ and $w \geq^{R\cup S}_a v$, and $d_a(v) = m$. As $d_a(w) = k > 0$, there is (as before) an $x \in Min_a([w]_{R^=} \cap [w]_a)$ such that $w \simeq^R_a x$ or $[w \simeq^S_a x$ and $(w \geq^R_a x$ or $w \leq^R_a x)]$, and such that $x \geq^{R\cup S}_a v$ and $d_a(x) = k - 1$. In case $w \geq^R_a x$, from that and $(w, w') \in R$ and [forth$_\geq$] (for R) follows that there is a x' such that $w' \geq^R_a x'$ and $(x, x') \in R$. As $(x, x') \in R$, also $(x, x') \in R \cup S$. Then, from the inductive hypothesis applied to $(k - 1, l, m)$, $(x, x') \in R \cup S$ and $x \geq^{R\cup S}_a v$ it follows that there is a v' such that $(v, v') \in R \cup S$ and $x' \geq^{R\cup S}_a v'$. We now use that, as w' and x' are either equally $\simeq^R_a$ plausible or equally $\simeq^S_a$ plausible, therefore $Min([w']_{(R\cup S)^=} \cap [w']_a) = Min([x']_{(R\cup S)^=} \cap [x']_a)$. Therefore, we do not only have $x' \geq^{R\cup S}_a v'$ but also $w' \geq^{R\cup S}_a v'$, which closes this case of the proof. (Please consult Figure 4.4 that illustrates the construction in detail.)

[back$_\geq$]: This case proceeds similarly to [forth$_\geq$] only now we use the inductive hypothesis on $(k, l - 1, m)$.

[forth$_\leq$]: In this case we use the third leg, and reduce the m in the (k, l, m) triple.

Let $v \in W$ and $w \leq^{R\cup S}_a v$ (i.e., $v \geq^{R\cup S}_a w$), and $d_a(v) = m$. As $d_a(v) = m > 0$, there is an $x \in Min_a([w]_{R^=} \cap [w]_a)$ such that $v \simeq_a^R x$ or $v \simeq_a^S x$, and such that $x \geq_a^{R\cup S} w$ and $d_a(x) = m-1$. From $x \geq_a^{R\cup S} w$ and $(w, w') \in R$ and [forth$_\leq$] and induction (we are now in case $(k, l, m-1)$), it follows that there is an x' such that $x' \geq_a^{R\cup S} w'$ and $(x, x') \in R \cup S$. We now proceed by subcases again. First, assume $(x, x') \in R$. Then, in case $v \simeq_a^R x$, from $(x, x') \in R$, $v \simeq_a^R x$ and [forth$_\geq$] follows that there is a v' such that $v' \simeq_a^R x'$ and $(v, v') \in R$. Then, from $v' \simeq_a^R x'$ and $x' \geq_a^{R\cup S} w'$ follows $v' \geq_a^{R\cup S} w'$ (the $\geq_a^R$ minimum of x' must be at most as plausible as the $\geq_a^{R\cup S}$ minimum of x'), in the other direction, $w' \leq_a^{R\cup S} v'$. Also from $(v, v') \in R$ follows $(v, v') \in R \cup S$. This establishes that v' satisfies the requirements. In the case that $(x, x') \in R$ and $v \simeq_a^S x$, we use that either $v \geq_a^R x$ or $x \geq_a^R v$ (see case [forth$_\geq$]) and then proceed similarly, while finally concluding from $v' \simeq_a^S x'$ and $x' \geq_a^{R\cup S} w'$ that $v' \geq_a^{R\cup S} w'$; and again we establish that v' satisfies the requirements. In case $(x, x') \in S$, we can also proceed with subcases $v \simeq_a^R x$ and $v \simeq_a^S x$, respectively; and again we establish that v' satisfies the requirements. This concludes the proof.

[back$_\leq$]: This case proceeds similarly to [forth$_\leq$] only now we use the inductive hypothesis on $(k, l, m-1)$ for the stipulated v' (instead of for the stipulated v). □

Proposition 4.9. *The union of two autobisimulations is an autobisimulation.*

Proof. From Lemma 4.8, and the observation that for any $w \in W$, $d_a(w)$ is finite[2], now follows directly Proposition 4.9. (For example, take $(w, w') \in R \cup S$, take an agent $a \in A$, and further assume a v such that $v \geq_a^{R\cup S}$. Then apply Lemma 4.8 on numerical parameters $d_a(w)$, $d_a(w')$, $d_a(v)$. We then find some v'. This is the one we need. □

[2] Let $x \in Min_a([w]_{(R\cup S)^=} \cap [w]_a)$. We then have $(w, x) \in (R\cup S)^=$. Given that $(R \cup S)^= = (R^= \cup S^=)^*$, we get $(w, x) \in (R^= \cup S^=)^*$, so that there is an n such that $(w, x) \in (R^= \cup S^=)^n$: the alternation is finite!

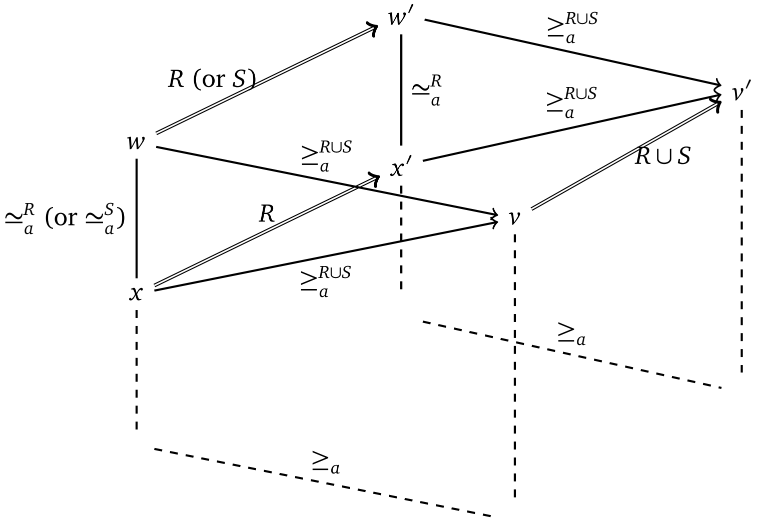

Figure 4.4: The union of two autobisimulations is an autobisimulation

Lemma 4.10. *The reflexive closure, transitive closure, and symmetric closure of an autobisimulation are autobisimulations.*

Proof. Let $M = (W, \geq, V)$ be a plausibility model and let R be an autobisimulation on M.

Reflexive Closure Let S be the reflexive closure of R. If $(w, w') \in S$, then $(w, w') \in R$ or $w = w'$. We now check the clauses for the autobisimulation. Clause [atoms] still applies. without loss of generality, let us now consider clause [forth$_\geq$] (the other three [forth] and [back] clauses can be shown similarly). If $(w, w') \in R$, then [forth$_\geq$] still holds because for all agents a, $\geq_a^R = \geq_a^S$, which is because $R^= = S^=$: the equivalence closure is also the equivalence closure of the reflexive closure. If $w = w'$, then [forth$_\geq$] obviously holds as well.

Transitive Closure Let S be the transitive closure of R. We show [forth$_\geq$]. Let $(w, w') \in S$ and consider $v \in W$ such that $w \geq_a^S v$. As $(w, w') \in S$ there is a finite sequence $w_0, \ldots, w_n$ for some natural number $n \geq 1$ such that $(w_i, w_{i+1}) \in R$ for all $i \leq n-1$, $w_0 = w$, and $w_n = w'$. For $n = 1$, [forth$_\geq$] holds, where we use that $\geq_a^R = \geq_a^S$ (the equivalence closure is also the equivalence closure of the transitive closure). Now assume we have proved [forth$_\geq$] for n, and consider a path of length $n+1$. We now have that $(w_0, w_n) \in S$ and $(w_n, w_{n+1}) \in R$. By induction, there is a v_n such that $w_n \geq_a^S v_n$ and $(v, v_n) \in S$.

From $w_n \geq_a^S v_n$ also follows $w_n \geq_a^R v_n$ (again, $\geq_a^R = \geq_a^S$). From $w_n \geq_a^R v_n$, $(w_n, w_{n+1}) \in R$ and [forth$_\geq$] for R it follows that there is a v_{n+1} such that $w_{n+1} \geq_a^R v_{n+1}$ and $(v_n, v_{n+1}) \in R$. From $(v, v_n) \in S$ and $(v_n, v_{n+1}) \in R$ follows $(v, v_{n+1}) \in S$ (S is the transitive closure of R). Also, from $w_{n+1} \geq_a^R v_{n+1}$ follows $w_{n+1} \geq_a^S v_{n+1}$ ($R^= = S^=$ so that $\geq_a^R = \geq_a^S$). Therefore, v_{n+1} is the required v'.

Symmetric Closure Let S be the symmetric closure of R. We show [forth$_\geq$]. We again use that $R^= = S^=$, now because the equivalence closure of a relation is the equivalence closure of its symmetric closure. Let $(w, w') \in S$ and consider $v \in W$ such that $w \geq_a^S v$. As S is symmetric, either $(w, w') \in R$ or $(w', w) \in R$. In the first case we apply [forth$_\geq$] for R and derive that there is a v' such that $w' \geq_a^R v'$ and $(v, v') \in R$, thus $w' \geq_a^S v'$ and $(v, v') \in S$. In the second case we apply [back$_\geq$] for R and derive that there is a v' such that $w' \geq_a^R v'$ and $(v', v) \in R$, thus also $w' \geq_a^S v'$ and $(v, v') \in S$.

□

Definition 4.11 (Maximal autobisimulation)**.** The *maximal autobisimulation* on a model is the union of all autobisimulations on that model.

Proposition 4.12. *The maximal autobisimulation is an autobisimulation, is maximal, and is an equivalence relation.*

Proof. Given $M = (W, \geq, V)$, let $\bigcup R$ denote the union of all autobisimulations on M.

- *The maximal autobisimulation is an autobisimulation.*
  From Proposition 4.9 that the union of two autobisimulations is an autobisimulation follows that the maximal autobisimulation is a bisimulation.[3]

- *The maximal autobisimulation is maximal.*
  Obvious.

- *The maximal autobisimulation is an equivalence relation.*

  - $\bigcup R$ *is reflexive*: for all w, $(w, w) \in \bigcup R$, because the identity relation R_{id} is a bisimulation (Lemma 4.10) and $R_{id} \subseteq \bigcup R$.

[3] An interesting direct proof along the same lines as Proposition 4.9 is also possible. We then observe that the d_a distance from a given w to a $\geq_a$ minimal element x of $[w]_{(\bigcup R)^=} \cap [w]_a)$ is, again finite, now employing that $(\bigcup R)^= = (\bigcup R^=)^*$ so that $(w, x) \in (\bigcup R^=)^n$ for some natural number n. Instead of an alternating S/R chain in the case of the proof of the union of two autobisimulations, we now have a chain *only containing a finite subset of autobisimulations from* $\bigcup R$; plus the initial bisimulation witnessing an assumed $(w, w') \in \bigcup R$, that makes for $n + 1$ autobisimulations. We now proceed by induction on triples as in the proof of Prop. 4.9.

- $\bigcup R$ *is transitive*: let $(w,w'),(w',w'') \in \bigcup R$. Then $(w,w') \in S$ and $(w',w'') \in T$ for some autobisimulations $S,T \subseteq \bigcup R$. From Proposition 4.9 follows that $S \cup T$ is also an autobisimulation, and from Lemma 4.10 follows that the transitive closure U of $S \cup T$ is also an autobisimulation. Given that $(w,w') \in S$ and $(w',w'') \in T$, we obviously have that $(w,w') \in S \cup T$ and $(w',w'') \in S \cup T$, and therefore also $(w,w') \in U$ and $(w',w'') \in U$. As U is transitive, $(w,w'') \in U$. From that and $U \subseteq \bigcup R$ follows $(w,w'') \in \bigcup R$.
- $\bigcup R$ *is symmetric*: let $(w,w') \in \bigcup R$. Then there is an $S \subseteq \bigcup R$ such that $(w,w') \in S$. From Lemma 4.10 follows that the symmetric closure T of S is also a bisimulation. As $(w,w') \in S$, it follows that $(w',w) \in T$, and from that and $T \subseteq \bigcup R$ follows $(w',w) \in \bigcup R$.

□

Our later results for the correspondence between bisimulation and modal equivalence assume that all relations $\geq_a$ are (pre)image-finite. This amounts to requiring that all equivalence classes of $\sim_a$ are finite, while still allowing infinite domains. We therefore also need that:

Lemma 4.13. *If a plausibility relation $\geq_a$ on a given model is (pre)image-finite, and R is a binary relation on that model, then $\geq_a^R$ is also (pre)image-finite.*

Proof. We recall that if $\geq_a$ is image-finite and preimage-finite, then $\geq_a$ and $\leq_a$ are image-finite, so $\geq_a \cup \leq_a := \sim_a$ as well; and all equivalence classes induced by $\sim_a$ are finite. Now observe that this also holds in the other direction. If a relation is image-finite, then so is any subset of the relation. Therefore, for any choice of R, as $\geq_a^R \subseteq \sim_a$ and $\leq_a^R \subseteq \sim_a$, $\geq_a^R$ and $\leq_a^R$ are image-finite, and therefore, $\geq_a^R$ is image-finite and preimage-finite. □

Definition 4.14 (Bisimulation). Let $M = (W, \geq, V)$ and $M' = (W', \geq', V')$ be plausibility models and let $M'' = M \sqcup M'$ be the disjoint union of the two. Given an autobisimulation R on M'', if $R' = R \cap (W \times W')$ is non-empty, then R' is called a *bisimulation* between M and M'. A bisimulation between (M,w) and (M',w') is a bisimulation between M and M' containing (w,w').

Example 4.15. Take another look at M_C and M_R of Figure 4.1. Let $M' = M_C \sqcup M_R$ and consider possible autobisimulations here. From Figure 4.6 we have the existence of a maximal autobisimulation on M_R. For M_C, the maximal autobisimulation is just the identity. Naming them R_R and R_C respectively, we (trivially) have that $R_R \cup R_C$ is an autobisimulation on M'. The question is whether we can extend $R_R \cup R_C$ to an autobisimulation on M' connecting the submodels M_R to M_C. We can. This new autobisimulation is $R = R' \cup R_R \cup R_C$, where $R'(u_1) = R'(u_3) = \{v_1\}$,

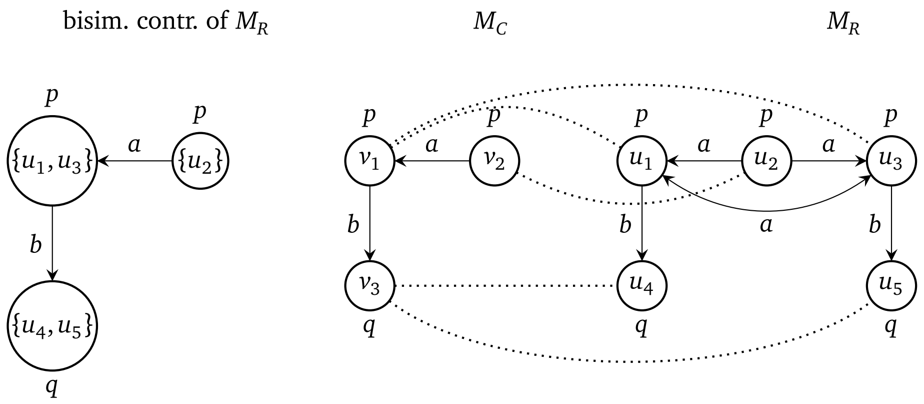

Figure 4.5: See Figure 4.1. The dotted edges show the maximal bisimulation between M_C and M_R. Model M_C is isomorphic to the bisimulation contraction of M_R (on the left) and to the bisimulation contraction of M_L (not depicted).

$R'(u_2) = \{v_2\}$ and $R'(u_4) = R'(u_5) = \{v_3\}$. Now we easily get a bisimulation between M_R and M_C as $R \cap (D(M_R) \times D(M_C)) = R'$. Figure 4.5 shows the bisimulation R'. ■

Definition 4.16 (Bisimulation contraction). Let $M = (W, \geq, V)$ be a plausibility model and let R be the maximal autobisimulation on M. The *bisimulation contraction* of M is the model $M' = (W', \geq', V')$ such that $W' = \{[w]_R \mid w \in W\}$, $V'([w]_R) = V(w)$, and for all agents a and worlds $w, v \in W$:

$$[w]_R \geq'_a [v]_R \quad \text{iff} \quad \text{for some } w' \in [w]_R \text{ and } v' \in [v]_R\text{: } w' \geq^R_a v'.$$

Example 4.17. We now compute the simulation contraction $M'_R = (W', \geq', V')$ of $M_R = (W, \geq, V)$. For $\geq'_a$ and $\geq'_b$ take the reflexive closures.

$$\begin{array}{lcl}
W' & = & \{\{u_1, u_3\}, \{u_2\}, \{u_4, u_5\}\} \\
\geq'_a & = & \{(\{u_2\}, \{u_1, u_3\})\} \\
\geq'_b & = & \{\{u_1, u_3\}, \{u_4, u_5\})\} \\
V'(\{u_1, u_3\}) & = & \{p\} \\
V'(\{u_2\}) & = & \{p\} \\
V'(\{u_4, u_5\}) & = & \{q\}
\end{array}$$

Model M_C is isomorphic to both the bisimulation contraction of M_L and the bisimulation contraction of M_R. ■

Proposition 4.18. *The bisimulation contraction of a plausibility model is a plausibility model and is bisimilar to that model.*

Proof. Let $M = (W, \geq, V)$ be a plausibility model, let R the maximal autobisimulation, and let $M = (W', \geq', V')$ be the bisimulation contraction.

The bisimulation contraction of a plausibility model is a plausibility model.
We have to show that $\geq'_a$ is a well-preorder on each $\sim'_a$ equivalence class.

- *Relation $\geq'_a$ is reflexive*:
  Suppose $[w]_R = [v]_R$. As $w \in [w]_R$ and $w \in [v]_R$, and $w \geq^R_a w$, by definition of $\geq'_a$ it follows that $[w]_R \geq'_a [v]_R$.
- *Relation $\geq'_a$ is transitive*:
  Assume that $[w]_R \geq'_a [v]_R$ and $[v]_R \geq'_a [u]_R$. Then there are $w' \in [w]_R$ and $v' \in [v]_R$ such that $w' \geq^R_a v'$, and there are $v'' \in [v]_R$ and $u'' \in [u]_R$ such that $v'' \geq^R_a u''$. We observe that for the maximal autobisimulation R, $R = R^=$, and that $[v]_a = [v']_a = [v'']_a$, so that we have that $Min_a([v]_R \cap [v]_a) = Min_a([v']_R \cap [v']_a) = Min_a([v'']_R \cap [v'']_a)$. From that, $w' \geq^R_a v'$, and $v'' \geq^R_a u''$ we conclude $Min_a([w']_R \cap [w']_a) \geq_a Min_a([v']_R \cap [v']_a) = Min_a([v'']_R \cap [v'']_a) \geq_a Min_a([u'']_R \cap [u'']_a)$. Therefore $Min_a([w']_R \cap [w']_a) \geq_a Min_a([u'']_R \cap [u'']_a)$, i.e., $w' \geq^R_a u''$, which by definition delivers the required $[w]_R \geq'_a [u]_R$.
- *Relation $\geq'_a$ satisfies that any non-empty subset has minimal elements*:
  Consider a $\sim'_a$ equivalence class $U' \subseteq W'$, and let $U \subseteq U'$ be a non-empty subset. Suppose the property does not hold. Then

  for all $u \in U$ there is a $u' \in U$ such that $u' <'_a u$,

  i.e.,
  $$\forall u \in U, \exists u' \in U, \forall v \in u, \forall v' \in u' : v' <^R_a v.$$
  Therefore,
  $$\forall u \in U, \exists u' \in U, \forall v \in u, \forall v' \in u' : Min_a([v']_R \cap [v']_a) <_a Min_a([v]_R \cap [v]_a).$$
  Let now $V = \bigcup U$, i.e., $V = \{v \in W \mid v \in u \text{ for some } u \in U\}$. As $u, u' \in U$ we have $u, u' \subseteq V$. We now have the following, where any $v' \in u'$ serves as a witness:
  $$\forall v \in V, \exists v' \in V : Min_a([v']_R \cap [v']_a) <_a Min_a([v]_R \cap [v]_a).$$
  By definition, this is
  $$\forall v \in V, \exists v' \in V : v' <^R_a v.$$
  Now $\geq_a$ is a well-preorder, and from Lemma 4.5 follows that $\geq^R_a$ is also a well-preorder. This contradicts the above.

The bisimulation contraction is bisimilar to the given model.
Consider the (functional) relation $S : W \to W'$ defined as $S = \{(w, [w]_R) \mid w \in W\}$. We prove that the relation S defines a autobisimulation. As it is a relation between two models, this therefore proves that it is a bisimulation. Consider $(w, [w]_R) \in S$.

- [atoms]: $V(w) = V'([w]_R)$ by the definition of bisimulation contraction.
- [forth$_\geq$]: Let $v \in W$ and $w \geq^S_a v$. We show that $[w]_R \geq'^S_a [v]_R$. As $(v, [v]_R) \in S$, this then demonstrates that $[v]_R \in W'$ satisfies the autobisimulation requirements for [forth$_\geq$]. All other clauses for [back] and [forth] are shown similarly.

  The relation S is rather particular. The proof uses the equivalence closure $S^=$ of S, and in particular, for a given w, the elements of $[w]_{S^=}$ in an equivalence class for a in W, and similarly the elements of $[[w]_R]_{S^=}$ in an equivalence class for a in W'. Now in the former case, as all $v \in [w]_R$ are mapped by S to the same $[w]_R \in W'$, that is the set $[w]_R$: $[w]_{S^=} = [w]_R$. This we use in step $(*)$ of the proof. In the latter case, as no member of $[w]_R \in W'$ is bisimilar to $[w]_R$, we have that $[[w]_R]_{S^=} \cap [[w]_R]_a$ is a singleton $\{[w]_R\}$. This is used in step $(**)$ of the proof. Step (***) is also important, as the class $[w]_a$ may contain objects from *different* $\sim_a$ equivalence classes in W.

$$
\begin{aligned}
&w \geq^S_a v \Leftrightarrow Min_a([w]_{S^=} \cap [w]_a) \geq_a Min_a([v]_{S^=} \cap [v]_a) \Leftrightarrow (*)\\
&Min_a([w]_R \cap [w]_a) \geq_a Min_a([v]_R \cap [v]_a) \Leftrightarrow w \geq^R_a v \Rightarrow (***)\\
&\exists w \in [w]_R, \exists v \in [v]_R : w \geq^R_a v \Leftrightarrow [w]_R \geq'_a [v]_R \Leftrightarrow\\
&Min_a\{[w]_R\} \geq'_a Min_a\{[v]_R\} \Leftrightarrow (**)\\
&Min_a([[w]_R]_{S^=} \cap [[w]_R]_a) \geq'_a Min_a([[v]_R]_{S^=} \cap [[v]_R]_a)\\
&\Leftrightarrow [w]_R \geq'^S_a [v]_R
\end{aligned}
$$

□

Definition 4.19 (Normal plausibility relation, normal model)**.** Let $M = (W, \geq, V)$ be a plausibility model and let R be the maximal autobisimulation on M. For all agents a, the relation $\geq^R_a$ is the *normal plausibility relation* for agent a in M, for which we may also write $\succeq_a$. The model is *normal* if for all a, $\geq_a = \geq^R_a$. Any model M can be *normalised* by replacing all $\geq_a$ by $\geq^R_a$.

Example 4.20. The models M_L and M_R of Figure 4.1 are reproduced in Figure 4.6, along with the normalisation of M_L, based on the previously seen maximal autobisimulation. The maximal autobisimulation shown for M_R is completely analogous to the one we've seen for M_L. ■

Proposition 4.21. *The bisimulation contraction of a plausibility model is normal.*

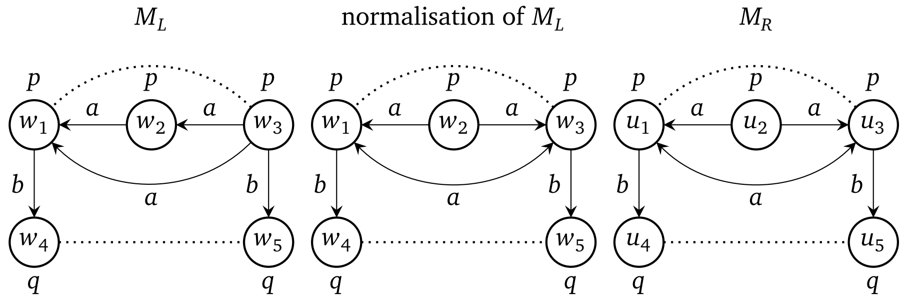

Figure 4.6: See Figure 4.1. The dotted lines show the maximal autobisimulations (modulo reflexivity).

Proof. Let M be a plausibility model, and let $M' = (W', \geq', V')$ be the bisimulation contraction of M. The maximal autobisimulation on M' is the identity relation R_{id}. For each agent a, we now have that $\geq'^{R_{id}}_a = \geq_a$. Therefore, M' is normal. □

4.3 Logical language and semantics

In this section we define the language and the semantics of our logics.

Definition 4.22 (Logical language)**.** For any countably infinite set of propositional symbols P and finite set of agents A we define language L^{CDS}_{PA} by:

$$\varphi ::= p \mid \neg\varphi \mid \varphi \wedge \varphi \mid K_a\varphi \mid B^{\varphi}_a\varphi \mid B^n_a\varphi \mid \Box_a\varphi$$

where $p \in P$, $a \in A$, and $n \in \mathbb{N}$.

The formula $K_a\varphi$ stands for 'agent a knows (formula) φ', $B^{\psi}_a\varphi$ stands for 'agent a believes φ on condition ψ', $B^n_a\varphi$ stands for 'agent a believes φ to degree n', and $\Box_a\varphi$ stands for 'agent a safely believes φ'. (The semantics of these constructs is defined below.) The duals of K_a, B^{φ}_a and $\Box_a$ are denoted $\widehat{K}_a$, $\widehat{B}^{\varphi}_a$ and $\Diamond_a$. We use the usual abbreviations for the boolean connectives as well as for $\top$ and $\bot$, and the abbreviation B_a for $B^{\top}_a$. In order to refer to the type of modalities in the text, we call K_a a *knowledge modality*, B^{ψ}_a a *conditional belief modality*, B^n_a a *degree of belief modality*, and $\Box_a$ a *safe belief modality*.

In L^{CDS}_{PA}, if A is clear from the context, we may omit that and write L^{CDS}_P, and if P is clear from the context, we may omit that as well, so that we get L^{CDS}. The letter C stands for 'conditional', D for 'degree', and S for 'safe'. Let X be any subsequence of

CDS, then L^X is the language with, in the inductive definition, only the modalities X (and with knowledge K_a) for all agents. In our work we focus on the *logic of conditional belief* with language L^C, the *logic of degrees of belief* with language L^D, and the *logic of safe belief* with language L^S.

Definition 4.23 (Satisfaction Relation). Let $M = (W, \geq, V)$ be a plausibility model for P and A, let $\succeq$ be the normal plausibility relation for M, and let $w \in W$, $p \in P$, $a \in A$, and $\varphi, \psi \in L^{CDS}$. Then:

$$\begin{array}{lll}
M, w \models p & \text{iff} & p \in V(w) \\
M, w \models \neg\varphi & \text{iff} & M, w \not\models \varphi \\
M, w \models \varphi \wedge \psi & \text{iff} & M, w \models \varphi \text{ and } M, w \models \psi \\
M, w \models K_a\varphi & \text{iff} & M, v \models \varphi \text{ for all } v \in [w]_a \\
M, w \models B_a^{\psi}\varphi & \text{iff} & M, v \models \varphi \text{ for all } v \in Min_a([\![\psi]\!]_M \cap [w]_a) \\
M, w \models B_a^{n}\varphi & \text{iff} & M, v \models \varphi \text{ for all } v \in Min_a^n[w]_a \\
M, w \models \Box_a\varphi & \text{iff} & M, v \models \varphi \text{ for all } v \text{ with } w \succeq_a v
\end{array}$$

where

$$\begin{array}{lll}
Min_a^0[w]_a & = & Min_{\succeq_a}[w]_a \\
Min_a^{n+1}[w]_a & = & \begin{cases} [w]_a & \text{if } Min_a^n[w]_a = [w]_a \\ Min_a^n[w]_a \cup Min_{\succeq_a}([w]_a \setminus Min_a^n[w]_a) & \text{otherwise} \end{cases}
\end{array}$$

and where $[\![\varphi]\!]_M = \{w \in W \mid M, w \models \varphi\}$.

We write $M \models \varphi$ (φ is valid on M) to mean that $M, w \models \varphi$ for all $w \in W$.

Definition 4.24 (Modal equivalence). Consider the language L_P^X, for X a subsequence of CDS. Given are models $M = (W, \geq, V)$ and $M' = (W', \geq', V')$, and $w \in W$ and $w' \in W'$. We say that (M, w) and (M', w') are *modally equivalent* in L_P^X, notation $(M, w) \equiv_P^X (M', w')$, if and only if for all $\varphi \in L^X$, $M, w \models \varphi$ if and only if $M', w' \models \varphi$. If P is obvious from the context we may write $(M, w) \equiv^X (M', w')$.

The logic of conditional belief

The logic L^C of conditional belief appears in [Stalnaker, 1996, Baltag and Smets, 2008a, van Benthem, 2007, Baltag and Smets, 2008b], where particularly the latter two are foundational for dynamic belief revision (older roots are Lewis' counterfactual conditionals [Lewis, 1973]). An axiomatisation is found in [Stalnaker, 1996]. In this logic, defeasible belief $B_a\varphi$ is definable as $B_a^{\top}\varphi$, while $K_a\varphi$ is definable as $B_a^{\neg\varphi}\bot$.

Example 4.25. Consider Figure 4.1. In the the plausibility model M_C we have, for instance: $M_C \models K_a p \rightarrow (B_a B_b q \wedge \neg K_a B_b q)$: If a knows p (true in v_1 and v_2), a believes, but does not know, that b believes q. Another example is $M_C \models B_a^{\neg B_b q} K_b \neg q$:

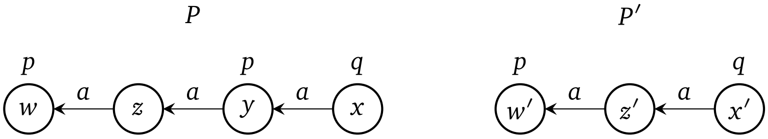

Figure 4.7: A plausibility model P and its bisimulation contraction P'.

Conditional on b not believing q, a believes that b knows $\neg q$. Only in v_2 does $\neg B_b q$ hold; there $K_b \neg q$ holds. A final example is $M_C \models K_a p \to B_a^{\widehat{K}_b q} B_b q$: From v_1 and v_2 (where $K_a p$ holds), formula $\widehat{K}_b q$ only holds in v_1, and conditional to that, the one and only most plausible world v_1 satisfies $B_b q$. We can repeat this exercise in M_L and M_R, as all three models are bisimilar and therefore, as will be proved in the next section, logically equivalent. ■

The logic of degrees of belief

The logic L^D of degrees of belief, also known as the logic of graded belief, goes back to [Grove, 1988, Spohn, 1988], although these could more properly be said to be semantic frameworks to model degrees of belief. Logics of degrees of belief have seen some popularity in artificial intelligence and AGM style belief revision, see e.g. [van der Hoek, 1992, van der Hoek, 1993, Laverny, 2006]. Belief revision based on degrees of belief have been proposed by [Aucher, 2005, van Ditmarsch, 2005]. The typical distinction between conviction (arbitrarily strong belief) and knowledge, as in [Lenzen, 1978, Lenzen, 2003], is absent in our logic L^D, wherein the strongest form of belief defines knowledge. Reasoning with degrees of belief is often called quantitative, where conditional belief can then be called qualitative. In other communities both are called qualitative, and quantitive epistemic reasoning approaches are in that case those that combine knowledge and probabilities [Halpern, 2003]. The zeroth degree of belief $B_a^0 \varphi$ defines defeasible belief $B_a \varphi$. How Spohn's work relates to dynamic belief revision as in [Baltag and Smets, 2008a] is discussed in detail in [van Ditmarsch, 2008]. There have also been various proposals combining knowledge and belief ($B_a^\top \varphi$ or $B_a^0 \varphi$) in a single framework, without considering either conditional or degrees of belief, where the dynamics are temporal modalities, see [Kraus et al., 1990, Kraus and Lehmann, 1988, Friedman and Halpern, 1994]. For purposes of further discussions and the proofs in Section 4.4.2 we define beliefs layers follows:

Definition 4.26 (Belief Layers). Let $M = (W, \geq, V)$. For $w \in W$, $a \in A$ and $n \in \mathbb{N}$, the nth (belief) layer of w for a is defined as $E_a^n[w]_a = Min_{\succeq_a}([w]_a \setminus Min_a^{n-1}[w]_a)$, where we use the special case $Min_a^{-1}[w]_a = \emptyset$.

This immediately gives the following lemma:

Lemma 4.27. *For $M = (W, \geq, V)$, $w \in W$, $a \in A$ and $n \in \mathbb{N}$, we have $Min^n_a[w]_a = E^n_a[w]_a \cup Min^{n-1}_a[w]_a$. For n such that $Min^n_a[w]_a = [w]_a$ we have $E^n_a[w]_a = \emptyset$. We name the smallest such n the* maximum degree *(for a at w). If n is the maximum degree for a at w, we have $M, w \models K_a\varphi \longleftrightarrow B^n_a\varphi$.*

In [Aucher, 2005, van Ditmarsch, 2005, Laverny, 2006] different layers can contain bisimilar worlds. In our approach they cannot, because we define belief degrees on the normal plausibility relation. Unlike [Spohn, 1988] our semantics does not allow empty layers in between non-empty layers. If $E^n_a[w]_a \neq \emptyset$ and $E^{n+2}_a[w]_a \neq \emptyset$, then $E^{n+1}_a[w]_a \neq \emptyset$. Layers above the maximum degree will be empty, i.e. if there is a maximum degree n for a at w, as there is in (pre)image-finite models, then for all degrees $k > n$, we have $E^k_a[w]_a = \emptyset$.

Example 4.28. In Figure 4.1, we have that $M_C \models B^0_aB^0_bq$ but not $M_C \models B^1_aB^0_bq$. The maximum degree of belief for a in M_C is at either v_1 and v_2, where it is 1, so $M_C \models K_a\varphi \longleftrightarrow B^1_a\varphi$. This is also true in the other two models. Consider now the models P and P' in Figure 4.7 and an alternative definition of B^n_a not using $\succeq_a$ but $\geq_a$ (as in [Aucher, 2005, van Ditmarsch, 2005, Laverny, 2006, Baltag and Smets, 2008b]). In the $\geq_a$-semantics we have $P \models B^2_a\neg q$, as q is false in $\{y, z, w\}$. Only when we reach the third degree of belief does q become uncertain: $P \not\models B^3_a\neg q$. With $\succeq_a$-semantics, 2 is maximum degree so $P \not\models B^2_a\neg q$. This can be seen in the bisimilar model P', where $P' \not\models B^2_a\neg q$. ■

The logic of safe belief

The logic L^S of safe belief goes back to Stalnaker [Stalnaker, 1996] and has been progressed by Baltag and Smets (for example, how it relates to conditional belief and knowledge) in [Baltag and Smets, 2008b], which also gives a detailed literature review involving the roots of conditional belief, degrees of belief, and safe belief. An agent has *safe belief* in a formula φ iff the agent will continue to believe φ no matter what *true* information conditions its belief, i.e. $M, w \models \Box_a\varphi$ iff $M, w \models B^\psi_a\varphi$ for all $\Box$-free ψ s.t. $M, w \models \psi$. In [Baltag and Smets, 2008b] safe belief is defined as $M, w \models \Box_a\varphi$ iff $M, v \models \varphi$ for all v s.t. $w \geq_a v$. For both [Stalnaker, 1996] and [Baltag and Smets, 2008b] true information are subsets of the domain containing the actual world. When this is what true information is, there is a correspondence between the two definitions (as indeed noted by Baltag and Smets). The complications of this choice are addressed in detail in [Demey, 2011]. For us, there is not a correspondence between the two definitions, because we can only condition on modally definable subsets. When we, as we do, define safe belief using $\succeq_a$, this correspondence is reestablished.

Example 4.29. Consider for a final time the models of Figure 4.1. We have $M_C, v_1 \models \Box_a \widehat{K}_b q$, whereas $M_C, v_2 \not\models \Box_a \widehat{K}_b q$. Now consider M_L and the $\geq_a$-version of safe belief for which we have $M_L, w_3 \not\models \Box_a \widehat{K}_b q$. For [Stalnaker, 1996, Baltag and Smets, 2008b] this is as it should be: For the subset $\{w_2, w_3\}$ (which includes the actual world w_3 as required) we have $Min_a(\{w_2, w_3\} \cap [w_3]_a) = \{w_2\}$ where $M_L, w_2 \not\models \widehat{K}_b q$. Using the $\succeq_a$-version of safe belief, we have $M_L, w_3 \models \Box_a \widehat{K}_b q$. For us, this is as it should be: As our conditional belief picks using $[\![\psi]\!] \cap [w]_a$, any set containing w_3 must include the modally equivalent world w_1. This corresponds to first normalising M_L to get M_R. In *that* model, u_2 is strictly less plausible than u_3. ■

The semantics we propose for degrees of belief and safe belief are non-standard. Still, as we show in the following, these non-standard semantics and the standard semantics for conditional belief are all bisimulation invariant. This makes the results in Section 4.5 showing a non-trivial expressivity hierarchy between these logics even more remarkable.

4.4 Bisimulation characterisation for L^C, L^D and L^S

4.4.1 Bisimulation correspondence for conditional belief

That the logic of conditional belief contains an infinity of modal operators (due to an infinity of conditional formulas), is why we have the extra step of first going via autobisimulations. There isn't a correspondence between the infinite number of conditional belief modalities and the 'non-normal' plausibility relations $\geq_a$, so we cannot define bisimulations with respect to them (if we want correspondence between bisimulation and modal equivalence). This is what we see in the above example: In M_L, for any ψ for which w_2 'sees' w_1 via B^ψ, w_2 also 'sees' w_3, even though w_2 is considered more plausible by $\geq_a$. Defining the normal plausibility relation reestablishes the correspondence between modality and relation, after which a definition of bisimulations fall into place.

In the following we prove that bisimilarity implies modal equivalence and vice versa. This shows that our notion of bisimulation is proper for the language and models at hand.

Theorem 4.30. *Bisimilarity implies modal equivalence for L^C.*

Proof. Assume $(M_1, w_1) \leftrightarrows (M_2, w_2)$. Then, by definition, there exists an autobisimulation R on the disjoint union of M_1 and M_2 with $(w_1, w_2) \in R$. Let $M = (W, \geq, V)$

denote the disjoint union of M_1 and M_2. We then need to prove that (M, w_1) and (M, w_2) are modally equivalent in L_C. We will show that for all φ in L_C, for all $(w, w') \in R$, if $M, w \models \varphi$ then $M, w' \models \varphi$. This implies the required (the other direction being symmetric). The proof is by induction on the syntactic complexity of φ. The propositional cases are easy, so we only consider the cases $\varphi = K_a\psi$ and $\varphi = B_a^\gamma\psi$. Consider first $\varphi = K_a\psi$. In this case we assume $M, w \models K_a\psi$, that is, $M, v \models \psi$ for all v with $w \sim_a v$. Let v' be chosen arbitrarily with $w' \sim_a v'$. We need to prove $M, v' \models \psi$. From Lemma 4.5 we have that $\geq_a^R$ is a well-preorder on each $\sim_a$-equivalence class. Since $w' \sim_a v'$ we hence get that either $w' \geq_a^R v'$ or $v' \geq_a^R w'$. We can assume $w' \geq_a^R v'$, the other case being symmetric. Then since $(w, w') \in R$ and $w' \geq_a^R v'$, [back$_\geq$] gives us a v s.t. $(v, v') \in R$ and $w \geq_a^R v$. Lemma 4.5 now implies $w \sim_a v$, and hence $M, v \models \psi$. Since $(v, v') \in R$, the induction hypothesis gives us $M, v' \models \psi$, and we are done.

Now consider the case $\varphi = B_a^\gamma\psi$. This case is more involved. Assume $M, w \models B_a^\gamma\psi$, that is, $M, v \models \psi$ for all $v \in Min_a(\llbracket\gamma\rrbracket_M \cap [w]_a)$. Letting $v' \in Min_a(\llbracket\gamma\rrbracket_M \cap [w']_a)$, we need to show $M, v' \models \psi$ (if $Min_a(\llbracket\gamma\rrbracket_M \cap [w']_a)$ is empty there is nothing to show). The standard approach is to use one of the back conditions to find a v with $(v, v') \in R$ and then show that $M, v \models \psi$. From this $M, v' \models \psi$ will follow, using the induction hypothesis. It is tempting to directly apply [back$_\geq$] to $w' \geq_a^R v'$ (or [back$_\leq$] to $w' \leq_a^R v'$) to produce such a v with $(v, v') \in R$. But unfortunately, we will not be able to conclude that such a v is in $Min_a(\llbracket\gamma\rrbracket_M \cap [w]_a)$, and hence not that $M, v \models \psi$. More work is needed. Instead we will first find a y in $Min_a(\llbracket\gamma\rrbracket_M \cap [w]_a)$, then find a y' with $(y, y') \in R$, and only then apply [back$_\geq$] to $y' \geq_a^R v'$ to produce the required v. The point is here that our initial choice of y in $Min_a(\llbracket\gamma\rrbracket_M \cap [w]_a)$ will ensure that v is in $Min_a(\llbracket\gamma\rrbracket_M \cap [w]_a)$.

As mentioned, we want to start out choosing a y in $Min_a(\llbracket\gamma\rrbracket_M \cap [w]_a)$, so we need to ensure that this set is non-empty. By choice of v' we have $v' \in \llbracket\gamma\rrbracket_M$ and $v' \sim_a w'$. From $v' \sim_a w'$ we get that $w' \geq_a^R v'$ or $w' \leq_a^R v'$, using Lemma 4.5. Since also $(w, w') \in R$, we can apply [back$_\geq$] or [back$_\leq$] to get a u such that $(u, v') \in R$ and either $w \geq_a^R u$ or $w \leq_a^R u$. From $(u, v') \in R$ and $v' \in \llbracket\gamma\rrbracket_M$, we get $u \in \llbracket\gamma\rrbracket_M$, using the induction hypothesis. From the fact that either $w \geq_a^R u$ or $w \leq_a^R u$ we get $w \sim_a u$, using Lemma 4.5. Hence we have $u \in \llbracket\gamma\rrbracket_M \cap [w]_a$. This shows the set $\llbracket\gamma\rrbracket_M \cap [w]_a$ to be non-empty. Hence also $Min_a(\llbracket\gamma\rrbracket_M \cap [w]_a)$ is non-empty, and we are free to choose a y in that set. Since $y \sim_a w$, Lemma 4.5 gives us that either $y \geq_a^R w$ or $w \geq_a^R y$, so we can apply [forth$_\leq$] or [forth$_\geq$] to find a y' with $(y, y') \in R$ and either $y' \geq_a^R w'$ or $w' \geq_a^R y'$.

Claim 1. $y' \geq_a^R v'$.

Proof of claim 1. We need to prove $Min_a([y']_{R^=} \cap [y']_a) \geq_a Min_a([v']_{R^=} \cap [v']_a)$. We first prove that $[y']_{R^=} \cap [y']_a \subseteq \llbracket\gamma\rrbracket_M \cap [w']_a$:

- $[y']_{R^=} \cap [y']_a \subseteq [\![\gamma]\!]_M$: Assume $y'' \in [y']_{R^=} \cap [y']_a$. Then $(y', y'') \in R^=$. Since we also have $(y, y') \in R$, we get $(y, y'') \in R^=$. From $(y, y'') \in R^=$ and $y \in [\![\gamma]\!]_M$ a finite sequence of applications of the induction hypothesis gives us $y'' \in [\![\gamma]\!]_M$.
- $[y']_{R^=} \cap [y']_a \subseteq [w']_a$: Assume $y'' \in [y']_{R^=} \cap [y']_a$. Then $y'' \sim_a y'$. Since we have either $y' \geq_a^R w'$ or $w' \geq_a^R y'$, we must also have $y' \sim_a w'$, by Lemma 4.5. Hence $y'' \sim_a y' \sim_a w'$ implying $y'' \in [w']_a$.

Since v' is chosen minimal in $[\![\gamma]\!]_M \cap [w']_a$ and $[y']_{R^=} \cap [y']_a \subseteq [\![\gamma]\!]_M \cap [w']_a$ we get $Min_a([y']_{R^=} \cap [y']_a) \geq_a \{v'\} \geq_a Min_a([v']_{R^=} \cap [v']_a)$, as required. This concludes the proof of the claim.

By choice of y' we have $(y, y') \in R$, and by Claim 1 we have $y' \geq_a^R v'$. We can now finally, as promised, apply [back$_\geq$] to these premises to get a v s.t. $(v, v') \in R$ and $y \geq_a^R v$.

Claim 2. $Min_a([v]_{R^=} \cap [v]_a) \subseteq Min_a([\![\gamma]\!]_M \cap [w]_a)$.

Proof of claim 2. Let $x \in Min_a([v]_{R^=} \cap [v]_a)$. We need to prove $x \in Min_a([\![\gamma]\!]_M \cap [w]_a)$. We do this by proving $x \in [\![\gamma]\!]_M$, $x \in [w]_a$ and $\{x\} \leq_a Min_a([\![\gamma]\!]_M \cap [w]_a)$:

- $x \in [\![\gamma]\!]_M$: By choice of x we have $(v, x) \in R^=$. From $(v, x) \in R^=$ and $(v, v') \in R$ we get $(v', x) \in R^=$. From $(v', x) \in R^=$ and $v' \in [\![\gamma]\!]_M$ a finite sequence of applications of the induction hypothesis gives us $x \in [\![\gamma]\!]_M$.
- $x \in [w]_M$: By choice of x we have $x \sim_a v$. Since $y \geq_a^R v$, Lemma 4.5 implies $v \sim_a y$. By choice of y we have $y \sim_a w$, so in total we get $x \sim_a v \sim_a y \sim_a w$, as required.
- $\{x\} \leq_a Min_a([\![\gamma]\!]_M \cap [w]_a)$:

$$\begin{array}{lll} \{x\} & \leq_a Min_a([v]_{R^=} \cap [v]_a) & \text{by choice of } x \\ & \leq_a Min_a([y]_{R^=} \cap [y]_a) & \text{since } y \geq_a^R v \\ & \leq_a \{y\} & \\ & \leq_a Min_a([\![\gamma]\!]_M \cap [w]_a) & \text{since } y \in Min_a([\![\gamma]\!]_M \cap [w]_a). \end{array}$$

This concludes the proof of the claim.

Now we are finally ready to prove $M, v' \models \psi$. Let $z \in Min_a([v]_{R^=} \cap [v]_a)$. Then $z \in Min_a([\![\gamma]\!]_M \cap [w]_a)$, by Claim 2. Hence $M, z \models \psi$, by assumption. Since

$(v,z) \in R^=$ and $(v,v') \in R$ we get $(z,v') \in R^=$, and hence a finite sequence of applications of the induction hypothesis gives us $M, v' \models \psi$.

□

We proceed now to show the converse, that modal equivalence with regard to L^C implies bisimulation. The proof has the same structure as the Hennessy-Millner approach, though appropriately modified for our purposes. Given a pair of image-finite models M and M', the standard approach is to construct a relation $R \subseteq \mathcal{D}(M) \times \mathcal{D}(M')$ s.t. $(w,w') \in R$ if $M,w \equiv^C M',w'$. Using $\Diamond$-formulas, it is then shown that R fulfils the requirements for being a bisimulation, as such formulas denote what is true at worlds accessible by whatever accessibility relation is used in the model. This means that modally equivalent worlds have modally equivalent successors, which is then used to show that R fulfils the required conditions. For our purposes this will not do, as we only have $\widehat{K}_a$-formulas (i.e. for $\sim_a$). Instead, our equivalent to $\Diamond$-formulas are of the form $\widehat{B}_a^{\psi}\varphi$, each such formula corresponding to accessibility to the most plausible ψ-worlds from all worlds in an equivalence class. What we want are formulas corresponding to specific links between worlds, so we first establish that such formulas exists. We thus have formulas with the same function as $\Diamond$-formulas serve in the standard approach.

Theorem 4.31. *On the class of (pre)image-finite models, modal equivalence with respect to L^C implies bisimilarity.*

Proof. Assume $(M_1, w) \equiv^C (M_2, w')$ and that both models are (pre)image-finite. We wish to show that $(M_1, w) \leftrightarroweq (M_2, w')$. Let $M = M_1 \sqcup M_2$ be the disjoint union of M_1 and M_2. We then need to show that $Q = \{(v,v') \in D(M) \times D(M) \mid M, v \equiv^C M, v'\}$ is an autobisimulation on M. Note that as $\equiv^C$ is an equivalence relation, so is Q. We first show that $\Diamond$-like formulas talking about the $\geq_a^Q$-relations between specific worlds in M exist.

Claim 1. Let w and w' be worlds of the (pre)image-finite model $M = (W, \geq, V)$ where $w \geq_a^Q w'$. Further let $\varphi \in L^C$ be any formula true in w'. There then exists a formula $\psi \in L^C$ such that $([w]_Q \cup [w']_Q) \cap [w]_a = [\![\psi]\!]_M \cap [w]_a$ and $M, w \models \widehat{B}_a^{\psi}\varphi$.

Proof of Claim 1. If two worlds s and s' are not modally equivalent, there exists some distinguishing formula $\Psi_{s,s'}$ with $M, s \models \Psi_{s,s'}$ and $M, s' \not\models \Psi_{s,s'}$. As $\sim_a$ is image-finite (since both $\geq_a$ and its converse are) the following formula is finite:

$$\Psi_t = \bigwedge \{\Psi_{t,t'} \mid t \sim_a t' \land (t,t') \notin Q\}$$

The formula Ψ_t distinguishes t from all the worlds in $[t]_a$ that it is not modally equivalent to. If there are no such worlds, Ψ_t is the empty conjunction equivalent to $\top$.

We now return to our two original worlds w and w'. With the assumption that $M,w' \models \varphi$, we show that $\psi = \Psi_w \vee \Psi_{w'}$ is a formula of the kind whose existence we claim. First note that $[\![\Psi_w]\!]_M \cap [w]_a$ contains only those worlds in $[w]_a$ that are modally equivalent to w, exactly as $[w]_Q \cap [w]_a$ does. As $[\![\Psi_w]\!]_M \cup [\![\Psi_{w'}]\!]_M = [\![\Psi_w \vee \Psi_{w'}]\!]_M$ we have $([w]_Q \cup [w']_Q) \cap [w]_a = [\![\Psi_w \vee \Psi_{w'}]\!]_M \cap [w]_a$. To get $M,w \models \widehat{B}_a^{\psi}\varphi$ we need to show that $\exists v \in Min_a([\![\Psi_w \vee \Psi_{w'}]\!]_M \cap [w]_{\sim_a})$ s.t. $M,v \models \varphi$. Pick an arbitrary $v \in Min_a([w']_Q \cap [w']_a)$. We will now show that this has the required properties.

Let $T = [\![\Psi_w \vee \Psi_{w'}]\!]_M \cap [w]_{\sim_a}$. Since $T = ([w]_Q \cup [w']_Q) \cap [w]_a$, Lemma 4.4 gives $u \simeq_a^Q w$ or $u \simeq_a^Q w'$ for all $u \in T$. Together with $w \geq_a^Q w'$, this gives $w' \in Min_{\geq_a^Q} T$. Choose $u \in T$ arbitrarily. We then have $u \geq_a^Q w'$ and, by definition, that $Min_a([u]_Q \cap [u]_a) \geq_a Min_a([w']_Q \cap [w']_a)$. By choice of v we can then conclude $\{v\} \leq_a Min_a([w']_Q \cap [w']_a) \leq_a Min_a([u]_Q \cap [u]_a) \leq_a \{u\}$. As u was chosen arbitrarily in T, this shows $v \in Min_a T$. As $v \in [w']_Q$ we have $M,v \equiv^C M,w'$ and by assumption of $M,w' \models \varphi$ that $M,v \models \varphi$. We now have $v \in Min_a([\![\Psi_w \vee \Psi_{w'}]\!]_M \cap [w]_{\sim_a})$ and $M,v \models \varphi$, completing the proof of the claim.

We now proceed to show that Q fulfils the conditions for being an autobisimulation on M (Definition 4.2). [atoms] is trivial. Next we show [forth$_\geq$]. Let $(w,w') \in Q$ (i.e. $(M,w) \equiv^C (M,w')$) and $w \geq_a^Q v$. We then have that [forth$_\geq$] is fulfilled if $\exists v' \in W$, s.t. $w' \geq_a^Q v'$ and $(v,v') \in Q$ (i.e. $(M,v) \equiv^C (M,v')$). To this end, we show that assuming for all $v' \in W$, $w' \geq_a^Q v'$ implies $(M,v) \not\equiv^C (M,v')$ leads to a contradiction. This is analogous to how Q is shown to be a bisimulation in standard Hennesey-Millner proofs.

From Lemma 4.13, we have that (pre)image-finiteness of the model gives that $\geq_a^Q$ is image-finite, so the set of $\geq_a^Q$-successors of w', $S = \{v' \mid w' \geq_a^Q v'\} = \{v_1', \dots, v_n'\}$ is also finite. Having assumed that v and none of the v_i's are modally equivalent, we have that there exists a number of distinguishing formulae $\varphi^{v_i'}$, one for each v_i', such that $M,v \models \varphi^{v_i'}$ and $M,v_i' \not\models \varphi^{v_i'}$. Therefore, $M,v \models \varphi^{v_1'} \wedge \cdots \wedge \varphi^{v_n'}$. For notational ease, let $\varphi = \varphi^{v_1'} \wedge \cdots \wedge \varphi^{v_n'}$.

With $M,v \models \varphi$, Claim 1 gives the existence of a formula ψ, such that $([w]_Q \cup [v]_Q) \cap [w]_a = [\![\psi]\!]_M \cap [w]_a$ and $M,w \models \widehat{B}_a^{\psi}\varphi$. Due to modal equivalence of w and w', we must have $M,w' \models \widehat{B}_a^{\psi}\varphi$. This we have iff $\exists u' \in Min_a([\![\psi]\!]_M \cap [w']_a)$, s.t. $M,u' \models \varphi$. By construction of φ, no world v_i' exists such that $w' \geq_a^Q v_i'$ and $M,v_i' \models \varphi$, so we must have $u' >_a^Q w'$. As $u' \in [w']_a$, the definition of $>_a^Q$ gives $Min_a([u']_Q \cap [w']_a) >_a Min_a([w']_Q \cap [w']_a)$, so we get $\exists w'' \in Min_a([w']_Q \cap [w']_a)$ s.t. $u' >_a w''$. As $u' \in Min_a([\![\psi]\!]_M \cap [w']_a)$, we must therefore have $w'' \notin [\![\psi]\!]_M$, and then also $w' \notin [\![\psi]\!]_M$. But as $M,w \models \psi$, we get the sought after contradiction (we initially assumed $(M,w) \equiv^C (M,w')$). We get [back$_\geq$] immediately from Q being an equivalence relation.

Now we get to [forth$_\leq$]. Let $(w,w') \in Q$ and $w \leq^Q_a v$. We have that [forth$_\leq$] is fulfilled if $\exists v' \in W$, s.t. $w' \leq^Q_a v'$ and $(v,v') \in Q$.

Claim 2. There exists a $v' \in [w']_a$ satisfying $(v,v') \in Q$.

Proof of Claim 2. Suppose not. Then v does not have a modally equivalent world in $[w']_a$. Thus there must be some formula φ holding in v that holds nowhere in $[w']_a$. Since $v \in [w]_a$ (using Lemma 4.5), this implies that $M,w \models \widehat{K}_a\varphi$ and $M,w' \not\models \widehat{K}_a\varphi$. However, this contradicts $(w,w') \in Q$, concluding the proof of the claim.

Let v' be chosen as guaranteed by Claim 2. It now suffices to show $w' \leq^Q_a v'$. From $(v,v') \in Q$ and $v \geq^Q_a w$, [forth$_\geq$] gives a w'' s.t. $v' \geq^Q_a w''$ and $(w,w'') \in Q$. From $v' \geq^Q_a w''$ we get $v' \sim_a w''$, using Lemma 4.5. Since $v' \in [w']_a$ we further get $w' \sim_a v' \sim_a w''$. Since $(w,w'') \in Q$ and $(w,w') \in Q$ we also get $(w',w'') \in Q$. From $w' \sim_a w''$ and $(w',w'') \in Q$ Lemma 4.4 gives us $w' \simeq^Q_a w''$. From this and $v' \geq^Q_a w''$ we get $v' \geq^Q_a w'$ and hence $w' \leq^Q_a v'$, as required. This concludes proof of [forth$_\leq$]. As for [back$_\geq$] getting to [back$_\leq$] is easy and left out.

□

Theorem 4.32 (Bisimulation characterization for L^C)**.** *Let $(M,w),(M',w')$ be epistemic plausibility models. On the class of image-finite models:*

$$(M,w) \leftrightarrows (M',w') \textit{ iff } (M,w) \equiv^C (M',w')$$

Proof. From Theorem 4.30 and Theorem 4.31. □

4.4.2 Bisimulation correspondence for degrees of belief

We now show bisimulation characterisation results for the logic of degrees of belief L^D. Let $M = (W, \geq, V)$. Recalling Definition 4.23, for some world $w \in W$, the set $Min^0_a[w]_a$ is the minimal worlds with respect to $\succeq_a$ in the $\sim_a$-equivalence class of w. For a given w and a, we refer to the generalised definition $Min^n_a[w]_a$ as (belief) sphere n of w for a. The distinction between Min^n_a and Min_a is important to keep straight! The former Min—used to give semantics of the B^n_a modality of L^D—is with respect to the relation $\succeq_a$. The latter Min is with respect to $\geq_a$, used to give the semantics of L^C. Dealing as we do in this section with L^D, we first state some necessary (and beautiful) observations about the properties of what we call beliefs spheres. When convenient we will simply say that v is in (belief) sphere n for a, understanding that this actually means $v \in Min_a^n[v]_a$.

It follows easily from the definitions, that for any world w, sphere n for a is wholly contained within sphere $n+1$ for a, i.e. $Min_a^n[w]_a \subseteq Min_a^{n+1}[w]_a$.

Lemma 4.33. *Let $M = (W, \geq, V)$ be a plausibility model and consider $w, v \in W$. If $w \sim_a v$ and $w \notin Min_a^n[w]_a$, we have the following two properties:*

(i) If $v \in Min_a^n[w]_a$, then $w \succ_a v$.

(ii) If $v \in Min_a^{n+1}[w]_a$ then $w \succeq_a v$.

Proof. The truth of (i) easily comes from the definition of Min_a^n. For (ii), we consider two exhaustive cases for v. Either $v \in Min_a^{n+1}[w]_a \setminus Min_a^n[w]_a$ in which case $w \succeq_a v$ follows from $\succeq_a$-minimality, since by assumption $w \in [w]_a \setminus Min_a^n[w]_a$. Otherwise $v \in Min_a^n[w]_a$, and so from $w \notin Min_a^n[w]_a$ and (i) it follows that $w \succ_a v$ and hence also $w \succeq_a v$. □

Now getting to the meat of this section, showing bisimulation correspondence for L^D, we first show that bisimilar worlds belong to spheres of all the same degrees.

Lemma 4.34. *If $(M_1, w_1) \leftrightarroweq (M_2, w_2)$ then for all $n \in \mathbb{N}$, $w_1 \in Min_a^n[w_1]_a$ iff $w_2 \in Min_a^n[w_2]_a$.*

Proof. Assume $(M_1, w_1) \leftrightarroweq (M_2, w_2)$. By definition there exists an autobisimulation R on the disjoint union of M_1 and M_2 with $(w_1, w_2) \in R$. Denote by $R_{\max}$ the extension of R into the maximal autobisimulation (so $\succeq_a = \geq_a^{R_{\max}}$). We are going to show by contradiction that for any $(w, w') \in R_{\max}$ (which includes (w_1, w_2)) and $n \in \mathbb{N}$, $w \in Min_a^n[w]_a$ iff $w' \in Min_a^n[w']_a$. Suppose that this does not hold. Then there must be some pair of worlds w and w' such that $(w, w') \in R_{\max}$ and either i) $w \in Min_a^n[w]_a$ and $w' \notin Min_a^n[w']_a$, or ii) $w \notin Min_a^n[w]_a$ and $w' \in Min_a^n[w']_a$ for some n. Let n be the smallest natural number for which we have either i) or ii). Because the cases are symmetrical, we deal only with i). Using the alternative definition $Min_a^n[w]_a = E_a^n[w]_a \cup Min_a^{n-1}[w]_a$ we can deal with both $n > 0$ and $n = 0$ simultaneously.

By assumption of the smallest n we have $w \notin Min_a^{n-1}[w]_a$, since $w' \notin Min_a^n[w']_a$ implies $w' \notin Min_a^k[w']_a$ for all $0 \leq k \leq n$ (so we could otherwise have chosen a smaller n). Therefore $w \in E_a^n[w]_a$ and $w' \notin E_a^n[w']_a$. Because $w' \in [w'] \setminus Min_a^n[w']_a$, we know that n is not the maximum degree, so there must be *some* world $v' \in E^n_a[w']_a$ which by definition means that $v' \notin Min^{n-1}_a[w']_a$. With $v' \in E^n_a[w']_a \subseteq Min^n_a[w']_a$ and and $w' \notin Min^n_a[w']_a$, Lemma 4.33 gives $w' \succ_a v'$, i.e. $w' \succeq_a v'$ and $v' \not\succeq_a w'$. By [back$_\geq$] there is a v s.t. $w \succeq_a v$ and $(v, v') \in R_{\max}$. Because $v' \notin Min^{n-1}_a[w']_a$ we cannot have $v \in Min^{n-1}_a[w]_a$, as we could then again have chosen a smaller n making either i) or ii) true. Thus $v \in [w]_a \setminus Min^{n-1}_a[w]_a$. As $w \in Min^n_a[w]_a$, Lemma 4.33 gives $v \succeq_a w$, so by [forth$_\geq$] there is a u' s.t. $v' \succeq_a u'$ and $(w, u') \in R_{\max}$.

With $(w, w') \in R_{\max}$ and $(w, u') \in R_{\max}$, we have $(w', u') \in R_{\max}$. As $w' \sim_a u'$ (we have $w' \succeq_a v'$ and $v' \succeq_a u'$), Lemma 4.4 gives $w' \simeq^{R_{\max}}_a u'$, i.e. $w' \succeq_a u'$ and $w' \preceq_a u'$. As $w' \notin Min^n_a[w']_a$, we then have $u' \notin Min^n_a[w']_a$. As $u' \notin Min^n_a[w']_a$ while $v' \in E^n_a[w']_a \subseteq Min^n_a[w']_a$, Lemma 4.33 then gives $u' \succ_a v'$. But this contradicts $v' \succeq_a u'$, concluding the proof.

□

Now we can prove that bisimilarity implies modal equivalence for L^D.

Theorem 4.35. *Bisimilarity implies modal equivalence for L^D.*

Proof. Assume $(M_1, w_1) \underline{\leftrightarrow} (M_2, w_2)$. Then there exists an autobisimulation R on the disjoint union of M_1 and M_2 with $(w_1, w_2) \in R$. Let $R_{\max}$ be the extension of R into the maximal autobisimulation and let $M = (W, \geq, V)$ denote the disjoint union of M_1 and M_2. We need to prove that $(M, w_1) \equiv^D (M, w_2)$.

We will show that for all $(w, w') \in R_{\max}$, for all $\varphi \in L^D$, $M, w \models \varphi$ iff $M, w' \models \varphi$ (which then also means that it holds for all $(w, w') \in R$). We proceed by induction on the syntactic complexity of φ. The propositional and knowledge cases are already covered by Theorem 4.30, so we only go for $\varphi = B^n_a \psi$.

Assume $M, w \models B^n_a \psi$. We need to prove that $M, w' \models B^n_a \psi$, that is $M, v' \models \psi$ for all $v' \in Min^n_a[w']_a$. Picking an arbitrary $v' \in Min^n_a[w']_a$, we have $[w']_a = [v']_a$ from Lemma 4.5, and either $w' \succeq_a v'$ or $w' \preceq_a v'$ (so we also have $v' \in Min^n_a[v']_a$). Using [back$_\geq$] or [back$_\leq$] as appropriate, we get that there is a v such that $w \succeq_a v$ or $w \preceq_a v$, and $(v, v') \in R_{\max}$. From this, $v' \in Min^n_a[v']_a$, and Lemma 4.34 we get $v \in Min^n_a[v]_a$, allowing us to conclude $v \in Min^n_a[w]_a$ from $[w]_a = [v]_a$. With the original assumption of $M, w \models B^n_a \psi$ we get $M, v \models \psi$. As $(v, v') \in R_{\max}$, the induction hypothesis gives $M, v' \models \psi$. As v' was chosen arbitrarily in $Min^n_a[w']_a$ this gives $M, w' \models B^n_a \psi$. Showing that $M, w' \models B^n_a \psi$ implies $M, w \models B^n_a \psi$ is completely symmetrical and therefore left out. □

We now get to showing that modal equivalence for the language of degrees of belief implies bisimilarity on the class of (pre)image-finite models. Trouble is, that the B_a^n modality uses the maximal autobisimulation for deriving the relation $\succeq_a$. This makes it difficult to go the Hennesey-Millner way of showing by contradiction that the modal equivalence relation Q is an autobisimulation.

Instead, we establish that modal equivalence for L^D implies modal equivalence for L^C. We go about this by way of a model and world dependent translation of L^C formulas into L^D formulas (Definition 4.37). This translation has two properties. First, the translated formula is true at M,w iff the untranslated formula is (Lemma 4.38)—a quite uncontroversial property. More precisely, letting $M = (W, \geq, R)$ be a plausibility model, then for any $w \in W$, $\gamma \in L^C$ where $\sigma_{M,w}(\gamma)$ is the translation at M,w: $M,w \models \gamma \Leftrightarrow M,w \models \sigma_{M,w}(\gamma)$. Assume further that we have some M',w' such that $(M,w) \equiv^D (M',w')$. As $\sigma_{M,w}(\gamma)$ is a formula of L^D we can conclude $M',w' \models \sigma_{M,w}(\gamma)$. So in all we get that

$$M,w \models \gamma \Leftrightarrow M,w \models \sigma_{M,w}(\gamma) \Leftrightarrow M',w' \models \sigma_{M,w}(\gamma) \qquad (*)$$

The second property is that the translation of γ is the same for worlds modally equivalent for L^D (Lemma 4.39): If $(M,w) \equiv^D (M',w')$ then $\sigma_{M,w}(\gamma) = \sigma_{M',w'}(\gamma)$. This then gives

$$M',w' \models \sigma_{M,w}(\gamma) \Leftrightarrow M',w' \models \sigma_{M',w'}(\gamma) \Leftrightarrow M',w' \models \gamma \qquad (**)$$

Combining (*) and (**) gives that if $(M,w) \equiv^D (M',w')$ then $M,w \models \gamma$ iff $M',w' \models \gamma$ for any $\gamma \in L^C$, i.e. that $(M,w) \equiv^C (M',w')$. As shown in the previous section, modal equivalence for L^C implies bisimilarity for (pre)image-finite models (Theorem 4.31), and we can therefore finally conclude that modal equivalence for L^D implies bisimilarity for (pre)image-finite models (Theorem 4.40).

Lemma 4.36. *For a (pre)image-finite model M, a world $w \in D(M)$, agent $a \in A$ and a formula ψ of L^C, if $[\![\psi]\!]_M \cap [w]_a \neq \emptyset$, there is a unique natural number k for which $Min_a([\![\psi]\!]_M \cap [w]_a) \subseteq E_a^k[w]_a$ $(= Min_{\succeq_a}([w]_a \setminus Min_a^{k-1}[w]_a))$.*

Proof. Let $S = [\![\psi]\!]_M \cap [w]_a$. We first show that all worlds in $Min_a S$ are equiplausible with respect to $\succeq_a$.

Take any two worlds $v_1, v_2 \in Min_a S$. We wish to show $v_1 \simeq_a^R v_2$, i.e. $Min_a([v_1]_R \cap [v_1]_a) \simeq_a Min_a([v_2]_R \cap [v_2]_a)$, where R is the maximal autobisimulation on M. With Theorem 4.30 (bisimilarity implies modal equivalence for L^C) and for $i = 1,2$, we have that $[v_i]_R \subseteq [\![\psi]\!]_M$. Hence $[v_i]_R \cap [v_i]_a = [v_i]_R \cap [w]_a \subseteq [\![\psi]\!]_M \cap [w]_a = S$. With $v_i \in Min_a S$ and $v_i \in [v_i]_R \cap [v_i]_a \subseteq S$, we have $v_i \in Min_a([v_i]_R \cap [v_i]_a)$ (if an element of a set A is minimal in a set $B \supseteq A$, then it is also minimal in A). From this we can conclude that $Min_a([v_i]_R \cap [v_i]_a) \simeq_a \{v_i\}$. Since $v_1 \simeq_a v_2$ we get $Min_a([v_1]_R \cap [v_1]_a) \simeq_a \{v_1\} \simeq_a \{v_2\} \simeq_a Min_a([v_2]_R \cap [v_2]_a)$, concluding the proof of $v_1 \simeq_a^R v_2$.

Due to (pre)image-finiteness of M, $[w]_a$ is finite. This means that for any $v \in [w]_a$ there is a unique natural number k for which $v \in E_a^k[w]_a$. As all worlds in $Min_a S$ are $\succeq_a$-equiplausible, we have that $Min_a S \subseteq E_a^k[w]_a$ for some unique k. □

Having established that if $[\![\psi]\!]_M \cap [w]_a \neq \emptyset$ then there does indeed exist a unique k st. $Min_a([\![\psi]\!]_M \cap [w]_a) \subseteq E_a^k[w]_a$, we have that the following translation is well-defined.

Definition 4.37 (Translation $\sigma_{M,w}$)**.** Let $M = (W, \geq, V)$ be a pre(image)-finite plausibility model and $\gamma \in L^C$ be given. We write $\sigma_{M,w}(\gamma)$ for the *translation* of γ at M, w into a formula of L^D defined as follows:

$$\sigma_{M,w}(p) = p$$

$$\sigma_{M,w}(\neg\varphi) = \neg\sigma_{M,w}(\varphi)$$

$$\sigma_{M,w}(\varphi_1 \wedge \varphi_2) = \sigma_{M,w}(\varphi_1) \wedge \sigma_{M,w}(\varphi_2)$$

$$\sigma_{M,w}(B_a^\psi \varphi) = \begin{cases} B_a^k \bigvee\{\sigma_{M,v}(\psi \to \varphi) \mid v \in [w]_a\} \wedge B_a^k \bigvee\{\sigma_{M,v}(\psi) \mid v \in [w]_a\} & \text{if } [\![\psi]\!]_M \cap [w]_a \neq \emptyset \\ K_a \bigvee\{\sigma_{M,v}(\neg\psi) \mid v \in [w]_a\} & \text{if } [\![\psi]\!]_M \cap [w]_a = \emptyset \end{cases}$$

where k is the natural number such that $Min_a([\![\psi]\!]_M \cap [w]_a) \subseteq E_a^k[w]_a$. As $K_a\varphi$ is definable in L^C as $B_a^{\neg\varphi}\bot$, we need no $K_a\varphi$-case in the translation.

We need (pre)image-finiteness M because the translation of $\sigma_{M,w}(B_a^\psi \varphi)$ is based on either $[w]_a$ or $Min_a([\![\psi]\!]_M \cap [w]_a)$. For $\sigma_{M,w}(B_a^\psi \varphi)$ to be finite, we need finiteness of $[w]_a$.

We now get to showing the first of the promised properties of the translation, namely that the translated formula is true at M, w iff the untranslated formula is.

Lemma 4.38. *Given a (pre)image-finite plausibility model $M = (W, \geq, V)$ and $\gamma \in L^C$ we have $M, w \models \gamma$ iff $M, w \models \sigma_{M,w}(\gamma)$ for all $w \in W$.*

Proof. We show both directions by induction on the modal depth of γ. For the base case of a modal depth of 0, we have $\sigma_{M,w}(\gamma) = \gamma$ easily, giving $M, w \models \gamma$ iff $M, w \models \sigma_{M,w}(\gamma)$. The p-, $\neg$-, $\wedge$-cases being quite easy, we deal only with $\gamma = B_a^\psi \varphi$ in the induction step. For that case there are to subcases; whether $\sigma_{M,w}(\gamma)$ is a K_a-formula or not.

(⇒) : $M, w \models \gamma$ implies $M, w \models \sigma_{M,w}(\gamma)$.

Take first the case $[\![\psi]\!]_M \cap [w]_a = \emptyset$ where $\sigma_{M,w}(B_a^\psi \varphi) = K_a \bigvee\{\sigma_{M,v}(\neg\psi) \mid v \in [w]_a\}$. If $[\![\psi]\!]_M \cap [w]_a = \emptyset$, then $M, v \models \neg\psi$ for all $v \in [w]_a$. Applying the induction hypothesis gives $M, v \models \sigma_{M,v}(\neg\psi)$ for all $v \in [w]_a$. Then we also have $M,u \models \bigvee\{\sigma_{M,v}(\neg\psi) \mid v \in [w]_a\}$ for all $u \in [w]_a$ and finally $M,w \models K_a \bigvee\{\sigma_{M,v}(\neg\psi) \mid v \in [w]_a\}$.

Now take the case $[\![\psi]\!]_M \cap [w]_a \neq \emptyset$. Letting $S = Min_a([\![\psi]\!]_M \cap [w]_a)$ and k be chosen as in the translation, i.e. such that $S \subseteq E^k_a[w]_a$, we wish to prove that $M,w \models B^\psi_a \varphi$ implies $M,w \models B^k_a \bigvee\{\sigma_{M,v}(\psi \rightarrow \varphi) \mid v \in [w]_a\} \wedge \widehat{B}^k_a \bigvee\{\sigma_{M,v}(\psi) \mid v \in [w]_a\}$. We first show $M,w \models \widehat{B}^k_a \bigvee\{\sigma_{M,v}(\psi) \mid v \in [w]_a\}$. Because $M,v \models \psi$ for all $v \in S$, the induction hypothesis gives $M,v \models \sigma_{M,v}(\psi)$ for all $v \in S$. From this we can conclude $M,u \models \bigvee\{\sigma_{M,v}(\psi) \mid v \in S\}$ for all $u \in S$, and thus also $M,u \models \bigvee\{\sigma_{M,v}(\psi) \mid v \in [w]_a\}$ for all $u \in S$. From Lemma 4.36 we have $S \subseteq Min^k_a[w]_a$, so $M,u \models \bigvee\{\sigma_{M,v}(\psi) \mid v \in [w]_a\}$ for some $u \in Min^k_a[w]_a$. This gives $M,w \models \widehat{B}^k_a \bigvee\{\sigma_{M,v}(\psi) \mid v \in [w]_a\}$. Next is $M,w \models B^k_a \bigvee\{\sigma_{M,v}(\psi \rightarrow \varphi) \mid v \in [w]_a\}$.

Claim. If $M,w \models B^\psi_a \varphi$, then for all $v \in E^k_a[w]_a \cap [\![\psi]\!]_M$, $M,v \models \varphi$.

Proof of claim. We show the claim by contradiction, assuming that at least one world in $E^k_a[w]_a \cap [\![\psi]\!]_M$ is a $\neg\varphi$-world. Let v be this $\psi \wedge \neg\varphi$-world. As $v \in E^k_a[w]_a$, we have $\{v\} \simeq^{R_{\max}}_a E^k_a[w]_a \simeq^{R_{\max}}_a S$, and specifically that $\forall s \in S : v \simeq^{R_{\max}}_a s$. This means $\forall s \in S : Min([v]_{R_{\max}} \cap [v]_a) \simeq_a Min_a([s]_{R_{\max}} \cap [s]_a)$. Because $\forall s \in S : Min_a([s]_{R_{\max}} \cap [s]_a) \simeq_a S$, we have $Min([v]_{R_{\max}} \cap [v]_a) \simeq_a S$ and thus some $v' \in Min([v]_{R_{\max}} \cap [v]_a)$ such that $\{v'\} \simeq_a S$. Combining $v' \in [v]_{R_{\max}}$ with Theorem 4.32 gives $M,v \equiv^C M,v'$ and thus that $M,v' \models \psi \wedge \neg\varphi$. Putting $v' \in [\![\psi]\!]_M$ together with $\{v'\} \simeq_a S$, means that $v' \in S$. As $M,v' \models \neg\varphi$, we have a contradiction of $M,w \models B^\psi_a \varphi$, concluding the proof of the claim.

With $M,w \models B^\psi_a \varphi$, we now have $M,v \models \varphi$ for all $v \in E^k_a[w]_a \cap [\![\psi]\!]_M$, and thus $M,v \models \psi \rightarrow \varphi$ for all $v \in E^k_a[w]_a$. Lemma 4.36 gives $S \subseteq E^k_a[w]_a$, and by definition we have $E^k_a[w]_a \cap Min^{k-1}_a[w]_a = \emptyset$, that is, there are no ψ-worlds below layer k, so $M,v \models \psi \rightarrow \varphi$ for all $v \in Min^k_a[w]_a$. Using the induction hypothesis gives $M,v \models \sigma_{M,v}(\psi \rightarrow \varphi)$ for all $v \in Min^k_a[w]_a$ and therefore $M,w \models B^k_a \bigvee\{\sigma_{M,v}(\psi \rightarrow \varphi) \mid v \in [w]_a\}$, finalising left-to-right direction of the proof.

($\Leftarrow$) : $M,w \models \sigma_{M,w}(\gamma)$ implies $M,w \models \gamma$.

We show the stronger claim that $M,w \models \sigma_{M,w'}(\gamma)$ for some $w' \in D(M)$ implies $M,w \models \gamma$. Let $\gamma = B^\psi_a \varphi$ and suppose that $M,w \models \sigma_{M,w'}(\gamma)$ for some $w' \in D(M)$. We then need to show $M,w \models B^\psi_a \varphi$. First take the case where $[\![\psi]\!]_M \cap [w']_a = \emptyset$. Then $\sigma_{M,w'}(B^\psi_a \varphi) = K_a \bigvee\{\sigma_{M,v'}(\neg\psi) \mid v' \in [w']_a\}$, i.e. $M,w \models K_a \bigvee\{\sigma_{M,v'}(\neg\psi) \mid v' \in [w']_a\}$. This means that $M,v \models \bigvee\{\sigma_{M,v'}(\neg\psi) \mid v' \in [w']_a\}$ for all $v \in [w]_a$, i.e. for any $v \in [w]_a$ there is a $v' \in [w']_a$ such that $M,v \models \sigma_{M,v'}(\neg\psi)$. Applying the induction hypothesis, we get $M, v \models \neg\psi$ for all $v \in [w]_a$. Thus $[\![\psi]\!]_M \cap [w]_a = \emptyset$ and we trivially have $M, w \models B_a^\psi \varphi$.

Now take the case $[\![\psi]\!]_M \cap [w']_a \neq \emptyset$. Letting $S' = Min_a([\![\psi]\!]_M \cap [w']_a)$ and k' be s.t. $S' \subseteq E_a^{k'}[w']_a$, we have $M, w \models B_a^{k'} \bigvee\{(\sigma_{M,v'}(\psi \to \varphi) \mid v' \in [w']_a\} \land \widehat{B}_a^{k'} \bigvee\{\sigma_{M,v'}(\psi) \mid v' \in [w']_a\}$. From $M, w \models B_a^{k'} \bigvee\{\sigma_{M,v'}(\psi \to \varphi) \mid v' \in [w']_a\}$ we have $M, v \models \bigvee\{\sigma_{M,v'}(\psi \to \varphi) \mid v' \in [w']_a\}$ for all $v \in Min_a^{k'}[w]_a$, i.e. for any $v \in [w]_a$ there is a $v' \in [w']_a$ such that $M, v \models \sigma_{M,v'}(\psi \to \varphi)$. Applying the induction hypothesis, we get $M, v \models \psi \to \varphi$ for all $v \in Min_a^{k'}[w]_a$. From $M, w \models \widehat{B}_a^{k'} \bigvee\{\sigma_{M,v'}(\psi) \mid v' \in [w']_a\}$ we have $M, v \models \bigvee\{\sigma_{M,v'}(\psi) \mid v' \in [w']_a\}$ for some $v \in Min_a^{k'}[w]_a$, i.e. there is a $v \in [w]_a$ and a $v' \in [w']_a$ such that $M, v \models \sigma_{M,v'}(\psi)$. Applying the induction hypothesis gets us $M, v \models \psi$. Thus we have $M, w \models B_a^{k'}(\psi \to \varphi) \land \widehat{B}_a^{k'} \psi$ (where $\psi, \varphi \in L^C$).

From $M, w \models \widehat{B}_a^{k'} \psi$ we have that $[\![\psi]\!]_M \cap [w]_a \neq \emptyset$, so Lemma 4.36 gives the existence of a k, s.t. $Min_a([\![\psi]\!]_M \cap [w]_a) \subseteq Min_a^k[w]_a$. We also have from $M, w \models \widehat{B}_a^{k'} \psi$ that $k \leq k'$, so $Min_a([\![\psi]\!]_M \cap [w]_a) \subseteq Min_a^{k'}[w]_a$. With $M, w \models B_a^{k'}(\psi \to \varphi)$ we get $M, v \models \psi \to \varphi$ for all $v \in Min_a([\![\psi]\!]_M \cap [w]_a)$, then $M, v \models \varphi$ for all $v \in Min_a([\![\psi]\!]_M \cap [w]_a)$, and finally $M, w \models B_a^\psi \varphi$. □

We have now gotten to the second of the two promised properties; that the translation is the same for worlds modally equivalent for L^D.

Lemma 4.39. *Given (pre)image-finite plausibility models M and M', for any $w \in D(M)$ and $w' \in D(M')$, if $(M, w) \equiv^D (M', w')$ then for any formula $\gamma \in L^C$, $\sigma_{M,w}(\gamma) = \sigma_{M',w'}(\gamma)$.*

Proof. We show this by another induction on the modal depth of γ. For the base case of modal depth 0 we trivially have $\sigma_{M,w}(\gamma) = \sigma_{M',w'}(\gamma)$.

For the induction step we, as before, only deal with $\gamma = B_a^\psi \varphi$. Note first that every world in $[w]_a$ is modally equivalent to at least one world in $[w']_a$. If that wasn't the case, there would be some formula L^D-formula φ true somewhere in $[w]_a$ and nowhere in $[w']_a$. Then $M, w \models \widehat{K}_a \varphi$ while $M', w' \not\models \widehat{K}_a \varphi$, contradicting $(M, w) \equiv^D (M', w')$. A completely analogous argument gives that every world in $[w']_a$ is modally equivalent to at least one world in $[w]_a$. Thus $[\![\psi]\!]_M \cap [w]_a = \emptyset$ iff $[\![\psi]\!]'_M \cap [w']_a = \emptyset$. We thus have two cases, either both $\sigma_{M,w}(B_a^\psi \varphi)$ and $\sigma_{M',w'}(B_a^\psi \varphi)$ are K_a-formulas, or both are B_a^k-formulas.

We deal first with the case where both translations are K_a-formulas. Here we have $\sigma_{M,w}(B_a^\psi \varphi) = K_a \bigvee\{\sigma_{M,v}(\neg\varphi) \mid v \in [w]_a\}$ and $\sigma_{M',w'}(B_a^\psi \varphi) = K_a \bigvee\{\sigma_{M',v'}(\neg\varphi) \mid v' \in [w']_a\}$. As already shown, for all $v \in [w]_a$ there is a $v' \in [w']_a$ such that

$(M,w) \equiv^D (M',v')$, and vice versa. The induction hypothesis gives $\sigma_{M,v}(\neg\varphi) = \sigma_{M',v'}(\neg\varphi)$ for all these vs and v's. Then $\bigvee\{\sigma_{M,v}(\neg\varphi) \mid v \in [w]_a\} = \bigvee\{\sigma_{M',v'}(\neg\varphi) \mid v' \in [w']_a\}$ and thus $\sigma_{M,w}(B_a^\psi\varphi) = \sigma_{M',w'}(B_a^\psi\varphi)$.

Take now the case where both translations are B_a^k-formulas. A similar argument as above gives $\bigvee\{\sigma_{M,v}(\psi \to \varphi) \mid v \in [w]_a\} = \bigvee\{\sigma_{M',v'}(\psi \to \varphi) \mid v' \in [w']_a\}$ and $\bigvee\{\sigma_{M,v}(\psi) \mid v \in [w]_a\} = \bigvee\{\sigma_{M',v'}(\psi) \mid v' \in [w']_a\}$. Letting k and k' be the indices chosen in the translation of $\sigma_{M,w}(B_a^\psi\varphi)$ and $\sigma_{M',w'}(B_a^\psi\varphi)$ respectively, we have $\sigma_{M,w}(B_a^\psi\varphi) = \sigma_{M',w'}(B_a^\psi\varphi)$ if $k = k'$. Assume towards a contradiction that $k > k'$. Lemma 4.36 now gives $Min_a(\llbracket\psi\rrbracket_M \cap [w]_a) \cap Min_a^{k'}[w]_a = \emptyset$, so $M,v \models \neg\psi$ for all $v \in Min_a^{k'}[w]_a$. With Lemma 4.38 we have $M,v \models \sigma_{M,v}(\neg\psi)$ for all $v \in Min_a^{k'}[w]_a$ and thus also that $M,w \models B_a^{k'} \bigvee\{\sigma_{M,v}(\neg\psi) \mid v \in [w]_a\}$. From Lemma 4.36 we also have $Min_a(\llbracket\psi\rrbracket_M \cap [w']_a) \subseteq Min_a^{k'}[w']_a$, so $M',v' \not\models \neg\psi$ for some $v' \in Min_a^{k'}[w']_a$. From here we use Lemma 4.38 to conclude $M',v' \not\models \sigma_{M',v'}(\neg\psi)$ for some $v' \in Min_a^{k'}[w']_a$ and thus $M',w' \not\models B_a^{k'} \bigvee\{\sigma_{M',v'}(\neg\psi) \mid v' \in [w']_a\}$. By the work done so far, this also means $M',w' \not\models B_a^{k'} \bigvee\{\sigma_{M,v}(\neg\psi) \mid v \in [w]_a\}$ which contradicts $(M,w) \equiv^D (M',w')$. The case when $k' > k$ is completely symmetrical, and the proof if thus concluded. □

Theorem 4.40. *On the class of (pre)image-finite models, modal equivalence for L^D implies bisimilarity.*

Proof. Let $M = (W, \geq, V)$ and $M' = (W', \geq', V')$ be two (pre)image-finite plausibility models. We first show that if $(M,w) \equiv^D (M',w')$ then $(M,w) \equiv^C (M',w')$. Assume $(M,w) \equiv^D (M',w,')$ and let γ be any formula of L^C.

$$\begin{aligned} M,w \models \gamma &\Leftrightarrow M,w \models \sigma_{M,w}(\gamma) && \text{(Lemma 4.38)}\\ &\Leftrightarrow M',w' \models \sigma_{M,w}(\gamma) && \text{(by assumption)}\\ &\Leftrightarrow M',w' \models \sigma_{M',w'}(\gamma) && \text{(Lemma 4.39)}\\ &\Leftrightarrow M',w' \models \gamma && \text{(Lemma 4.38)} \end{aligned}$$

Putting this together with Theorem 4.31 (modal equivalence for L^C implies bisimilarity, which requires (pre)image-finiteness), we have that two worlds which are modally equivalent in L^D are also modally equivalent in L^C and therefore bisimilar. □

Theorem 4.41 (Bisimulation characterization for L^D)**.** *Let $(M,w),(M',w')$ be epistemic plausibility models. On the class of (pre)image-finite models:*

$$(M,w) \leftrightarroweq (M',w') \textit{ iff } (M,w) \equiv^D (M',w')$$

Proof. From Theorem 4.35 and Theorem 4.40. □

4.4.3 Bisimulation correspondence for safe belief

We now show bisimulation characterisation results for the logic of degrees of belief L^S.

Theorem 4.42. *Bisimilarity implies modal equivalence for L^S.*

Proof. Assume $M_1 \underline{\leftrightarrow} M_2$. Then there is an autobisimulation R' on the disjoint union $M_1 \sqcup M_2$ with $R' \cap (\mathcal{D}(M_1) \times \mathcal{D}(M_2)) \neq \emptyset$. Extend R' into the maximal autobisimulation R on $M_1 \sqcup M_2$. Define $R_1 = R \cap (\mathcal{D}(M_1) \times \mathcal{D}(M_1))$ and $R_2 = R \cap (\mathcal{D}(M_2) \times \mathcal{D}(M_2))$.

Claim. Let $i \in \{1,2\}$ and $w \in \mathcal{D}(M_i)$. Then

(i) R_i is the maximal autobisimulation on M_i.

(ii) $Min_a([w]_{R^=} \cap [w]_a) = Min_a([w]_{R_i^=} \cap [w]_a)$.

(iii) For any v, $w \geq_a^R v$ iff $w \geq_a^{R_i} v$.

Proof of claim. To prove (i), let S_i denote the maximal autobisimulation on M_i. If we can show $S_i \subseteq R_i$ we are done. Since S_i is an autobisimulation on M_i, it must also be an autobisimulation on $M_1 \sqcup M_2$. Thus, clearly, $S_i \subseteq R$, since R is the maximal autobisimulation on $M_1 \sqcup M_2$. Hence, since $S_i \subseteq \mathcal{D}(M_i) \times \mathcal{D}(M_i)$, we get $S_i = S_i \cap (\mathcal{D}(M_i) \times \mathcal{D}(M_i)) \subseteq R \cap (\mathcal{D}(M_i) \times \mathcal{D}(M_i)) = R_i$. This shows $S_i \subseteq R_i$, as required.

We now prove (ii). Since $w \in \mathcal{D}(M_i)$ we get $[w]_a \subseteq \mathcal{D}(M_i)$. Since $R_i = R \cap (\mathcal{D}(M_i) \times \mathcal{D}(M_i))$ this implies $[w]_R \cap [w]_a = [w]_{R_i} \cap [w]_a$. Now note that since R is the maximal autobisimulation on $M_1 \sqcap M_2$ and R_i is the maximal autobisimulation on M_i, we have $R = R^=$ and $R_i = R_i^=$, by Proposition 4.12. Hence from $[w]_R \cap [w]_a = [w]_{R_i} \cap [w]_a$ we can conclude $[w]_{R^=} \cap [w]_a = [w]_{R_i^=} \cap [w]_a$, and then finally $Min_a([w]_{R^=} \cap [w]_a) = Min_a([w]_{R_i^=} \cap [w]_a)$.

We now prove (iii). Note that if $w \geq_a^R v$ or $w \geq_a^{R_i}$ then $w \sim_a v$ (by Lemma 4.5). So in proving $w \geq_a^R v \Leftrightarrow w \geq_a^{R_i} v$ for $w \in \mathcal{D}(M_i)$, we can assume that also $v \in \mathcal{D}(M_i)$. We then get:

$$\begin{aligned} w \geq_a^R v \quad &\Leftrightarrow \quad Min_a([w]_{R^=} \cap [w]_a) \geq_a Min_a([v]_{R^=} \cap [v]_a) \\ &\Leftrightarrow \quad Min_a([w]_{R_i^=} \cap [w]_a) \geq_a Min_a([v]_{R_i^=} \cap [v]_a) \quad \text{by (ii), since } w, v \in \mathcal{D}(M_i) \\ &\Leftrightarrow \quad w \geq_a^{R_i} v. \end{aligned}$$

This completes the proof of the claim.

We will now show that for all φ and all $(w_1, w_2) \in R \cap (\mathcal{D}(M_1) \times \mathcal{D}(M_2))$, if $M_1, w_1 \models \varphi$ then $M_2, w_2 \models \varphi$ (the other direction being symmetric). The proof is by induction on the syntactic complexity of φ. The propositional and knowledge cases are already covered by Theorem 4.30, so we only need to consider the case $\varphi = \Box_a \psi$. Hence assume $M_1, w_1 \models \Box_a \psi$ and $(w_1, w_2) \in R \cap (\mathcal{D}(M_1) \times \mathcal{D}(M_2))$. We need to prove $M_2, w_2 \models \Box_a \psi$. Pick an arbitrary $v_2 \in \mathcal{D}(M_2)$ with $w_2 \succeq_a v_2$. If we can show $M_2, v_2 \models \psi$, we are done. By (i), R_2 is the maximal autobisimulation on M_2. Hence $w_2 \succeq_a v_2$ by definition means $w_2 \geq_a^{R_2} v_2$. Using (iii), we can from $w_2 \geq_a^{R_2} v_2$ conclude $w_2 \geq_a^R v_2$. Since R is an autobisimulation, we can now apply [back$_{\geq}$] to $(w_1, w_2) \in R$ and $w_2 \geq_a^R v_2$ to get a v_1 with $w_1 \geq_a^R v_1$ and $(v_1, v_2) \in R$. Using (iii) again we can conclude from $w_1 \geq_a^R v_1$ to $w_1 \geq_a^{R_1} v_1$, since $w_1 \in \mathcal{D}(M_1)$. By (i), R_1 is the maximal autobisimulation on M_1, so $w_1 \geq_a^{R_1} v_1$ is by definition the same as $w_1 \succeq_a v_1$. Since we have assumed $M_1, w_1 \models \Box_a \psi$, and since $w_1 \succeq_a v_1$, we get $M_1, v_1 \models \psi$. Since $(v_1, v_2) \in R$, the induction hypothesis gives us $M_2, v_2 \models \psi$, and we are done. □

As for the previous logics, the converse also holds, that is, modal equivalence with regard to L^S implies bisimulation. This is going to be proved as follows. First we prove that any conditional belief formula φ_C can be translated into a logically equivalent safe belief formula φ_S. This implies that if two pointed models (M, w) and (M', w') are modally equivalent in L^S, they must also be modally equivalent in L^C: Any formula $\varphi_C \in L^C$ is true in (M, w) *iff* its translation $\varphi_S \in L^S$ is true in (M, w) *iff* φ_S is true in (M', w') *iff* φ_C is true in (M', w'). Now we can reason as follows: If two pointed models (M, w) and (M', w') are modally equivalent in L^S then they are modally equivalent in L^C and hence, by Theorem 4.31, bisimilar. This is the result we were after. We postpone the full proof until Section 4.5.1, which is where we provide the translation of conditional belief formulas into safe belief formulas (as part of a systematic investigation of the relations between the different languages and their relative expressivity). Here we only state the result:

Theorem 4.43. *On the class of (pre)image-finite models, modal equivalence for L^S implies bisimilarity.*

Proof. See Section 4.5.1. □

As for the two previous languages, L^C and L^D, we now get the following bisimulation characterization result.

Theorem 4.44 (Bisimulation characterization for L^S). *Let $(M, w), (M', w')$ be plausibility models. On the class of (pre)image-finite models:*

$$(M, w) \leftrightarrows (M', w') \text{ iff } (M, w) \equiv^S (M', w')$$

Proof. From Theorem 4.42 and Theorem 4.43. □

4.5 Expressivity

By combining Theorems 4.30, 4.35 and 4.42 from the previous section we immediately have the following result.

Corollary 4.45. *Let a plausibility model $M = (W, \geq, V)$ be given and let R be an autobisimulation on M. For any $(w, w') \in R$ we have that $(M, w) \equiv^{CDS} (M, w')$.*

Still there's more to the story than bisimulation and modal equivalence. In what follows we will gauge the relative expressive power (expressivity) of the logics under consideration. Abstractly speaking, expressivity is a yardstick for measuring whether two logics are able to capture the same properties of a class of models. More concretely in our case, we'll for instance be interested in determining whether the conditional belief modality can be expressed using the degrees of belief modality (observe that the translation in Section 4.4.2 depends on a particular model). With such results at hand we can for instance justify the inclusion or exclusion of a modality, and it also sheds light upon the strengths and weaknesses of our doxastic notions. To start things off we now formally introduce the notion of expressivity found in [van Ditmarsch et al., 2007].

Definition 4.46. Let L and L' be two logical languages interpreted on the same class of models.

- For $\varphi \in L$ and $\varphi' \in L'$, we say that φ and φ' are *equivalent* ($\varphi \equiv \varphi'$) iff they are true in the same pointed models of said class.[4]
- L' is *at least as expressive* as L ($L \leqq L'$) iff for every $\varphi \in L$ there is a $\varphi' \in L'$ s.t. $\varphi \equiv \varphi'$.
- L and L' are *equally expressive* ($L \equiv L'$) iff $L \leqq L'$ and $L' \leqq L$.
- L' is *more expressive* than L ($L < L'$) iff $L \leqq L'$ and $L' \nleqq L$.
- L and L' are *incomparable* ($L \bowtie L'$) iff $L \nleqq L'$ and $L' \nleqq L$.

Below we will show several cases where $L \nleqq L'$; i.e. that L' is *not* at least as expressive as L. Our primary modus operandi (obtained by logically negating $L \leqq L'$) will be to show that there is a $\varphi \in L$, where for any $\varphi' \in L'$ we can find two pointed models $(M, w), (M', w')$ such that

$$M, w \models \varphi, \quad M', w' \not\models \varphi \quad \text{and} \quad (M, w \models \varphi' \Leftrightarrow M', w' \models \varphi')$$

[4] With our usage of $\equiv$ it is clear from context whether we're referring to modal equivalence, formulas or languages.

In other words, for some $\varphi \in L$, no matter the choice of $\varphi' \in L'$, there will be models which φ distinguishes but φ' does not, meaning that $\varphi \not\equiv \varphi'$.

Our investigation will be concerned with the 7 distinct languages that are obtained by considering each L^X such that X is a non-empty subsequence of CDS. In Section 4.5.1 our focus is on safe belief, and in Section 4.5.2 on degrees of belief. Using these results, we provide in Section 4.5.3 a full picture of the relative expressivity of each of these logics, for instance showing that we can formulate 5 distinct languages up to equal expressivity. We find this particularly remarkable in light of the fact that our notion of bisimulation is the right fit for all our logics.

4.5.1 Expressivity of Safe Belief

Our first result relates to expressing conditional belief in terms of safe belief. Similar results can be found elsewhere in the literature, for instance in [Demey, 2011, Fact 31] and [Baltag and Smets, 2008b] where it is stated without proof. Below we prove that the identity found in [Demey, 2011] is also a valid identity in our logics, which is not a given as our semantics differ in essential ways. In particular the semantics of safe belief in [Demey, 2011] is a standard modality for $\geq_a$, whereas our semantics uses the derived relation $\succeq_a$. A more in-depth account of this matter is provided in Section 4.6. Returning to the matter at hand, we point out that our work in Section 4.4 actually serves our investigations here, as evident from the crucial role of Corollary 4.45 in the following proof.

Proposition 4.47. *The formula* $B_a^{\psi}\varphi \leftrightarrow (\widehat{K}_a\psi \to \widehat{K}_a(\psi \land \Box_a(\psi \to \varphi)))$ *is valid.*

Proof. We let $M = (W, \geq, V)$ be any plausibility model with $w \in W$, and further let $\succeq_a$ denote the normal plausibility relation for an agent a in M. We will show that $M, w \models B_a^{\psi}\varphi \leftrightarrow (\widehat{K}_a\psi \to \widehat{K}_a(\psi \land \Box_a(\psi \to \varphi)))$. To this end we let $X = Min_a([\![\psi]\!]_M \cap [w]_a)$. Immediately we have that if $X = \emptyset$ then no world in $[w]_a$ satisfies ψ, thus trivially yielding both $M, w \models B_a^{\psi}\varphi$ and $M, w \models \widehat{K}_a\psi \to \widehat{K}_a(\psi \land \Box_a(\psi \to \varphi))$. For the remainder we therefore assume X is non-empty. We now work under the assumption that $M, w \models B_a^{\psi}\varphi$ and show that this implies $M, w \models \widehat{K}_a\psi \to \widehat{K}_a(\psi \land \Box_a(\psi \to \varphi))$.

Claim 1. Let $x \in X$ be arbitrarily chosen, then $M, x \models \psi \land \Box_a(\psi \to \varphi)$.

Proof of claim 1. From $x \in X$ we have first that $M, x \models \psi \land \varphi$ and $w \sim_a x$. Since $M, x \models \psi$ this means we have proven Claim 1 if $M, x \models \Box_a(\psi \to \varphi)$ can be shown. To that effect, consider any $y \in W$ s.t. $x \succeq_a y$, for which we must prove $M, y \models \psi \to \varphi$. When $M, y \not\models \psi$ this is immediate, and so we may assume

$M, y \models \psi$. Since $x \succeq_a y$ we have $Min_a([x]_{R^=} \cap [x]_a) \geq_a Min_a([y]_{R^=} \cap [y]_a)$ with R being the maximal autobisimulation on M. Using the fact that R is maximal we have worlds x', y' in M such that $(y, y') \in R$, $x \geq_a x'$ and $x' \geq_a y'$. Applying Corollary 4.45 and $M, y \models \psi$ it follows that $M, y' \models \psi$. Using $\geq_a$-transitivity we have $x \geq_a y'$ and hence $w \sim_a x \sim_a y'$, allowing the conclusion that $y' \in X$. By assumption this means $M, y' \models \psi \wedge \varphi$, and so applying once more Corollary 4.45 it follows that $M, y \models \psi \rightarrow \varphi$ thus completing the proof of this claim.

To show $M, w \models \widehat{K}_a\psi \rightarrow \widehat{K}_a(\psi \wedge \Box_a(\psi \rightarrow \varphi))$ we take any $x \in X$, for which we have $w \sim_a x$ by definition of X. Combining this with Claim 1 it follows that $M, w \models \widehat{K}_a(\psi \wedge \Box_a(\psi \rightarrow \varphi))$. Consequently this also means that $M, w \models \widehat{K}_a\psi \rightarrow \widehat{K}_a(\psi \wedge \Box_a(\psi \rightarrow \varphi))$, thus completing the proof of this direction.

For the converse assume now that $M, w \models \widehat{K}_a\psi \rightarrow \widehat{K}_a(\psi \wedge \Box_a(\psi \rightarrow \varphi))$. As $X \neq \emptyset$ there is a world $u \in W$ s.t. $w \sim_a u$ and $M, u \models \psi \wedge \Box_a(\psi \rightarrow \varphi)$. Therefore we have $M, u' \models \psi \rightarrow \varphi$ for all $u \succeq_a u'$.

Claim 2. Let $x \in X$ be arbitrarily chosen, then $M, x \models \varphi$.

Proof of claim 2. From $x \in X$ we have that $M, x \models \psi$. If $u \succeq_a x$ it follows that $M, x \models \psi \rightarrow \varphi$, hence also $M, x \models \varphi$. Otherwise we have $u \not\succeq_a x$. Let R denote the maximal bisimulation on M and consider any $x' \in Min_a([x]_{R^=} \cap [x]_a)$. As $R^=$ and $\sim_a$ are both reflexive, we have $x \geq_a x'$. From $u \not\succeq_a x$ we therefore have a $u' \in Min_a([u]_{R^=} \cap [u]_a)$ s.t. $u' \not\geq_a x'$, $u \sim_a u'$ and $(u, u') \in R$ (thus also $x' >_a u'$). Since $u' \sim_a u$ and $u \sim_a w$ we have also $u' \sim_a w$, and additionally from $x \geq_a x'$ and $x' >_a u'$ we can conclude that $x \geq_a u'$ and $u' \not\geq_a x$. Using $M, u \models \psi$ and $(u, u') \in R$ we apply Corollary 4.45 which implies $M, u' \models \psi$. As $x \in X$, $u' \sim_a w$ and $x \geq_a u'$ it must be the case that $u' \in X$. From $u' \not\geq_a x$ we also have that $x \notin X$. This contradicts our assumption of $x \in X$, thereby showing $M, x \models \varphi$ and completing the proof of the claim.

Recalling that $M, w \models B_a^{\psi}\varphi$ iff $M, x \models \varphi$ for all $x \in X$, Claim 2 readily shows this direction, and thereby completes the proof. □

This result shows there is an equivalence-preserving translation from formulas in L^C to formulas in L^S, and so we have the following results.

Corollary 4.48. *For any $\varphi \in L^C$ there is a formula $\varphi' \in L^S$ s.t. $\varphi \equiv \varphi'$.*

Corollary 4.49. *$L^C \leq L^S$, $L^S \equiv L^{CS}$ and $L^{DS} \equiv L^{CDS}$.*

From Corollary 4.49 we have that any expressivity result for L^S also holds for L^{CS}, and similarly for L^{DS} and L^{CDS}. In other words, the conditional belief modality is superfluous in terms of expressivity when the safe belief modality is at our disposal. What is more, we can now finally give a full proof of Theorem 4.43.

of Theorem 4.43. Let (M,w) and (M',w') be (pre)image-finite plausibility models which are modally equivalent in L^S. For any $\varphi_C \in L^C$ it follows from Corollary 4.48 that there is a $\varphi_S \in L^S$ s.t. $\varphi_C \equiv \varphi_S$. Therefore

$$M,w \models \varphi_C \Leftrightarrow M,w \models \varphi_S \overset{\equiv^S}{\Longleftrightarrow} M',w' \models \varphi_S \Leftrightarrow M',w' \models \varphi_C$$

and hence $(M,w) \equiv^C (M',w')$. Using Theorem 4.31 we can conclude $(M,w) \leftrightarroweq (M',w')$ as required. □

We now proceed to show that L^{CD} is not at least as expressive as L^S. In doing so we need only work with $A = \{a\}$, meaning that the result holds even in the single-agent case. This is also true for our results in Section 4.5.2.

Lemma 4.50. *Let p,q be two distinct symbols in P, and let $M = (W,\geq,V)$ and $M' = (W',\geq',V')$ denote the two plausibility models presented in Figure 4.8. Then for $P' = P \setminus \{q\}$ we have that $(M,w_3) \equiv^{CD}_{P'} (M',w'_3)$.*

Proof. We prove the stronger result that for any $\varphi \in L^{CD}_{P'}$:

$$\text{for each } i \in \{1,2,3\} : (M,w_i \models \varphi) \Leftrightarrow (M',w'_i \models \varphi)$$

We proceed by induction on φ and let $i \in \{1,2,3\}$. When φ is a propositional symbol r in P', we have that $r \neq q$ and so $r \in V(w_i)$ iff $r \in V'(w'_i)$, thus completing the base case. Negation and conjunction are readily shown using the induction hypothesis.

For $\varphi = K_a\psi$ we have that $M,w_i \models K_a\psi$ iff $M,v \models \psi$ for all $v \in \{w_1,w_2,w_3\}$, since $[w_i]_a = \{w_1,w_2,w_3\}$. Applying the induction hypothesis to each element this is equivalent to $M',v' \models \psi$ for all $v' \in \{w'_1,w'_2,w'_3\}$ iff $M',w'_i \models K_a\psi$ (as $[w'_i]_a = \{w'_1,w'_2,w'_3\}$), which completes this case. Continuing to consider $\varphi = B^{\gamma}_a\psi$ we can simplify notation slightly, namely $Min_a([\![\gamma]\!]_M \cap [w_i]_a) = Min_a[\![\gamma]\!]_M$, since $[w_i]_a = W$ and $[\![\gamma]\!]_M \subseteq W$. The same holds for each world w'_i of M'.

Claim 1. For M and M' we have that $w_i \in Min_a[\![\gamma]\!]_M$ iff $w'_i \in Min_a[\![\gamma]\!]_{M'}$.

Proof of Claim 1. For M we have that $w_3 >_a w_2$ and $w_2 >_a w_1$, and similarly

$w'_3 >'_a w'_2$ and $w'_2 >'_a w'_1$ for M'. Thus the claim follows from the argument below.

$$\begin{aligned} w_i \in Min_a[\![\gamma]\!]_M &\Leftrightarrow \\ M, w_i \models \gamma \text{ and there is no } j < i \text{ s.t. } M, w_j \models \gamma &\overset{\text{(IH)}}{\Longleftrightarrow} \\ M', w'_i \models \gamma \text{ and there is no } j < i \text{ s.t. } M', w'_j \models \gamma &\Leftrightarrow \\ w'_i \in Min_a[\![\gamma]\!]_{M'} & \end{aligned}$$

We now have that $M, w_i \models B^\gamma_a\psi$ iff $M, v \models \psi$ for all $v \in Min_a[\![\gamma]\!]_M$. Applying both the induction hypothesis and Claim 1, we have that this is equivalent to $M, v' \models \psi$ for all $v' \in Min_a[\![\gamma]\!]_{M'}$ iff $M', w'_i \models B^\gamma_a\psi$.

Finally we consider the case of $\varphi = B^n_a\psi$. To this end we note that the union of $\{(w'_1, w'_3)\}$ and the identity relation on M' is the maximal bisimulation on M' (this relation cannot be extended and still satify [atoms]). As w'_1 and w'_3 are bisimilar, it follows from Corollary 4.45 that $M', w'_1 \models \psi$ iff $M', w'_3 \models \psi$ $(*)$.

Claim 2. For $n \in \mathbb{N}$ we have that $M, w \models \psi$ for all $w \in Min^n_a[w_i]$ iff $M', w' \models \psi$ for all $w' \in Min^n_a[w'_i]$.

Proof of Claim 2. We treat three exhaustive cases for n.

- $n = 0$: $M, w \models \psi$ for all $w \in Min^0_a[w_i] \Leftrightarrow M, w_1 \models \psi \overset{\text{(IH)}}{\Longleftrightarrow} M', w'_1 \models \psi \overset{(*)}{\Longleftrightarrow} M', w'_3 \models \psi$. Therefore $M, w \models \psi$ for all $w \in Min^0_a[w_i]$ is equivalent to $M', w' \models \psi$ for all $w' \in \{w'_1, w'_3\}$, and as $Min^0_a[w'_i] = \{w'_1, w'_3\}$ this concludes this case.
- $n = 1$: Since $Min^1_a[w_i] = \{w_1, w_2\}$ we have that $M, w \models \psi$ for all $w \in \{w_1, w_2\} \overset{\text{(IH)}}{\Longleftrightarrow} M', w' \models \psi$ for all $w' \in \{w'_1, w'_2\}$. Using $(*)$ this is equivalent to $M', w' \models \psi$ for all $w' \in \{w'_1, w'_2, w'_3\}$. By this argument and the fact that $Min^1_a[w'_i] = \{w'_1, w'_2, w'_3\}$, we can conclude $M, w \models \psi$ for all $w \in Min^1_a[w_i]$ $\Leftrightarrow M', w' \models \psi$ for all $w' \in Min^1_a[w'_i]$ as required.
- $n \geq 2$: For $m \geq 2$ we have that $Min^m_a[w_i] = \{w_1, w_2, w_3\}$ and $Min^m_a[w'_i] = \{w'_1, w'_2, w'_3\}$, hence this is exactly as the case of $\varphi = K\psi$.

We have that $M, w_i \models B^n_a\psi$ iff $M, w \models \psi$ for all $w \in Min^n_a[w_i]$. Applying Claim 2 this is equivalent to $M', w' \models \psi$ for all $w' \in Min^n_a[w'_i]$ iff $M', w'_i \models B^n_a\psi$, thereby

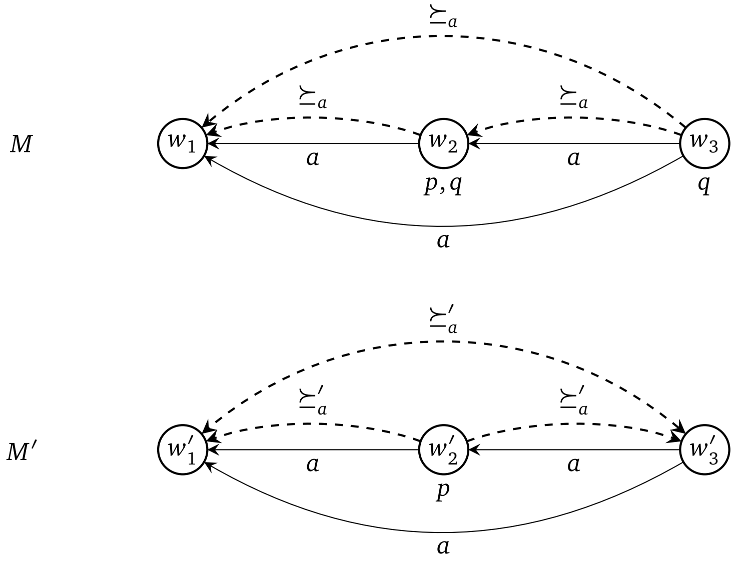

Figure 4.8: Two single-agent plausibility models and their normal plausibility relations (dashed arrows). As usual reflexive arrows are omitted.

completing the final case of the induction step. It follows that $(M, w_3) \equiv^{CD}_{P'} (M', w'_3)$ as required. □

Proposition 4.51. $L^S \not\leq L^{CD}$.

Proof. Consider the formula $\Diamond_a p$ of L^S with $p \in P$, and take some arbitrary formula $\varphi_{CD} \in L^{CD}_P$. As φ_{CD} is finite and P is countably infinite, there will be *some* $q \neq p$ not occurring in φ_{CD}. Letting $P' = P \setminus \{q\}$ this means that $\varphi_{CD} \in L^{CD}_{P'}$. This choice of p and q can always be made, and consequently there also exists models M and M' as given in Figure 4.8. The maximal bisimulation on M is the identity as no two worlds have the same valuation. At the same time $\{(w'_1, w'_1), (w'_1, w'_3), (w'_3, w'_3), (w'_2, w'_2)\}$ is the maximal bisimulation on M'. This gives rise to the normal plausibility relations $\succeq_a$ (for M) and $\succeq'_a$ (for M') depicted in Figure 4.8 using dashed edges.

Since $w_3 \succeq_a w_2$ and $M, w_2 \models p$ it follows that $M, w_3 \models \Diamond_a p$. Furthermore we have that the image of w'_3 under $\succeq'_a$ is $\{w'_1, w'_3\}$. This means that there is no $v' \in W'$ s.t. $w'_3 \succeq'_a v'$ and $M', v' \models p$, and consequently $M', w'_3 \not\models \Diamond_a p$. At the same time we have by Lemma 4.50 that $M, w_3 \models \varphi_{CD}$ iff $M', w'_3 \models \varphi_{CD}$. Therefore using the formula $\Diamond_a p$ of L^S, for any formula of $\varphi_{CD} \in L^{CD}$ there are models which $\Diamond_a p$ distinguishes but φ_{CD} does not, and so $\Diamond_a p \not\equiv \varphi_{CD}$. Consequently we have $L^S \not\leq L^{CD}$ as required. □

To further elaborate on this result, what is really being to put to use here is the ability of the safe belief modality to (at least in part) talk about propositional symbols that do not occur in a formula. This is an effect of the derived relation $\succeq_a$ depending on the maximal bisimulation.

4.5.2 Expressivity of Degrees of Belief

We have now settled that safe belief is more expressive than conditional belief, and further that the combination of the conditional belief modality and the degrees of belief modality does not allow us to express the safe belief modality. A hasty conclusion would be that the safe belief modality is the one modality to rule them all, but this is not so. In fact L^S (equivalent to L^{CS} cf. Corollary 4.49) falls short when it comes to expressing degrees of belief, which we now continue to prove.

Lemma 4.52. *Let p, q be two distinct symbols in P, and let $M = (W, \geq, V)$ and $M' = (W', \geq', V')$ denote the two plausibility models presented in Figure 4.9. Then for $P' = P \setminus \{q\}$ we have that $(M, x_1) \equiv^S_{P'} (M', x')$.*

Proof. We show first show the following result for formulas without the conditional belief modality, namely for $i \in \{1,2\} : (M, x_i) \equiv^S_{P'} (M', x')$ and $(M, y) \equiv^S_{P'} (M', y')$. We proceed by induction on $\varphi \in L^S_{P'}$, showing that:

$$\text{for } i \in \{1,2\} : M, x_i \models \varphi \text{ iff } M', x' \models \varphi \qquad \text{and} \qquad M, y \models \varphi \text{ iff } M', y' \models \varphi.$$

For the base case we have $\varphi = r$ for some $r \in P \setminus \{q\}$. Because $r \neq q$ it is clear that $r \in V(x_1)$ iff $r \in V'(x')$. Since we also have $V(x_2) = V'(x')$ and $V'(y) = V(y')$ this completes the base case. The cases of negation and conjunction are readily established using the induction hypothesis, and $\varphi = K_a\psi$ is shown just as we did in the proof of Lemma 4.50. Before proceeding we recall that $A = \{a\}$ and note that for any $w \in W$ we have $[w]_a = \{x_1, x_2, y\}$, as well as $[w']_a = \{x', y'\}$ for any $w' \in W'$. Moreover, the maximal bisimulation on M and M' respectively is the identity relation, meaning that $\geq_a = \succeq_a$ and $\geq'_a = \succeq'_a$. For the case of $\varphi = \Box_a\psi$ we can therefore argue as follows.

$$M, x_1 \models \Box_a\psi \Leftrightarrow M, x_1 \models \psi \overset{\text{(IH)}}{\Longleftrightarrow} M', x' \models \psi \Leftrightarrow M', x' \models \Box_a\psi$$

$$M, x_2 \models \Box_a\psi \Leftrightarrow (\forall i \in \{1,2\} : M, x_i \models \psi) \overset{\text{(IH)}}{\Longleftrightarrow} M', x' \models \psi \Leftrightarrow M', x' \models \Box_a\psi$$

$$M, y \models \Box_a\psi \Leftrightarrow (\forall w \in W : M, w \models \psi) \overset{\text{(IH)}}{\Longleftrightarrow} (\forall w' \in W' : M', w' \models \psi) \Leftrightarrow M', y' \models \Box_a\psi$$

In fact the last line is essentially the case of $K_a\psi$, as the image of y under $\succeq_a$ is W (and W' is the image of y' under $\succeq'_a$). Completing the induction step means that $M, x_1 \models \varphi$ iff $M', x' \models \varphi$, thus we can conclude $(M, x_1) \equiv^S_{P'} (M', x')$. □

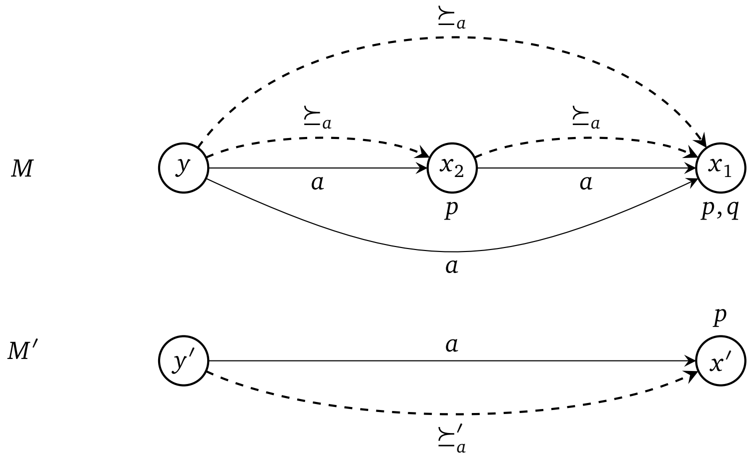

Figure 4.9: Two single-agent plausibility models and their normal plausibility relations (dashed arrows). As usual reflexive arrows are omitted.

Proposition 4.53. $L^D \nleq L^S$.

Proof. Consider the formula $B^1_a p \in L^D$ with $p \in P$, and additionally take any formula $\varphi_S \in L^S_P$. As φ_S is finite and P is countably infinite, there will be *some* $q \neq p$ which does not occur in φ_S. With $P' = P \setminus \{q\}$ we therefore have $\varphi_S \in L^S_{P'}$. As we can always make such a choice of p and q, this means that there always exists models (M, x_1), (M', x') of the form given in Figure 4.9.

Observe that the maximal bisimulation on M is the identity as no two worlds have the same valuation. The same goes for M', and so $Min^1_a[x_1]_a = \{x_1, x_2\}$ and $Min^1_a[x']_a = \{x', y'\}$. Consequently $M, x_1 \models B^1_a p$ whereas $M', x' \not\models B^1_a p$. Since $\varphi_S \in L^S_{P'}$ it follows from Lemma 4.52 that $M, x \models \varphi_S$ iff $M', x' \models \varphi_S$. What this proves is that using the formula $B^1_a p$ of L^D, no matter the choice of formula φ_S of L^S there will be models which $B^1_a p$ distinguishes but φ_S does not, hence $B^1_a p \not\equiv \varphi_S$. From this follows $L^D \nleq L^S$ as required. □

We find that this result is quite surprising. Again it is a consequence of our use of the maximal bisimulation when defining our semantics. The purpose of x_1 in model M (which is otherwise identical to M') is to inject an additional belief sphere, and to do so without adding any factual content which the safe belief modality might use to distinguish x_1 from x_2.

At this point it might seem as if all hope was lost for the conditional belief modality, however our final direct result somewhat rebuilds the reputation of this hardpressed modality. To this end we define for any $k \in \mathbb{N}$ the language L^{Dk}, which contains every formula of L^D for which if $B^n_a\varphi$ occurs then $n \leq k$. In other words formulas of L^{Dk} talk about belief to at most degree k, which comes in handy as we investigate the relative expressive power of L^D and L^C.

Lemma 4.54. *Let $k \in \mathbb{N}$ be given, and let (M^k, w_0) and (N^k, w'_0) denote the two plausibility models presented in Figure 4.10. Then we have that (M^k, w_0) and (N^k, w'_0) are modally equivalent in L^{Dk}.*

Proof. We prove a stronger version of this lemma, namely that $(M^k, w_i) \equiv^{Dk} (N^k, w'_i)$ for $0 \leq i \leq k$, $(M^k, x) \equiv^{Dk} (N^k, x')$ and $(M^k, y) \equiv^{Dk} (N^k, y')$.

Key to this proof is the fact that x (resp. y) has the same valuation as x' (resp. y'), and that x is more plausible than y whereas y' is more plausible than x'. We proceed by induction on $\varphi \in L^{Dk}$. In the base case φ is a propositional symbol, and so as the valuation of each w_i matches that of w'_i $(0 \leq i \leq k)$, x matches x' and y matches y' this completes the base case. The cases of negation and conjunction readily follow using the induction hypothesis, and for $\varphi = K_a\psi$ the argument is essentially that used in the proof of Lemma 4.50.

Lastly we consider $\varphi = B^j_a\psi$ for any $0 \leq j \leq k$, and recall that this is sufficient as $\varphi \in L^{Dk}_P$. As neither model contains two worlds with the same valuation, the maximal autobisimulation on either model is the identity, and so both models are normal. With the epistemic relation of agent a being total, we have for all $w \in W$ that $Min^j_a[w]_a = \{w_0, \dots, w_j\}$ and similarly for all $w' \in W'$ that $Min^j_a[w']_a = \{w'_0, \dots, w'_j\}$. We therefore have

$$\forall w \in W : M^k, w \models B^j_a\psi \Leftrightarrow \forall v \in \{w_0, \dots, w_j\} : M^k, v \models \psi \overset{\text{(IH)}}{\Longleftrightarrow}$$
$$\forall v' \in \{w'_0, \dots, w'_j\} : N^k, v' \models \psi \Leftrightarrow \ \forall w' \in W' : N^k, w' \models B^j_a\psi$$

as required. Observe that we can apply the induction hypothesis since $j \leq k$, and that importantly x, y are not in $Min^j_a[w]_a$, and x', y' are not in $Min^j_a[w']_a$. Thus we have shown that $(M^k, w_0) \equiv^{Dk} (N^k, w'_0)$ thereby completing the proof. □

Proposition 4.55. $L^C \not\leq L^D$.

Proof. Consider now $B^q_a r$ belonging to L^C and any formula $\varphi_D \in L^D$. Since φ_D is finite we can choose some $k \in \mathbb{N}$ such that $\varphi_D \in L^{Dk}$. Because $p_0, \dots, p_k, q, r$ is taken from the countably infinite set P, no matter the choice of k there exists pointed plausibility models (M^k, w_0) and (N^k, w'_0) as presented in Figure 4.10.

To determine the truth of $B^q_a r$ in (M^k, w_0) and (N^k, w'_0) respectively we point out that $[[q]]_{M^k} = \{x, y\}$ and $[[q]]_{N^k} = \{y', x'\}$. Therefore we have that $Min_a([[q]]_{M^k} \cap$

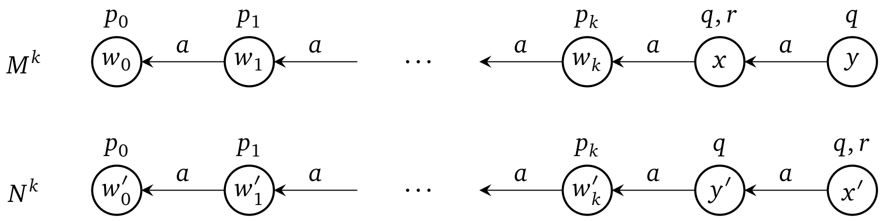

Figure 4.10: Two single-agent plausibility models. We've omitted reflexive arrows and for the sake of readability also some transitive arrows.

$[w_0]_a) = \{x\}$ and $Min_a([[q]]_{N^k} \cap [w'_0]_a) = \{y'\}$. Since $M^k, x \models r$ and $N^k, y' \not\models r$, it follows $M^k, w_0 \models B^q_a r$ whereas $N^k, w'_0 \not\models B^q_a r$. By Lemma 4.54 we have that $M^k, w_0 \models \varphi_D$ iff $N^k, w'_0 \models \varphi_D$. With this we have shown that taking the formula $B^q_a r$ of L^C, there is for any $\varphi_D \in L^D$ pointed plausibility models which $B^q_a r$ distinguishes but φ_D does not, thus $B^q_a r \not\equiv \varphi_D$. It follows that $L^C \not\leqq L^D$ as required. □

We've now shown that the degrees of belief modality cannot capture the conditional belief modality. What this really showcases is that for $B^\psi_a \varphi$, ψ potentially enables us to talk about worlds of arbitrarily large degree. This sets it apart from the degrees of belief modality, and causes for instance a difference in expressivity.

4.5.3 Mapping Out the Relative Expressive Power

With the results we've now shown, we're in fact able to determine the relative expressivity of all our languages. To this end we make use of the following facts related to expressivity, where we let L, L' and L'' denote logical languages interpreted on the same class of models:

(a) If L is a sublanguage of L' then $L \leqq L'$.

(b) If $L \leqq L'$ and $L' \leqq L''$ then $L \leqq L''$ (transitivity).

(c) If $L \equiv L'$ then $L \leqq L''$ iff $L' \leqq L''$ (transitivity consequence 1).

(d) If $L \leqq L'$ and $L'' \not\leqq L'$ then $L'' \not\leqq L$ (transitivity consequence 2).

(e) If $L \leqq L'$ and $L \not\leqq L''$ then $L' \not\leqq L''$ (transitivity consequence 3).

Now comes our main result, which shows the relative expressivity between the logic of conditional belief, the logic of degrees of belief and the logic of safe belief.

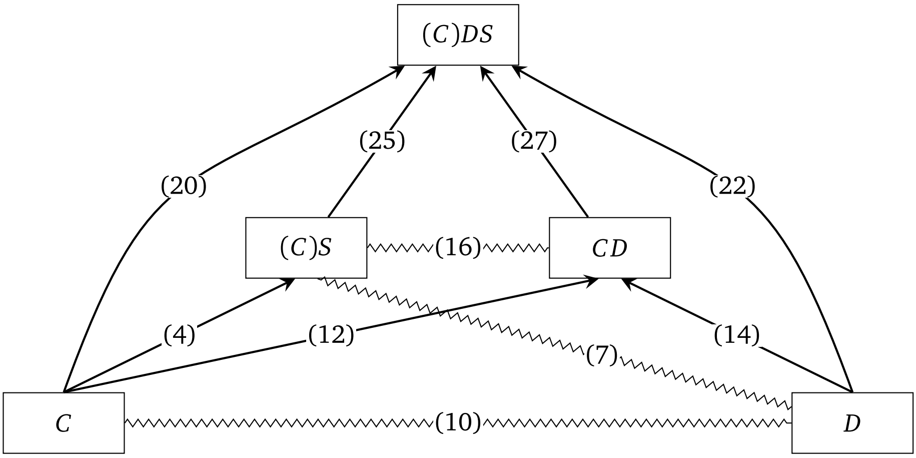

Figure 4.11: Summary of expressivity results for our logics. An arrow $X \longrightarrow X'$ indicates that $L^{X'}$ is more expressive than L^X. A zig-zag line between X and X' means that L^X and $L^{X'}$ are incomparable. The abbreviation $(C)DS$ means both CDS and DS, and similarly for $C(S)$ indicating both CS and S. Labels on arrows and zig-zag lines signify from where the result is taken in Table 4.1.

Theorem 4.56. $L^C < L^S$, $L^C \bowtie L^D$, $L^D \bowtie L^S$.

Proof. See the derivation of (4), (7) and (10) in Table 4.1. □

Beyond showing the above theorem, Table 4.1 fully accounts for the relative expressivity between L^C, L^D, L^S, L^{CD} and L^{DS}. Finally, using Corollary 4.49 and property (c) we have that any expressivity result for L^S holds for L^{CS} and similarly for L^{DS} and L^{CDS}. A more pleasing presentation of these results is found in Figure 4.11.

4.5.4 Reflection on bisimulation characterization and expressivity

Our bisimulation characterization results are that, on (pre)image-finite models:

$(M,w) \leftrightarroweq (M',w')$	iff	$(M,w) \equiv^C (M',w')$	Theorem 4.32
$(M,w) \leftrightarroweq (M',w')$	iff	$(M,w) \equiv^D (M',w')$	Theorem 4.41
$(M,w) \leftrightarroweq (M',w')$	iff	$(M,w) \equiv^S (M',w')$	Theorem 4.44

#	Result	Inferred from
(1)	$L^C \leqq L^S$	Corollary 4.49.
(2)	$L^S \nleqq L^{CD}$	Proposition 4.51.
(3)	$L^S \nleqq L^C$	$L^C \leqq L^{CD}$ from (a), $L^S \nleqq L^{CD}$ from (2) and applying (d).
(4)	$L^C < L^S$	$L^C \leqq L^S$ from (1), $L^S \nleqq L^C$ from (3).
(5)	$L^D \nleqq L^S$	Proposition 4.53.
(6)	$L^S \nleqq L^D$	$L^D \leqq L^{CD}$ from (a), $L^S \nleqq L^{CD}$ from (2) and applying (d).
(7)	$L^D \bowtie L^S$	$L^D \nleqq L^S$ from (5), $L^S \nleqq L^D$ from (6).
(8)	$L^C \nleqq L^D$	Proposition 4.55.
(9)	$L^D \nleqq L^C$	$L^C \leqq L^S$ from (1), $L^D \nleqq L^S$ from (5) and applying (d).
(10)	$L^C \bowtie L^D$	$L^C \nleqq L^D$ from (8), $L^D \nleqq L^C$ from (9).
(11)	$L^{CD} \nleqq L^C$	$L^D \leqq L^{CD}$ from (a), $L^D \nleqq L^C$ from (9) and applying (e).
(12)	$L^C < L^{CD}$	$L^C \leqq L^{CD}$ from (a), $L^{CD} \nleqq L^D$ from (13).
(13)	$L^{CD} \nleqq L^D$	$L^C \leqq L^{CD}$ from (a), $L^C \nleqq L^D$ from (8) and applying (e).
(14)	$L^D < L^{CD}$	$L^D \leqq L^{CD}$ from (a), $L^{CD} \nleqq L^D$ from (13).
(15)	$L^{CD} \nleqq L^S$	$L^D \leqq L^{CD}$ from (a), $L^D \nleqq L^S$ from (5) and applying (e).
(16)	$L^S \bowtie L^{CD}$	$L^S \nleqq L^{CD}$ from (2), $L^{CD} \nleqq L^S$ from (15).
(17)	$L^{CDS} \leqq L^{DS}$	$L^{CDS} \equiv L^{DS}$ from Corollary 4.49 and Definition 4.46.
(18)	$L^C \leqq L^{DS}$	$L^C \leqq L^{CDS}$ from (a), $L^{CDS} \leqq L^{DS}$ from (17) and applying (b).
(19)	$L^{DS} \nleqq L^C$	$L^S \leqq L^{DS}$ from (a), $L^S \nleqq L^C$ from (3) and applying (e).
(20)	$L^C < L^{DS}$	$L^C \leqq L^{DS}$ from (18), $L^{DS} \nleqq L^C$ from (19).
(21)	$L^{DS} \nleqq L^D$	$L^S \leqq L^{DS}$ from (a), $L^S \nleqq L^D$ from (6) and applying (e).
(22)	$L^D < L^{DS}$	$L^D \leqq L^{DS}$ from (a), $L^{DS} \nleqq L^D$ from (21).
(23)	$L^{CD} \leqq L^{DS}$	$L^{CD} \leqq L^{CDS}$ from (a), $L^{CDS} \leqq L^{DS}$ from (17) and applying (b).
(24)	$L^{DS} \nleqq L^S$	$L^{CD} \leqq L^{DS}$ from (23), $L^{CD} \nleqq L^S$ from (15) and applying (e).
(25)	$L^S < L^{DS}$	$L^S \leqq L^{DS}$ from (a), $L^{DS} \nleqq L^S$ from (24).
(26)	$L^{CD} \nleqq L^{DS}$	$L^S \leqq L^{DS}$ from (a), $L^S \nleqq L^{CD}$ from (2) and applying (e).
(27)	$L^{CD} < L^{DS}$	$L^{CD} \leqq L^{DS}$ from (23), $L^{CD} \nleqq L^{DS}$ from (26).

Table 4.1: Derivation of the relative expressivity of our logics. Each of the references (a), (b), (d) and (e) refer to properties stated at the start of Section 4.5.3. Bold faced numbers are illustrated in Figure 4.11.

In other words, bisimulation corresponds to modal equivalence in all three logics. Our expressivity results can be summarised as (Theorem 4.56)

$$\begin{array}{ccc} L^C & < & L^S \\ L^C & \bowtie & L^D \\ L^D & \bowtie & L^S \end{array}$$

The logic of conditional belief is less expressive than the logic of safe belief, the logic of conditional belief and the logic of degrees of belief are incomparable, as are the logic of degrees of belief and the logic of safe belief.

The former results seem to suggest that, in some sense, the three logics are the same, whereas the latter results seem to suggest that, in another sense, the three logics are different. It is therefore a good moment to explain how to interpret our results.

The bisimulation characterization results say that the information content of a given ((pre)image-finite) plausibility model is equally well described in the three logics. An obvious corollary of these results is

Corollary 4.57.

$$\begin{array}{lcl} (M,w) \equiv^C (M',w') & \textit{iff} & (M,w) \equiv^D (M',w') \\ (M,w) \equiv^C (M',w') & \textit{iff} & (M,w) \equiv^S (M',w') \\ (M,w) \equiv^D (M',w') & \textit{iff} & (M,w) \equiv^S (M',w') \end{array}$$

Now consider an even more specific case: a finite model; and consider a characteristic formula of that model (these can be shown to exist for plausibility models along the lines of [van Benthem, 2006, van Ditmarsch et al., 2012]—where we note that we take models, not pointed models). For a model M this gives us, respectively, formulas φ^C_M, φ^D_M, and φ^S_M. Then the bisimulation characterization results say that φ^C_M, φ^D_M, and φ^S_M are all equivalent. Now a characteristic formula is a very special formula with a unique model (modulo bisimilarity). For other formulas that do not have a singleton denotation (again, modulo bisimilarity) in the class of plausibility models, this equivalence cannot be achieved. That is the expressivity result. For example, given that $L^C < L^S$, there is a safe belief formula that is not equivalent to any conditional belief formula. This formula should then describe a property that has several non-bisimilar models. It is indeed the case that the formula $\Diamond_a p$ used in the proof of Proposition 4.51 demonstrating $L^C < L^S$ has many models! It is tempting to allow ourselves a simplication and to say that the expressivity hierarchy breaks down if we restrict ourselves to formulas with unique models.[5]

Finally, we must point out that in the publication on single-agent bisimulation [Andersen et al., 2013, p. 285], we posed the following conjecture:

In an extended version of the paper we are confident that we will prove

[5] If we consider infinitary versions of the modalities in our logical languages, in other words, common knowledge and common belief modalities, we preserve the bisimulation characterization results (for a more refined notion of bisimulation) but it is then to be expected that all three logics become equally expressive (oral communication by Tim French).

> *that the logics of conditional belief and knowledge, of degrees of belief and knowledge, and both with the addition of safe belief are all expressively equivalent.*

It therefore seems appropriate to note that we have proved our own confident selves resoundingly wrong!

4.6 Comparison and applications

We compare our bisimulation results to those in Demey's work [Demey, 2011], our expressivity results to those obtained in Baltag and Smets' [Baltag and Smets, 2008b], and finally discuss the relevance of our results for epistemic planning [Bolander and Andersen, 2011].

Bisimulation Prior to our work Demey discussed the model theory of plausibility models in great detail in [Demey, 2011]. Our results add to the valuable original results he obtained. Demey does not consider degrees of belief; he considers knowledge, conditional belief and safe belief. Our plausibility models are what [Demey, 2011] refers to as uniform and locally connected epistemic plausibility models; he also considers models with fewer restrictions on the plausibility function. But given [Demey, 2011, Theorem 35], these types of models are for all intents and purposes equivalent to ours. The semantics for conditional belief and knowledge are as ours, but his semantics for safe belief is different (namely as in [Baltag and Smets, 2008b]). The difference is that in his case an agent safely believes φ if φ is true in all worlds as least as plausible as the current world, whereas in our case it is like that but **in the normalised model**. This choice of semantics has several highly significant implications as we will return to shortly.

In line with his interpretation of safe belief as a standard modality, Demey's notion of bisimulation for plausibility models is also standard. For example, whereas we require that

$$[\text{forth}_{\geq}]\ \text{If } v \in W \text{ and } w \geq_a^R v,\ \exists v' \in W \text{ such that } w' \geq_a^R v' \text{ and } (v, v') \in R$$

he requires that

$$[\text{forth}_{\geq}]\ \text{If } v \in W \text{ and } w \geq_a v,\ \exists v' \in W \text{ such that } w' \geq_a v' \text{ and } (v, v') \in R$$

He obtains correspondence for bisimulation and modal equivalence in the logic of safe belief in [Demey, 2011, Footnote 12 and Theorem 32]. Our notion of bisimulation is less restrictive, as we will now illustrate by way of the examples in Figure 4.12.

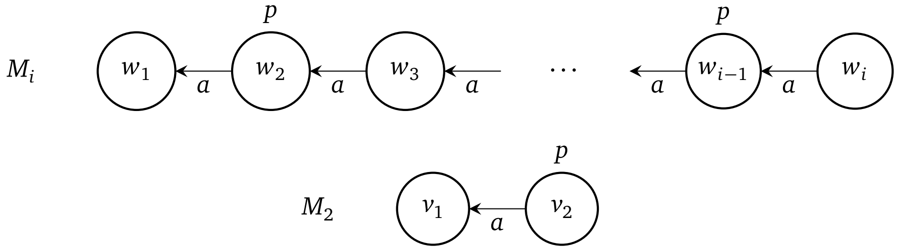

Figure 4.12: According to Demey's notion of bisimulation, model M_i (above) with alternating $\neg p$ and p worlds is a bisimulation contraction. In this particular case i is odd as p does not hold at w_i. According to our notion of bisimulation, all p worlds in model M_i are bisimilar and also all $\neg p$ worlds. Model M_2 (below) is the contraction.

Consider model M_i in Figure 4.12. This is a single-agent model on a single proposition p containing i worlds, where the image of a world w_j under $\geq_a$ is $\{w_1, \dots, w_j\}$. The valuation is such that if the index of a world is even then p holds, and otherwise p does not hold. Now, using Demey's notion of bisimulation entails that the maximal autobisimulation on M_i is the identity, and thus M_i is a bisimulation contraction. For example, we can find a formula that distinguishes (M_i, w_i) from (M_{i+2}, w_{i+2}). For safe belief $\Box$ we now have Demey's semantics (see above) $M, w \models \Box_a \varphi$ iff $M, v \models \varphi$ for all v with $w \geq_a v$. We now define $\varphi_0 = \top$ and for any natural number $n \geq 1$ we let:

$$\varphi_n = \begin{cases} \Diamond_a(\varphi_{n-1} \wedge p) & \text{if } n \text{ is even;} \\ \Diamond_a(\varphi_{n-1} \wedge \neg p) & \text{if } n \text{ is odd;} \end{cases}$$

for example

$$\varphi_4 = \Diamond_a(\Diamond_a(\Diamond_a(\Diamond_a(\top \wedge \neg p) \wedge p) \wedge \neg p) \wedge p).$$

We now have that for any $i \geq 1$, $M_i, w_i \models \varphi_i \wedge \neg \varphi_{i+1}$, which makes this a distinguishing formula between (M_i, w_i) from (M_{i+2}, w_{i+2}). In fact, the semantics of $\Box_a$ allow us to count the number of worlds in M_i. In this sense Demey's logic is immensely expressive.

Again referring to Figure 4.12, consider M_3, the model with a most plausible $\neg p$ world, a less plausible p world and an even less plausible $\neg p$ world. In the logic L^C of conditional belief w_1 and w_3 of M_3 are modally equivalent. Hence they also ought to be bisimilar. But in Demey's notion of bisimilarity they are not. Hence we have a mismatch between modal equivalence and bisimilarity, which is not supposed to happen: it is possible for two worlds to be modally equivalent but not bisimilar. Demey also was aware of this, of course. To remedy the problem one can either strengthen the notion of modal equivalence or weaken the notion of bisimilarity. Demey chose the former (namely by adding the safe belief modality to the conditional belief modality), we chose the latter. Thus we regain the correspondence between bisimilarity and modal equivalence. Baltag and Smets [Baltag and Smets, 2008b] achieve the same via a different route: they include in the language special propositional symbols, so-called S-propositions. The denotation of an S-proposition can be any subset of the domain. This therefore also makes the language much more expressive.

We believe that in particular for application purposes, weakening the notion of bisimulation, as we have done, is preferable over strengthening the logic, as in [Baltag and Smets, 2008b, Demey, 2011]. This come at the price of a more complex bisimulation definition (and, although we did not investigate this, surely a higher complexitity of determining whether two worlds are bisimilar), but, we venture to observe, also a very elegant bisimulation definition given the ingenious use of the bisimulation relation itself in the definition of the forth and back conditions of bisimulation. We consider this one of the highlights of our work.

Expressivity In [Baltag and Smets, 2008b] one finds many original expressivity results. Our results copy those, but also go beyond. We recall Table 4.1 for the full picture of our results, and the main results of those namely $L^C < L^S$, $L^C \bowtie L^D$, and $L^D \bowtie L^S$. The first, $L^C < L^S$, is originally found in [Baltag and Smets, 2008b, page 34, equation 1.7], and we obtained it using the same embedding translation. However, it may be worth to point out that in our case this translation still holds for the (in our opinion) more proper bisimulation preserving notion of safe belief. Baltag and Smets' S-propositions are arbitrary subsets of the domain, the (unnecessarily) far more expressive notion of safe belief. Baltag and Smets also discuss degrees of belief but do not obtain expressivity results for that, so $L^C \bowtie L^D$ may be considered novel and interesting. In artificial intelligence, the degrees of belief notion seems more widely in use than the conditional belief notion, so an informed reader had better be aware of the incomparability of both logics and may choose the logic to suit his or her needs. The result that $L^D \bowtie L^S$ could possibly also be considered unexpected, and therefore valuable.

Planning An application area of plausibility models is epistemic planning. A consequence of Demey's notion of bisimulation is that even for single-agent models on a finite set of propositions, the set of distinct, contraction-minimal pointed plausibility models is infinite. For example, we recall that in Figure 4.12 any two pointed plausibility models in $\{(M_i, w_i) \mid i \in \mathbb{N}\}$ are non-bisimilar. With our notion of bisimulation, there are in the single-agent case only finitely many distinct pointed plausibility models up to bisimulation. This was already reported in [Andersen et al., 2013]. Our motivation for this bisimulation investigation was indeed prompted by the application of doxastic logics in planning.

In planning, an agent attempt to find a sequence of action, a plan, that achieves a given goal. A planning problem implicitly represents a state-transition system, where transitions are induced by actions. By exploring this state-space we can reason about actions and synthesise plans. A growing community investigates planning by applying dynamic epistemic logics [Bolander and Andersen, 2011, Löwe et al., 2011b, Andersen et al., 2012], where actions are epistemic actions. Planning with doxastic modalities has also been considered [Andersen et al., 2014]. This is done by identifying states with (pointed) plausibility models, and the goal with a formula of the doxastic language. Epistemic actions can be public actions, like hard and soft announcements [van Benthem, 2007], but also non-public actions, such as event models [Baltag and Smets, 2008b].

With the state-space consisting of plausibility models, model theoretic results become pivotal when deciding the plan existence problem. Unlike Demey's approach, our framework leads to a finite state-space in the single-agent case and therefore the single-agent plan existence problem is decidable [Bolander and Andersen, 2011]. At the same time we know that even in a purely epistemic setting the multi-agent plan existence problem is undecidable [Bolander and Andersen, 2011]. But by placing certain restrictions on the planning problem it is possible to find decidable fragments even in the multi-agent case, for example, event models with propositional preconditions [Yu et al., 2013].

Acknowledgements

Hans van Ditmarsch is also affiliated to IMSc (Institute of Mathematical Sciences), Chennai, as research associate. He acknowledges support from European Research Council grant EPS 313360.

Chapter 5

Multi-agent plausibility planning

In this chapter I present work in progress on generalising plausibility planning to use multi-agent models (though there is still only one acting agent). The ideas are based on concepts from Chapter 3, Chapter 4 and [Bolander and Andersen, 2011]. Because of its nature as work in progress, there are a number of unresolved issues which will be discussed throughout the chapter. Chief amongst these is the absence of proofs for various properties. Luckily, many of the proofs will be straightforward generalisations from earlier the earlier work.

5.1 Planning so far

The two types of planning that have been introduced in Chapters 2 and 3[1] have two features in common, both of which will be modified in the following. Firstly, they both use single-agent models. In this chapter I show how to modify those techniques for planning with multi-agent models.

Secondly, both types of planning are 'epistemically safe', in that only those parts of the model that are known to become possible or impossible immediately after an action (i.e. the information cells) may be disregarded by the planner. When planning an action that is expected to sense the value of p, it is safe to discard one

[1] And their respective publications [Andersen et al., 2012] and[Andersen et al., 2014]

of the resulting information cells. If the actual outcome turns out to be $\neg p$, this will be discovered when executing the sensing action. Even when things don't turn out as expected, the resulting model will contain the actual world. So if a weak plan or a plausibility plan of either strength contains an action requiring p and the actual outcome was $\neg p$, the discrepancy will be observed before the action in question is ever executed.

Plausibility planning in particular can be viewed as a strategy that spends computational resources conservatively. If strong planning means paying (a lot) for a plan that is *guaranteed* to work, and weak planning means paying for a plan that *might* work, then strong plausibility planning means paying for a plan that is *guaranteed* to work for all *expected* and epistemically indistinguishable circumstances.[2] If we accept that artificial agents (and humans too) have hardware constraints and need to be expeditious – everybody's gotta act sometime – plausibility planning seems like the better choice. This is, of course, not the case in domains where we cannot recover from unexpected events, but if we wish to model and create agents that can work in real world scenarios we must accept, alas, this plight of people and robots alike. However, retaining *every* possibility until conclusive evidence shows it impossible is still a lot to ask. As long as I don't expect that my bicycle has been stolen, I don't include that possibility in my model. If humans don't do it, we should at least consider not requiring it of our agents.

For a planning method that allows discarding unexpected outcomes that have not yet been proven impossible, I modify plausibility planning to produce *default planning*.[3] In default planning, the agent only retains outcomes that that are believed, instead of retaining everything that is not (yet) known to be impossible.

Discarding *more* than what known to be impossible introduces the need for belief revision. If I discover, contrary to expectation, that my bicycle has in fact been stolen, I need to revise my model in such a way that it becomes compatible with this new information. Here it is done based on beliefs about what has happened in the past. By guessing a plausible explanation for how my bicycle was stolen, I can predict facts that aren't observed directly.

Plausibility planning with multi-agent models as defined in the following sections, is derived from concepts from previous work:

- Multi-agent plausibility models, bisimulation and the normal plausibility relation are adapted from Chapter 4.
- For planning we want $w \sim v \Rightarrow (w \geq v)$ or $(w \leq v)$ as in Chapter 3, but not

[2] And for weak plausibility planning: To work in *some expected* circumstances.

[3] The name is inspired by default logic for its ability to infer what holds *most of the time*.

$w \sim v \Leftrightarrow (w \geq v)$ or $(w \leq v)$ as in Chapter 4.

- In multi-agent planning, states are multi-pointed models and actions are multi-pointed event models as in [Bolander and Andersen, 2011]. Unable to point out a single world as the actual one (except in certain special situations), the planning agent maintains a set of worlds W_d, knowing only that the actual world is in W_d. When default planning, the agent believes that the actual world is in W_d.

5.2 Models & Language

Definition 5.1 (Plausibility model)**.** A *plausibility model* on a set of propositional symbols P and a set of agents A is a tuple $M = (W, \sim, \geq, V)$, where

- W is a finite set of *worlds*. $D(M) = W$ is called the domain.
- $\sim: A \to \mathcal{P}(W \times W)$ gives an indistinguishability relation for each $a \in A$. Each indistinguishability relation is an equivalence relation.
- $\geq: A \to \mathcal{P}(W \times W)$ is a plausibility relation for each $a \in A$, where $\geq(a)$ is a set of mutually disjoint well-preorders covering W, with the requirement that $w \sim_a v$ implies $(w \geq_a v)$ or $(v \geq_a w)$.
- $V : W \to 2^P$ is a *valuation*.

For $W_d \subseteq W$, (M, W_d) is a *multi-pointed plausibility model*, and W_d are the *designated worlds*.

$\sim_a$ is an equivalence relation on W called the *epistemic relation* (for agent a). For $\geq(a)$ we write $\geq_a$. If $w \geq_a v$ then v is *at least as plausible* as w (for agent a). If $w \geq_a v$ but $v \not\geq_a w$ we write $w >_a v$ (v is *more plausible* than w). For $w \geq_a v$ and $v \geq_a w$ we write $w \simeq_a v$ (w and v are *equiplausible*). Instead of $w \geq_a v$ I may write $v \leq_a w$ and $v <_a w$ instead of $w <_a v$. With $G \subseteq A$, I take $[w]_G$ to mean $[w]_{\sim_G}$, where $\sim_G$ is $(\bigcup_{a \in G} \sim_a)^=$, i.e. the equivalence closure of the union of $\sim_a$ for each agent a in G. [Fagin et al., 1995] calls this G-reachability: $[w]_G$ is the set of worlds reachable from w by only going through edges belonging to an agent in G.

Unlike in Chapter 4, but as in Chapter 3, $\sim_a$ is *not* the symmetric closure of $\geq_a$. Because $\geq(a)$ is a set of well-preorders, the symmetric closure of $\geq(a)$ is a set of equivalence relations. As with $[w]_a$ denoting the $\sim_a$-closure of w, I will use $[w]_{(\geq_a \cup \leq_a)}$ to denote the $(\geq_a \cup \leq_a)$-closure of w.

For this work I use an adaptation of the normal plausibility relation $\succeq_a$ defined in Chapter 4: With a plausibility model $M = (W, \sim, \geq, V)$, I write $w \geq^R_a v$ for $Min_a([w]_{R^=} \cap [w]_{(\geq_a \cup \leq_a)}) \geq_a Min_a([v]_{R^=} \cap [v]_{(\geq_a \cup \leq_a)})$ where R is an autobisimulation on M of the kind defined Definition 4.2, but with the addition of [forth$_\sim$]- and [back$_\sim$]-clauses:

Definition 5.2 (Autobisimulation). Let $M = (W, \sim, \geq, V)$ be a plausibility model. An *autobisimulation* on M is a non-empty relation $R \subseteq W \times W$ such that for all $(w, w') \in R$ and for all $a \in A$:

[atoms] $V(w) = V(w')$;

[forth$_\geq$] If $v \in W$ and $w \geq^R_a v$, there is a $v' \in W$ such that $w' \geq^R_a v'$ and $(v, v') \in R$;

[back$_\geq$] If $v' \in W$ and $w' \geq^R_a v'$, there is a $v \in W$ such that $w \geq^R_a v$ and $(v, v') \in R$;

[forth$_\leq$] If $v \in W$ and $w \leq^R_a v$, there is a $v' \in W$ such that $w' \leq^R_a v'$ and $(v, v') \in R$;

[back$_\leq$] If $v' \in W$ and $w' \leq^R_a v'$, there is a $v \in W$ such that $w \leq^R_a v$ and $(v, v') \in R$;

[forth$_\sim$] If $v \in W$ and $w \sim_a v$, there is a $v' \in W$ such that $w' \sim_a v'$ and $(v, v') \in R$;

[back$_\sim$] If $v' \in W$ and $w' \sim_a v'$, there is a $v \in W$ such that $w \sim_a v$ and $(v, v') \in R$

Now writing $w \succeq_a v$ for $w \geq^R_a v$, where R is the maximal autobisimulation, I proceed to the definition of minimal elements for $\succeq_a$.[4] With a plausibility model $M = (W, \sim, \geq, V)$, $w \in W$ and $a \in A$, any $S \subseteq [w]_{(\geq_a \cup \leq_a)}$ has minimal worlds with respect to $\succeq_a$. These are denoted $Min_a S$ and defined as

$$Min_a S = \{s \in S \mid \forall s' \in S : s' \succeq_a s\}.$$

This gives the following definition of the worlds believed at degree n:

$$\begin{aligned} Min^0_a S &= Min_a S \\ Min^{n+1}_a S &= \begin{cases} S & \text{if } Min^n_a S = S \\ Min^n_a S \cup Min_a(S \setminus Min^n_a S) & \text{otherwise} \end{cases} \end{aligned}$$

Note to the reader*: This isn't complete before proving that there indeed is a maximal autobismulation and that $\geq^R_a$ is a well-preorder. As the new bisimulation clauses are completely standard [Blackburn et al., 2001, pp. 64-65], the proofs are going to be very similar to those where bisimulations are as in Definition 4.2.*

Having abandoned $w \sim_a v \Leftarrow (w \geq v)$ or $(w \leq v)$ I should note that $[w]_{(\geq_a \cup \leq_a)}$ may contain more worlds than $[w]_a$, so it can be the case that $Min_a[w]_a \not\subseteq$

[4] Only interested in B^n_a, I do not need minimal elements for $\geq_a$.

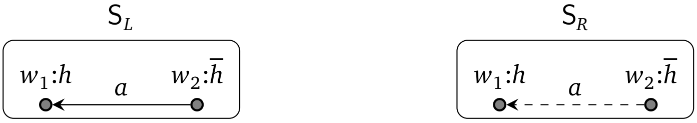

Figure 5.1: Two multi-pointed plausibility models, the left one with future indistinguishability, the right one with future distinguishability. If a world is grey it is in W_d. An a-edge from w to v indicates that $w \geq_a v$. A solid edge indicates $w \sim_a v$, whereas a dashed edge indicates $w \not\sim_a v$. A proposition written with a line over it ($\overline{h}$) indicates that it is false. Reflexive edges are omitted. Transitive edges may be omitted. This is the multi-agent version of the biased coin toss of Example 3.2 and 3.3.

$Min_a[w]_{(\geq_a \cup \leq_a)}$. This is in line with Chapter 3, where we had $[w]_{(\geq \cup \leq)} = W$, despite not (necessarily) having $[w]_\sim = W$.

This really is a threshold concept, so despite similarities with the definitions in Chapter 3, let's have a look at how it works for the new models.

Example 5.3. Figure 5.1 contains two plausibility models, both with world w_1 more plausible than w_2. In the left model they are epistemically indistinguishable, whereas the right have them distinguishable. We do not indicate the actual world, as is customary in modal logic. Because we are creating a framework that is to run on an agent situated in the environment being modelled, our models must be faithful to the point of view of the agent to whom $\sim_a$ and $\geq_a$ belong. The agent in the system does not know which world is the actual one. Indicating it with a single-pointed model would betray this intention, because *the modeller is modelling itself.*

Referring back to Examples 3.2 and 3.3 on page 51, the models describes reasoning about future outcomes of peeking at a hidden, biased coin. We can see the left model of Figure 5.1 as the reasoner's hypothesis about the situation the hidden coin toss. The right model is the agent's hypothesis about the possible situations after peeking at the coin. In this sense, $w_1 \not\sim_a w_2$ encodes *future* distinguishability. While it does not yet know the value of h, it knows that it will come to know it. Further, it believes (i.e. expects) that h will turn out to be true, in which case it will know it. The agent expects to come to know h. ■

What the agent knows now, and what it only knows it will come to know corresponds to the notions identified in [Petrick and Bacchus, 2002, Petrick and Bacchus, 2004][5] as *run time* and *plan time* knowledge. Plan time knowledge about h means

[5] Detailing the implementation of the PKS planner mentioned many pages ago.

that h or $\neg h$ is known to be true while planning, whereas run time knowledge about h means that either h or $\neg h$ will become known.

Definition 5.4 (Event model)**.** An *event model* on the language $L(P,A)$ (given in Definition 5.14) is a tuple $\mathcal{E} = (E, \sim, \geq, pre, post)$, where

- E is a finite set of *events*;
- $\sim: A \rightarrow \mathcal{P}(E \times E)$ gives an indistinguishability relation for each $a \in A$. Each indistinguishability relation is an equivalence relation.
- $\geq: A \rightarrow \mathcal{P}(E \times E)$ is a plausibility function, such that for each $a \in A$, $\geq(a)$ is a set of mutually disjoint well-preorders covering E, with the requirement that $e \sim_a f \Rightarrow (e \geq_a f)$ or $(e \geq_a f)$.
- $pre : E \rightarrow L(P,A)$ assigns to each event a *precondition*.
- $post : E \rightarrow (P \rightarrow L(P,A))$ assigns to each event a *postcondition*.

For $E_d \subseteq E$, $(\mathcal{E}, E_d)$ is a *multi-pointed event model*, and E_d are the *designated events*.

For planning and reasoning in the multi-agent case, the following defines a notion of states and actions for a particular agent.

Definition 5.5 (States and Actions)**.** With $M = (W, \sim, \geq, V)$ being a plausibility model on P and A and $W_d \subseteq W$, we call the pair $\mathsf{S} = (M, W_d)$ a *state* for agent a when the following conditions hold:

1. For all $i \in A$ and all $w, w' \in W$, $w \sim_i w'$ iff $(w \geq_i w'$ or $w \leq_i w')$.
2. For all $w \in W_d : [w]_a = W_d$.

S is called a *prestate* for agent a if W_d is closed under $(\geq_a \cup \leq_a)$. Thus, all states are prestates, but not all prestates are states. If $\mathsf{S} = (M, W_d)$ is a state, then M corresponds to a multi-agent plausibility model as defined in Chapter 4. When clear from the context, I will use M, M' (respectively W_d, W'_d), and similar, in reference to the model (respectively designated worlds) of $\mathsf{S} = (M, W_d)$, $\mathsf{S}' = (M', W'_d)$.

I call $\mathsf{A} = (\mathcal{E}, E_d)$, where $\mathcal{E} = (E, \sim, \geq, pre, post)$ and $E_d \subseteq E$, an action for i if E_d is closed under $(\geq_i \cup \leq_i)$.

The intuition for these definitions is that states represent the situation after all agents have found out which of the hypotheses encoded in a prestate was the true

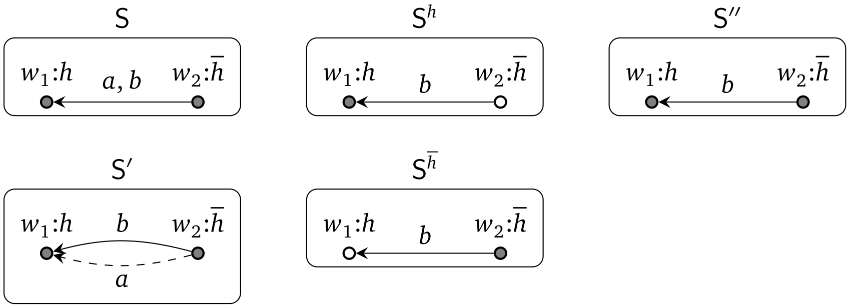

Figure 5.2: The state S represents both a and b being uncertain about, but believing, h. The prestate S' represents all outcomes of a privately learning the value of h, i.e peeking under the dice cup. Also shown are the induced states for both a and b: S^h and $\mathsf{S}^{\overline{h}}$ are a's possible states after peeking. S'' is the (only) possible state for b after the same. b hasn't learned the value of h, but knows that a has.

situation. This is why condition 1. requires $w \sim_j w'$ iff ($w \geq_j w'$ or $w \leq_j w'$) for all $j \in A$, instead of just for i. In this interpretation, there can be prestates which are also states. When that is the case, everything known at run time was already known at plan time. Additionally, if S is a state for the agent i, it can also be called i's *perspective*.

Example 5.6. Consider the models in Figure 5.2, where grey nodes indicate that a world is in W_d. Definition 5.5 gives that S (where $W_d = \{w_1, w_2\}$) is a state for both a (and b). Let's check: As $\sim$ and $\geq$ are reflexive, we need only check relations between worlds. As $w_1 \sim_a w_2$ and $w_1 \leq_a w_2$ (same for b), the first condition holds. The second condition also holds, as we have $[w_1]_a = [w_2]_a = W_d$; S is a state for both a and b.

S' is a prestate for both agents, as the prestate condition that W_d be closed under $(\geq_i \cup \leq_i)$ holds for $i = a$ and $i = b$ both. It is not a state, as condition 1. does not hold for a. This means that it cannot be a model of a actually having learnt the value of h. If a had sensed h or $\neg h$, we wouldn't have $w_2 \geq_a w_1$. Learning, for instance, the actual world is w_1, a would consider w_2 *impossible*, not just less plausible. ■

From a prestate, we can generate the possible *actual* states for a particular agent, by way of the actualisation operator $\mathcal{K}_i$. This operator gives all the possible states (or run time models) that a given prestate encodes.

Definition 5.7 ($\mathcal{K}_i$)**.** Let $\mathsf{S} = (M, W_d)$ be a state or prestate, where $(W, \sim, \geq, V)$ is a plausibility model on P,A and $W_d \subseteq W$. The actualisation of S for $a \in A$ is

$\mathcal{K}_a(\mathsf{S}) = \{((W, \sim, \geq \cap \sim, V) \restriction [w]_A, [w]_a) \mid w \in W_d\}$. Note that for any state $(M', W'_d) \in \mathcal{K}_i(M, W_d)$, $W'_d \subseteq W_d$.

This definition might be a bit hard to parse, so let's take some time to go through a detailed calculation of it.

Example 5.8. Continuing with the states of Figure 5.2, the actual state for a after sensing h is going to be one of those in $\mathcal{K}_a(\mathsf{S}')$. First note that

$$(W, \sim, \geq \cap \sim, V) \restriction [w_1]_A = (W, \sim, \geq \cap \sim, V) \restriction [w_2]_A$$

In S', we have $\sim_a = \{(w_1, w_1), (w_2, w_2)\}$ and $\geq_a = \{(w_1, w_1), (w_2, w_2), (w_1, w_2)\}$ so $\geq'_a = \geq_a \cap \sim_a = \{(w_1, w_1), (w_2, w_2)\}$. For b we have $\geq'_b = \geq_b \cap \sim_b = \geq_b$. As $[w_1]_A = [w_2]_A$, the only difference between the states in $\mathcal{K}_a(\mathsf{S}')$ is going to be the designated worlds. With $[w_1]_a = \{w_1\}$ and $[w_2]_a = \{w_2\}$ being two disjoint sets, we get that $\mathcal{K}_a(\mathsf{S}')$ will contain two states, i.e. $\mathcal{K}_a(\mathsf{S}') = \{\mathsf{S}^h, \mathsf{S}^{\overline{h}}\}$.

For $\mathcal{K}_b(\mathsf{S}')$, the only difference is that the designated worlds is the equivalence class $[w_1]_b = [w_2]_b = \{w_1, w_2\}$. Thus $\mathcal{K}_b(\mathsf{S}') = \{\mathsf{S}''\}$.

We have that S'' is not a state for a, as $[w_1]_a \neq \{w_1, w_2\}$. Similarly, neither S^h nor $\mathsf{S}^{\overline{h}}$ is a state for b, as $[w_1]_b \neq \{w_1, w_2\} \neq [w_2]_b$.

Generally we have that if S is a state for i, then $\mathcal{K}_i(\mathsf{S}) = \{\mathsf{S}\}$. In this particular example we have $\mathcal{K}_a(\mathsf{S}) = \mathcal{K}_b(\mathsf{S}) = \{\mathsf{S}\}$, $\mathcal{K}_a(\mathsf{S}^h) = \{\mathsf{S}^h\}$, $\mathcal{K}_a(\mathsf{S}^{\overline{h}}) = \{\mathsf{S}^{\overline{h}}\}$ and $\mathcal{K}_b(\mathsf{S}'') = \{\mathsf{S}''\}$. ■

Analogous to *most plausible information cell* for plausibility planning with single-agent models (Chapter 3 p. 56), I call a state $\mathsf{S}' \in \mathcal{K}_i(\mathsf{S})$ an *n-most plausible state* for a if $W'_d \cap Min^n_a W_d \neq \emptyset$. The n-most plausible states for a of a (pre)state S will be denoted $Min^n_a(\mathsf{S})$. Where convenient, I will use $Min_a(\mathsf{S})$ for $Min^0_a(\mathsf{S})$ and call $\mathsf{S}' \in Min_a(\mathsf{S})$ a most plausible state. Because $Min^n_a W_d$ is non-empty (remember that the weakest requirement for W_d is that it is closed under $(\geq_a \cup \leq_a)$), there will always be at least one n-most plausible state.

Beware that the n does *not* refer to a degree defined with respect to an ordering on states, but with respect to $\succeq_a$ on designated worlds in the prestate that induces the state.

Example 5.9. Figure 5.3 shows a prestate S with $\mathcal{K}_a(\mathsf{S}) = \{\mathsf{S}^{0-1}, \mathsf{S}^{2-3}\}$. The n-most plausible states of S are $Min^0_a(\mathsf{S}) = Min^1_a(\mathsf{S}) = \{\mathsf{S}^{0-1}\}$ and $Min^2_a(\mathsf{S}) = Min^3_a(\mathsf{S}) = \{\mathsf{S}^{0-1}, \mathsf{S}^{2-3}\} = Min^k_a(\mathsf{S})$ for $k \geq 3$. ■

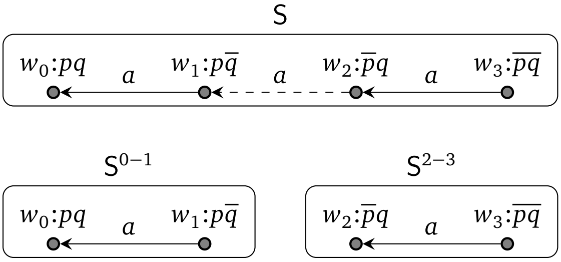

Figure 5.3: A prestate state S and the induced states for a.

Finally, note that an agent i can take the perspective of another agent j using $\mathcal{K}_j(\mathsf{S})$, where S is a (pre)state for i.

Example 5.10. Consider again the states in Figure 5.2. For both a-states S^h and $\mathsf{S}^{\overline{h}}$ we have $\mathcal{K}_b(\mathsf{S}^h) = \mathcal{K}_b(\mathsf{S}^{\overline{h}}) = \{\mathsf{S}''\}$. Regardless of whether a sees (or has seen) h or $\neg h$, a knows that S'' is b perspective.

For the b-state S'', we have $\mathcal{K}_a(\mathsf{S}'') = \{\mathsf{S}^h, \mathsf{S}^{\overline{h}}\}$. Though b does not know whether a's perspective is S^h or $\mathsf{S}^{\overline{h}}$, b believes that a's perspective is S^h as $Min^0_b(\mathcal{K}_a(\mathsf{S}'')) = \{\mathsf{S}^h\}$. ■

Definition 5.11 (Product Update). Let $\mathsf{S} = ((W, \sim, \geq, V), W_d)$ and $\mathsf{A} = ((E, \sim', \geq', pre, post), E_d)$ be a multi-pointed plausibility model on P respectively multi-pointed event model on $L(P,A)$. The *product update* of S with A is the multi-pointed plausibility model $(M, W_d) \otimes (\mathcal{E}, E_d) = ((W', \sim'', \geq'', V'), W'_d)$, where

- $W' = \{(w,e) \in W \times E \mid M, w \models pre(e)\}$,
- $\sim''_i = \{((w,e),(v,f)) \in W' \times W' \mid w \sim_i v \text{ and } e \sim'_i f\}$,
- $\geq''_i = \{((w,e),(v,f)) \in W' \times W' \mid e >'_i f \text{ or } (e \simeq'_i f \text{ and } w \geq_i v)\}$,
- $V'(p) = \{(w,e) \in W' \mid M, w \models post(e)(p)\}$ for each $p \in P$.
- $W'_d = \{(w,e) \in W' \mid w \in W_d \text{ and } e \in E_d\}$.

$\mathcal{K}_a(\mathsf{S}\otimes\mathsf{A})$ are the *possible* states for a after A happens at S. Similarly, $Min_a\mathcal{K}_a(\mathsf{S}\otimes\mathsf{A})$ are the *expected* states for a after A happens at S. For $\mathsf{S}' \in \mathcal{K}_a(\mathsf{S} \otimes \mathsf{A})$, respectively $\mathsf{S}' \in Min_a\mathcal{K}_a(\mathsf{S} \otimes \mathsf{A})$, S' is a possible, respectively expected state after A.

Proposition 5.12. *Let S be a state or prestate, and A action for agent a. Then $\mathsf{S} \otimes \mathsf{A}$ is a prestate for a. If E_d is closed under $\sim_a$ and S is a state for a, then $\mathsf{S} \otimes \mathsf{A}$ is also a state for a. Finally, all states in $\mathcal{K}_a(\mathsf{S} \otimes \mathsf{A})$ are states for a.*

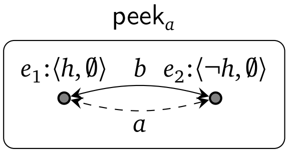

Figure 5.4: a peeks at the coin. b sees that a peeks, but does not know what a sees: peek$_a$is a private sensing of h by a.

Example 5.13. For the action peek$_a$shown in Figure 5.4 and the states of Figure 5.2, we have that $\mathsf{S} \otimes \text{peek}_a = \mathsf{S}'$, and that the possible states for a after peek$_a$ are $\{\mathsf{S}^h, \mathsf{S}^{\bar{h}}\}$, whereas S^h is the (only) expected state. ■

Definition 5.14 (Dynamic Language)**.** Let a countable set of propositional symbols P and agents A be given. The language $L(P,A)$ is given by the following BNF:

$$\varphi ::= p \mid \neg\varphi \mid \varphi \wedge \varphi \mid K_a\varphi \mid B_a^n\varphi \mid [\varphi]\varphi \mid [\mathcal{E}, e]\,\varphi \mid \mathcal{K}_a\varphi$$

where $p \in P$, $a \in A$, $\mathcal{E}$ is an *event model* on $L(P,A)$, and $e \in D(\mathcal{E})$. K_a is the *run time* knowledge modality, B_a^n the *plan time* degree-of-belief modality, $\mathcal{K}_a$ is the actualisation modality and $[\mathcal{E}, e]$ the dynamic modality.

I use the usual abbreviations for the other boolean connectives, as well as for the dual dynamic modality $\langle \mathcal{E}, e\rangle\,\varphi := \neg\,[\mathcal{E}, e]\,\neg\varphi$ and degree 0 plan time belief $B\varphi := B^0\varphi$. The duals of K_a and B_a^n are $\widehat{K}_a$ and $\widehat{B}_a^n$.

Definition 5.15 (Satisfaction Relation)**.** Let a plausibility model $M = (W, \sim, \geq, V)$ on P,A be given. The satisfaction relation is given by, for all $w \in W$:

$$\begin{array}{ll}
M, w \models p & \text{iff } w \in V(p) \\
M, w \models \neg\varphi & \text{iff } \textit{not } M, w \models \varphi \\
M, w \models \varphi \wedge \psi & \text{iff } M, w \models \varphi \text{ and } M, w \models \psi \\
M, w \models K_a\varphi & \text{iff } M, v \models \varphi \text{ for all } v \in [w]_a \\
M, w \models B_a^n\varphi & \text{iff } M, v \models \varphi \text{ for all } v \in Min_a^n[w]_{(\geq_a \cup \leq_a)} \\
M, w \models [\varphi]\,\psi & \text{iff } M, w \models \varphi \text{ implies } M \mid \varphi, w \models \psi \\
M, w \models [\mathcal{E}, e]\,\varphi & \text{iff } M, w \models pre(e) \text{ implies } M \otimes \mathcal{E}, (w, e) \models \varphi \\
M, w \models \mathcal{K}_a\varphi & \text{iff } M', w \models \varphi \text{ where } \{(M', W_d')\} = \mathcal{K}_a(M \upharpoonright [w]_A, [w]_a)
\end{array}$$

where $\varphi, \psi \in L(P,A)$ and $(\mathcal{E}, e)$ is a pointed event model. We write $M, W_d \models \varphi$ to mean $M, w \models \varphi$ for all $w \in W_d$, while $M \models \varphi$ means $M, W \models \varphi$. $M \mid \varphi$ is the restriction of M to the worlds satisfying φ, i.e. $M \restriction \{w \in W \mid M,w \models \varphi\}$. Satisfaction for the dynamic modality for multi-pointed event models $(\mathcal{E}, E_d)$ is introduced by abbreviation, viz. $[\mathcal{E}, E_d]\,\varphi := \bigwedge_{e \in E_d} [\mathcal{E}, e]\,\varphi$. Furthermore, $\langle \mathcal{E}, E_d \rangle\,\varphi := \neg\,[\mathcal{E}, E_d]\,\neg\varphi$.[6]

Note the overloading of $\mathcal{K}_a$ in $M,w \models \mathcal{K}_a\varphi$, and that $\mathcal{K}_a(M \restriction [w]_G, [w]_a)$ is (always going to be) the singleton set containing only the state $\{M', W'_d\}$. Also note that, now a language has been defined, plan time knowledge of φ can given by $\mathsf{S} \models \varphi$ (i.e. $M, W_d \models \varphi$), and run time knowledge by $\mathsf{S} \models K_a\varphi$. If we turn back to Chapter 3, we see that $\mathcal{K}_a$ is the multi-agent version of the localisation modality X from Definition 3.9 on page 56, where $M,w \models X\varphi$ iff $M \restriction [w]_\sim, w \models \varphi$.

As a sanity sanity check of the definitions so far, consider the following lemma as an analogue to Lemma 3.13 on page 59.

Lemma 5.16. *If* S *and* A *are a state and an action for agent a, then* $\mathsf{S} \models [\mathsf{A}]\varphi$ *iff* $\mathsf{S} \otimes \mathsf{A} \models \varphi$

Proof. Simply replace $D(M)$ with W_d and $D(\mathcal{E})$ with E_d in the proof of Lemma 3.13:

$$\begin{aligned}
&\mathsf{S} \models [\mathsf{A}]\varphi \Leftrightarrow \forall w \in W_d : M,w \models [\mathsf{A}]\varphi \Leftrightarrow\\
&\forall w \in W_d : \mathsf{S},w \models \bigwedge_{e \in E_d} [\mathsf{A},e]\varphi \Leftrightarrow\\
&\forall (w,e) \in W_d \times E_d : \mathsf{S},w \models [\mathsf{A},e]\varphi \Leftrightarrow\\
&\forall (w,e) \in W_d \times E_d : \mathsf{S},w \models pre(e) \text{ implies } \mathsf{S} \otimes \mathsf{A},(w,e) \models \varphi \Leftrightarrow\\
&\forall (w,e) \in \{(W_d \times E_d) | M,w \models pre(e)\} : \mathsf{S} \otimes \mathsf{A},(w,e) \models \varphi \Leftrightarrow\\
&\mathsf{S} \otimes \mathsf{A} \models \varphi.
\end{aligned}$$

□

And as a further indication that these definitions are as they should be, we should check how the dynamic modality relates to the possible states after doing an action. I exclude the B^n_a modality because it uses $\geq_a$. That is, B^n_a is plan time belief. Alternatively, we can require that B^n_a only occurs immediately after $\mathcal{K}_a$.

Lemma 5.17. *If* S *and* A *are a state and an action for agent a, and* φ *is a* B^n_a*-free formula, then* $\mathsf{S} \models [\mathsf{A}]\varphi$ *iff* $\forall \mathsf{S}' \in \mathcal{K}_a(\mathsf{S} \otimes \mathsf{A}) : \mathsf{S}' \models \varphi$

[6] Hence, $M,w \models \langle \mathcal{E}, E_d \rangle\,\varphi \Leftrightarrow M,w \models \neg\,[\mathcal{E}, E_d]\,\neg\varphi \Leftrightarrow M,w \models \neg(\bigwedge_{e \in E_d} [\mathcal{E}, e]\,\neg\varphi) \Leftrightarrow M,w \models \bigvee_{e \in E_d} \neg\,[\mathcal{E}, e]\,\neg\varphi \Leftrightarrow M,w \models \bigvee_{e \in E_d} \langle \mathcal{E}, e \rangle\,\varphi$.

Proof.

$$
\begin{aligned}
&\mathsf{S} \models [\mathsf{A}]\varphi \Leftrightarrow \mathsf{S} \otimes \mathsf{A} \models \varphi \Leftrightarrow^{\varphi \text{ is } B_a^n\text{-free.}}\\
&(M, W_d) \models \mathcal{K}_a\varphi \Leftrightarrow \forall w \in W_d : M, w \models \mathcal{K}_a\varphi \Leftrightarrow\\
&\forall w \in W_d : M', w \models \varphi \text{ where } \{(M', W_d')\} = \mathcal{K}_a(M \restriction [w]_a, [w]_a) \Leftrightarrow\\
&\forall w \in W_d, \forall w' \in W_d' : M', w' \models \varphi \text{ where } \{(M', W_d')\} = \mathcal{K}_a(M \restriction [w]_a, [w]_a) \Leftrightarrow\\
&\forall w \in W_d : M', W_d' \models \varphi \text{ where } \{(M', W_d')\} = \mathcal{K}_a(M \restriction [w]_a, [w]_a) \Leftrightarrow\\
&\forall w \in W_d : \mathsf{S}' \models \varphi \text{ where } \{\mathsf{S}'\} = \mathcal{K}_a(M \restriction [w]_a, [w]_a) \Leftrightarrow\\
&\forall \mathsf{S}' \in \mathcal{K}_a(\mathsf{S}) : \mathsf{S}' \models \varphi
\end{aligned}
$$

□

5.2.1 Buying Records – A Running Single-Agent Example

The following example, with amendments, will be the focus of discussion and motivation of the new techniques for planning. While the example in principle uses multi-agent models, only a single agent is present and acting (if we don't consider the shop vendor and pickpocket agents). To reduce visual clutter, I will take the liberty to drop the agent indexes on modalities and edges. Note that all worlds and events are going to be designated, so there is no need to point them out individually. Let's begin:

Our protagonist is perusing the vinyl collections at the Saturday Records Fair in Old Spitalfields Market. Having found a nice selection of classics, she approaches the vendor and asks if she can pay with her debit card. Eager to accommodate a new customer, he tells her that he does have a debit card reader, but that it's been a bit unreliable lately. Alternatively, there's a cashpoint at the other end of the market. Reluctant to make her way through the crowds, she plots her course of action.

The initial situation and our agent's actions can be seen in Figure 5.5. In the initial state S_0 she knows that she hasn't paid for the records and that she doesn't have any cash on hand ($\neg p \wedge \neg c$ is true in all possible worlds). She believes that there is a positive balance on the account associated with the debit card (a) and, perhaps a bit optimistically, that the card reader works (w), i.e. she believes ($a \wedge b$). Finally, she considers a-worlds more plausible than w-worlds (as $w_3 >_a w_2 >_a w_1$).

The actions Card, Cash and ATM represent trying to pay with the debit card, paying with cash, and trying to withdraw money at the cashpoint. The "trying to" part is important. I will return to this later.

For Card we see that if there's money on the debit card account (a) and the reader

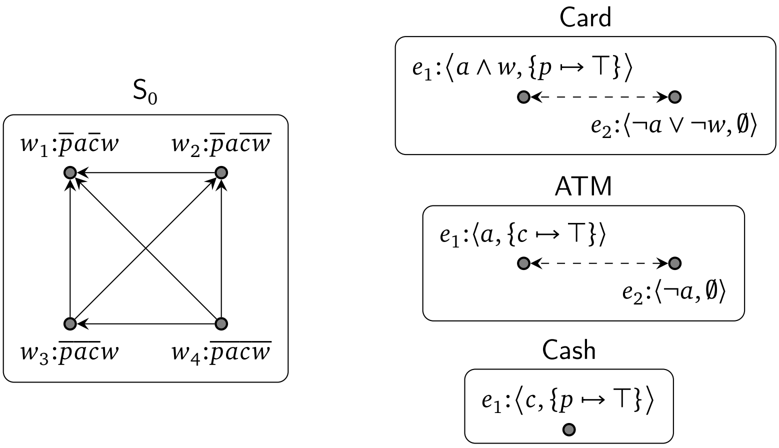
S_0
$w_1{:}\bar{p}a\bar{c}w$
$w_2{:}\bar{p}a\overline{cw}$
$w_3{:}\overline{pac}w$
$w_4{:}\overline{pacw}$
Card
$e_1{:}\langle a \wedge w, \{p \mapsto \top\}\rangle$
$e_2{:}\langle \neg a \vee \neg w, \emptyset\rangle$
ATM
$e_1{:}\langle a, \{c \mapsto \top\}\rangle$
$e_2{:}\langle \neg a, \emptyset\rangle$
Cash
$e_1{:}\langle c, \{p \mapsto \top\}\rangle$

Figure 5.5: The initial situation plus event models for the three available actions. Events are labelled $\langle pre(e), post(e)\rangle$, where $post(e)$ on the form $\{p_1 \mapsto \varphi_1, \ldots, p_n \mapsto \varphi_n\}$ means that $post(e)(p_i) = \varphi_i$ for all i and $post(e)(q) = q$ for $q \notin \{p_1, \ldots, p_n\}$

works (w), then she will successfully pay ($p \mapsto \top$). She can also pay by Cash provided she has some (c). This she can get by ATM, if the account has a positive balance (a again). All events in these actions are equiplausible and distinguishable.

Figure 5.6 illustrates the product update of the initial model S_0 with Card. This is the model $S_0 \otimes$ Card. The updated model is the agent's hypotheses about the possible outcomes of doing Card. The name of a world in $S_0 \otimes$Card indicates which world in S_0 it came from. As no event in any of the actions has its precondition satisfied in more than one world (no world w will be produce worlds w, e in the updated model), I will stick to this scheme unless otherwise stated.

In the prestate $S_0 \otimes$Card we have $w_1 \not\sim_a w_2 \sim_a w_3 \sim_a w_4$. This is consistent with the requirements on $\geq$ and $\sim$: Everything connected by $\sim_a$ must be $\geq_a$-comparable, whereas worlds that are $\geq_a$-comparable need not be $\sim_a$-connected. Knowledge prior to doing Card is that which holds in all worlds of the updated model, where we have $S_0 \otimes$ Card $\not\models p$ and $S_0 \otimes$ Card $\not\models \neg p$ – *imagining* using the debit card does not determine whether the records will actually be paid for. We do however have $S_0 \otimes$ Card $\models Bp$ – the agent *believes* that using the debit card will complete the transaction. We can get posterior knowledge either by the K or $\mathcal{K}$ on $S_0 \otimes$ Card. We have $S_0 \otimes$ Card $\models Kp \vee K\neg p$ (alternatively $S_0 \otimes$ Card $\models \mathcal{K}p \vee \mathcal{K}\neg p$) – actually using the debit card, as opposed to just hypothesising, will determine whether the records will be paid for. If the goal is to achieve p, then it seems reasonable for our protagonist's first move to be Card. She believes that this will achieve her goal.

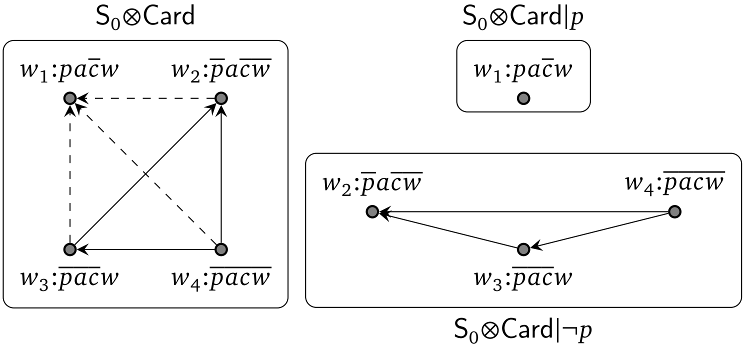
$\mathsf{S}_0 \otimes \mathsf{Card}$
$w_1{:}pa\overline{c}w$
$w_2{:}\overline{p}a\overline{cw}$
$w_3{:}\overline{pac}w$
$w_4{:}\overline{pacw}$
$\mathsf{S}_0 \otimes \mathsf{Card} | p$
$w_1{:}pa\overline{c}w$
$w_2{:}\overline{p}a\overline{cw}$
$w_4{:}\overline{pacw}$
$w_3{:}\overline{pac}w$
$\mathsf{S}_0 \otimes \mathsf{Card} | \neg p$

Figure 5.6: The results of attempting to pay with the debit card, before and after receiving feedback from the environment.

A peculiar feature is that $\mathsf{S}_0 \otimes \mathsf{Card} \models K\neg p \rightarrow Bp$ – if she knows $\neg p$, she believes p? This seems nonsensical, but remember that K represents run time knowledge, whereas B represents plan time belief. Using B and K together has its purposes, but we must take care not to mix present and future tense. B is about the now, K about the future. Using $\mathcal{K}$ we have that $\mathsf{S}_0 \otimes \mathsf{Card} \not\models \mathcal{K}(K\neg p \rightarrow Bp)$, because $\mathcal{K}$ simulates what the state will be when *actually* doing Card, thus pruning worlds distinguishable in the from the plausibility relation. That $\mathsf{S}_0 \otimes \mathsf{Card} \models \mathcal{K}(K\neg p \rightarrow B\neg p)$ should put our minds at ease again.

Our agent decides on, and does, Card. What will the environment's response be? That is not up to the agent, but the possible responses in the form of percepts can be seen as a set of formulae distinguishing the states in $\mathcal{K}(\mathsf{S}_0 \otimes \mathsf{Card})$. Here the possible responses are (for instance) p and $\neg p$. It may be useful to think of the percepts associated with the most plausible state(s) of $\mathsf{S}_0 \otimes \mathsf{Card}$ as the expected response. We can then say that while both p and $\neg p$ are considered possible responses to Card, p is the expected response. If indeed things go as expected, then doing Card will result in the model $\mathsf{S}_0 \otimes \mathsf{Card} \mid p$, and the agent will have achieved her goal.

Suppose instead that the unexpected happens, and that the transaction does not go through. Let $\mathsf{S}_1 = \mathsf{S}_0 \otimes \mathsf{Card} \mid \neg p$ be the corresponding model. Then the agent knows that either the machine isn't working ($\overline{w}$), that there is no money on the account ($\overline{a}$) or both ($\overline{aw}$). Moreover, she believes that the machine *is not* working, and that there *is* cash on the card – $\mathsf{S}_1 \models K(\neg w \vee \neg a) \wedge B(a \wedge \neg w)$.

What should our protagonist do now? It does not require much thought to see that she should withdraw some money at the ATM and then pay with Cash. Figure 5.7 shows the update $\mathsf{S}_1 \otimes \mathsf{ATM}$ and actualisations under the percepts c and $\neg c$. The expected outcome is $\mathsf{S}_2 = \mathsf{S}_1 \otimes \mathsf{ATM} \mid c$. She can then use Cash to pay for

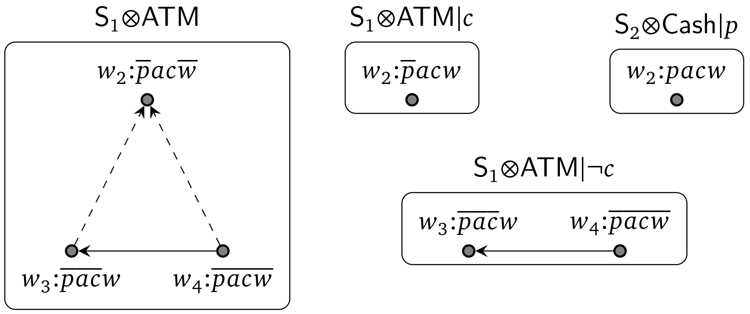
$\mathsf{S}_1 \otimes \mathsf{ATM}$
$w_2{:}\overline{p}ac\overline{w}$
$w_3{:}\overline{pac}w$
$w_4{:}\overline{pacw}$
$\mathsf{S}_1 \otimes \mathsf{ATM}|c$
$w_2{:}\overline{p}acw$
$\mathsf{S}_2 \otimes \mathsf{Cash}|p$
$w_2{:}pacw$
$\mathsf{S}_1 \otimes \mathsf{ATM}|\neg c$
$w_3{:}\overline{pac}w$
$w_4{:}\overline{pacw}$

Figure 5.7: Paying with the debit card has failed, so the agent tries to withdraw cash. We use $\mathsf{S}_1 = \mathsf{S}_0 \otimes \mathsf{Card}|\neg p$ and $\mathsf{S}_2 = \mathsf{S}_1 \otimes \mathsf{ATM}|c$.

the records. The corresponding model is $\mathsf{S}_2 \otimes \mathsf{Cash} \mid p$, where the goal of making p true has been achieved. We even have $\mathsf{S}_2 \otimes \mathsf{Cash} \models p$, so, if ATM succeeds in withdrawing cash, she knows at plan time that Cash will achieve her goal.

5.3 Plausibility Planning with Multi-agent Models

In Chapter 3, we showed how to plan with single-agent plausibility models. Here I show how to do plausibility planning with multi-agent models. The definitions that follow are the multi-agent generalisations of the corresponding concepts from Chapter 3, Section 3.3.3 and 3.4, starting on page 61.

Definition 5.18 (Applicability). An action $\mathsf{A} = (\mathcal{E}, E_d)$ is said to be *applicable* in a state $\mathsf{S} = (M, W_d)$ if $M, W_d \models \langle \mathcal{E}, E_d \rangle \top$. If S is a state and A an action, both for agent a, we say that A is applicable in S for a.

Unfolding the definition of $\langle \mathcal{E}, E_d \rangle$, we see that applicability still works for multi-agent models:

$$
\begin{aligned}
&M, W_d \models \langle \mathcal{E}, E_d \rangle \top \Leftrightarrow \forall w \in W_d : M, w \models \langle \mathcal{E}, E_d \rangle \top \Leftrightarrow \\
&\forall w \in W_d : M, w \models \vee_{e \in E_d} \langle \mathcal{E}, e \rangle \top \Leftrightarrow \\
&\forall w \in W_d, \exists e \in E_d : M, w \models \langle \mathcal{E}, e \rangle \top \Leftrightarrow \\
&\forall w \in W_d, \exists e \in E_d : M, w \models pre(e) \text{ and } M \otimes \mathcal{E}, (w, e) \models \top \Leftrightarrow \\
&\forall w \in W_d, \exists e \in E_d : M, w \models pre(e).
\end{aligned}
$$

Now it says that the action A is applicable in S if all worlds considered possible by the agent (it is one of those in W_d) have at least one matching event among those the agent considers possible in the action. This concept of applicability is equivalent to the one in [Bolander and Andersen, 2011].

Definition 5.19 (Plan Language)**.** Given a finite set AL of actions on $L(P,A)$, the *plan language* $L(P,A,AL)$ is given by:

$$\pi ::= \mathsf{A} \mid \mathsf{skip} \mid \mathsf{if}\ \varphi\ \mathsf{then}\ \pi\ \mathsf{else}\ \pi \mid \pi;\pi$$

where $\mathsf{A} \in AL$ and $\varphi \in L(P,A)$. We name members π of this language *plans*, and use $\mathsf{if}\ \varphi\ \mathsf{then}\ \pi$ as shorthand for $\mathsf{if}\ \varphi\ \mathsf{then}\ \pi\ \mathsf{else}\ \mathsf{skip}$.

The reading of the plan constructs are "do A", "do nothing", "if φ then π, else π'", and "first π then π'" respectively. In the translations provided in 5.20, the condition of the if-then-else construct becomes a K-formula, ensuring that branching depends only on worlds which are distinguishable to the agent.

Definition 5.20 (Translation)**.** Let α be one of s, w, sp or wp. The *α-translation* is a function from $L(P,A,AL) \times L(P,A)$ into $L(P,A)$:

$$[\mathsf{A}]_\alpha \varphi := \langle \mathsf{A} \rangle \top \wedge \begin{cases} [\mathsf{A}]\, \mathcal{K}_a K_a \varphi & \text{if } \alpha = s \\ \widehat{K}_a \langle \mathsf{A} \rangle\, \mathcal{K}_a K_a \varphi & \text{if } \alpha = w \\ [\mathsf{A}]\, B_a \mathcal{K}_a K_a \varphi & \text{if } \alpha = sp \\ [\mathsf{A}]\, \widehat{B}_a \mathcal{K}_a K_a \varphi & \text{if } \alpha = wp \end{cases}$$

$$[\mathsf{skip}]_\alpha \varphi := \varphi$$

$$\left[\mathsf{if}\ \varphi'\ \mathsf{then}\ \pi\ \mathsf{else}\ \pi'\right]_\alpha \varphi := (K_a \varphi' \to [\pi]_\alpha \varphi) \wedge (\neg K_a \varphi' \to [\pi']_\alpha \varphi)$$

$$[\pi;\pi']_\alpha \varphi := [\pi]_\alpha([\pi']_\alpha \varphi)$$

As before, $[\cdot]_s$ is *strong translation*, $[\cdot]_w$ the *weak translation*, $[\cdot]_{sp}$ the *strong plausibility* translation and $[\cdot]_{wp}$ the *weak plausibility* translation.

Lemma 5.21. *Let* S *and* A *be a state and action for agent a, and* φ *a formula of* $L(P,A)$*. Then:*

1. $\mathsf{S} \models [\mathsf{A}]_s \varphi$ iff $\mathsf{S} \models \langle \mathsf{A} \rangle \top$ *and for each* $\mathsf{S}' \in \mathcal{K}_a(\mathsf{S} \otimes \mathsf{A}) : \mathsf{S}' \models \varphi$.
2. $\mathsf{S} \models [\mathsf{A}]_w \varphi$ iff $\mathsf{S} \models \langle \mathsf{A} \rangle \top$ *and for some* $\mathsf{S}' \in \mathcal{K}_a(\mathsf{S} \otimes \mathsf{A}) : \mathsf{S}' \models \varphi$.
3. $\mathsf{S} \models [\mathsf{A}]_{sp} \varphi$ iff $\mathsf{S} \models \langle \mathsf{A} \rangle \top$ *and for each* $\mathsf{S}' \in Min_a(\mathcal{K}_a(\mathsf{S} \otimes \mathsf{A})) : \mathsf{S}' \models \varphi$.
4. $\mathsf{S} \models [\mathsf{A}]_{wp} \varphi$ iff $\mathsf{S} \models \langle \mathsf{A} \rangle \top$ *and for some* $\mathsf{S}' \in Min_a(\mathcal{K}_a(\mathsf{S} \otimes \mathsf{A})) : \mathsf{S}' \models \varphi$.
5. $\mathsf{S} \models [\mathsf{if}\ \varphi'\ \mathsf{then}\ \pi\ \mathsf{else}\ \pi']_\alpha \varphi$ iff
   $(\mathsf{S} \models \varphi'$ *implies* $\mathsf{S} \models [\pi]_\alpha \varphi)$ *and* $(\mathsf{S} \not\models \varphi'$ *implies* $\mathsf{S} \models [\pi']_\alpha \varphi)$.

Note to the reader: *A proof that the translations give the properties stated in the lemma above would be preferable. Given Lemmas 5.16 and 5.17, the proof should be a reasonably easy generalisation of the proof of Lemma 3.17.*

Definition 5.22 (Planning Problems and Solutions)**.** Let P be a finite set of propositional symbols and A a finite set of agents. A planning problem on P,A for $a \in A$ is a triple $\mathcal{P} = (\mathsf{S}_0, AL, \varphi_g)$ where

- S_0 is a state for a on P,A called the *initial state*.
- AL is a finite set of actions for a on $L(P,A)$ called the *action library*.
- $\varphi_g \in L(P,A)$ is the *goal (formula)*.

A plan $\pi \in L(P,A,AL)$ is an α*-solution* to $\mathcal{P}$ if $\mathsf{S}_0 \models [\pi]_\alpha \varphi_g$. For a specific choice of $\alpha = s/w/sp/wp$, π is called a *strong/weak/strong plausibility/weak plausibility*-solution respectively.

A sound and complete algorithm for synthesising $s/w/sp/wp$-solutions for planning problemson single-agent models was given in Chapter 3, Section 3.4. Based on constructing an AND-OR-tree (a *planning tree*), a definition using multi-agent models is the same when using states, actions and applicability as defined in this chapter.

Definition 5.23 (Planning Tree)**.** A *planning tree* for an agent a is a finite, labelled AND-OR tree in which each node n is labelled by a state $S(n)$ for a, and each edge (n,m) leaving an OR-node is labelled by an action $\mathsf{A}(n,m)$ for as.

A planning tree for $\mathcal{P} = (\mathsf{S}_0, AL, \varphi_g)$ is constructed as follows: Let the initial planning tree T_0 consist of just one OR-node $root(T_0)$ with $\mathsf{S}(root(T_0)) = \mathsf{S}_0$ (the root labels the initial state). A planning tree for $\mathcal{P}$ is then any tree that can be constructed from T_0 by repeated applications of the following tree expansion rule:

Definition 5.24 (Tree Expansion Rule)**.** Let T be a planning tree for a planning problem $\mathcal{P} = (\mathsf{S}_0, AL, \varphi_g)$ and an agent i. The tree expansion rule is defined as follows. Pick an OR-node n in T and an action $\mathsf{A} \in AL$ applicable in $\mathsf{S}(n)$ (both for a), where A does not label any existing outgoing edges from n. Then:

1. Add a new AND-node m to T with $\mathsf{S}(m) = \mathsf{S}(n) \otimes \mathsf{A}$, and add an edge (n,m) with $\mathsf{A}(n,m) = \mathsf{A}$.
2. For $\mathsf{S}' \in \mathcal{K}_a \mathsf{S}(m))$, add an OR-node m' with $\mathsf{S}(m') = \mathsf{S}'$ and add the edge (m,m').

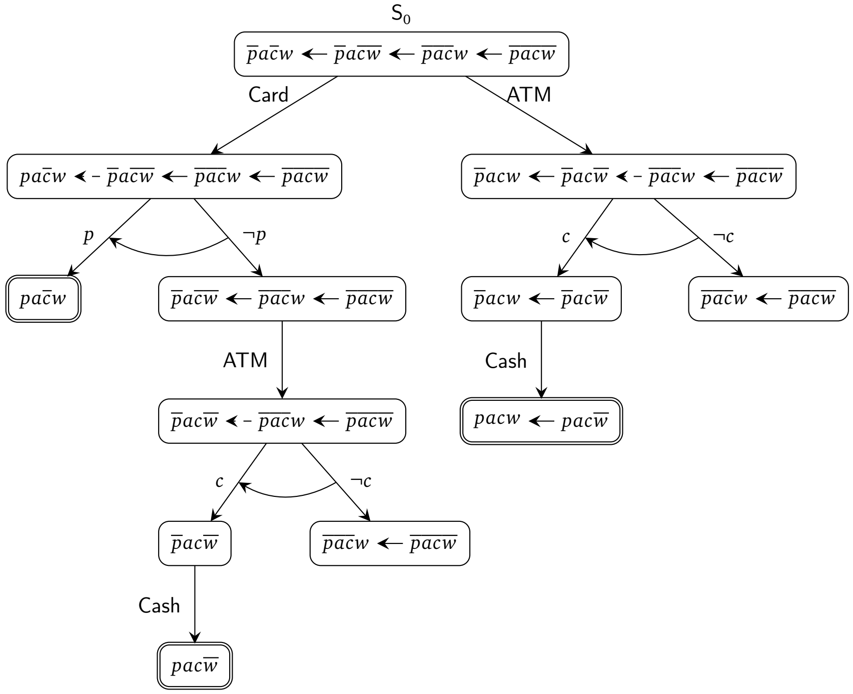
S_0
$\overline{p}a\overline{c}w \leftarrow \overline{p}a\overline{cw} \leftarrow \overline{pac}w \leftarrow \overline{pacw}$
Card
ATM
$pa\overline{c}w \dashleftarrow \overline{p}a\overline{cw} \leftarrow \overline{pac}w \leftarrow \overline{pacw}$
$\overline{p}acw \leftarrow \overline{p}ac\overline{w} \dashleftarrow \overline{pac}w \leftarrow \overline{pacw}$
p
$\neg p$
c
$\neg c$
$pa\overline{c}w$
$\overline{p}a\overline{cw} \leftarrow \overline{pac}w \leftarrow \overline{pacw}$
$\overline{p}acw \leftarrow \overline{p}ac\overline{w}$
$\overline{pac}w \leftarrow \overline{pacw}$
ATM
Cash
$\overline{p}ac\overline{w} \dashleftarrow \overline{pac}w \leftarrow \overline{pacw}$
$pacw \leftarrow pac\overline{w}$
c
$\neg c$
$\overline{p}ac\overline{w}$
$\overline{pac}w \leftarrow \overline{pacw}$
Cash
$pac\overline{w}$

Figure 5.8: A planning tree T for $(\mathsf{S}_0, \{\mathsf{Card}, \mathsf{ATM}, \mathsf{Cash}\}, p)$. The plan $\pi = \mathsf{Card}$ is a strong plausibility solution ($\mathsf{S}_0 \models [\pi]_{sp} p$). Each node contains a visually compacted and simplified plausibility model. Reflexive and transitive edges are left out. Most plausible children of AND-nodes are shown by the direction on AND-edges of the corresponding percepts. Doubly drawn nodes are nodes where the goal formula holds.

A planning tree for the problem $(\mathsf{S}_0, \{\mathsf{Card}, \mathsf{ATM}, \mathsf{Cash}\}, p)$ is shown in Figure 5.8.

Important to note here is that planning with multi-agent models can at best be semidecidable [Bolander and Andersen, 2011]. To ensure at least that, the tree expansion rule must be used to build the tree breadth-first. I leave out definitions of solved nodes, plan extraction and the planning algorithm, as they are easy generalisations of the single-agent versions (Definitions 3.30 and 3.31, and Section 3.4.3).

5.3.1 Restrictions & Revisions

The intended reading of the ordering in states and actions really is that of defeasible belief [Baltag and Smets, 2008b]. I don't say, and do not wish to say, things like "agent a believes that w_1 is fourteen times as likely as w_2". A translating between plausibilities and probabilities is neither intended nor desirable. We really are supposed to understand $w_2 >_a w_1$ as "Until given evidence to the contrary, agent a is sure that w_1 is the case". If S_0 really is the agent's view of the world, then we might well ask: *Why would she bother with all but the most plausible worlds and events?*

In Figure 5.8 we see how an agent would plan if only discarding 'epistemically safe' alternatives. For larger domains, it is easy to imagine that keeping track of all epistemic possibilities is too computationally expensive to be feasible. Wanting to explore the actual implementation of agents using epistemic and doxastic planning, this issue must be addressed. We want an agent that puts her money where her mind is. If she really trusts her beliefs, she will only consider less plausible worlds and events, once she comes to know that the most plausible ones are not the case.

Definition 5.25 (Belief restrictions). Let $\mathsf{S} = (M, W_d) = ((W, \sim, \geq, V), W_d)$ be a (pre)state, $X \subseteq D(M)$ and $G \subseteq A$. First note that for $x \in X$, the set of worlds plan time believed at x to degree k by least one agent in G is

$$\bigcup_{i \in G} Min_i^k([x]_{(\geq_i \cup \leq_i)})$$

Now the set of worlds plan time believed to degree k at at least one world in X by at least one agent in G is

$$\mathcal{B}^k_{G,\mathsf{S}}(X) = \bigcup_{x \in X} \bigcup_{i \in G} Min_i^k([x]_{(\geq_i \cup \leq_i)})$$

Letting $n \geq 0$, the set of worlds believed to degree k by G and depth n is defined as

$$\mathcal{B}^{k,0}_{G,\mathsf{S}}(X) = \mathcal{B}^k_{G,\mathsf{S}}(X)$$
$$\mathcal{B}^{k,n+1}_{G,\mathsf{S}}(X) = \mathcal{B}_{G,\mathsf{S}}(\mathcal{B}^{k,n}_{G,\mathsf{S}}(X))$$

Finally, the (pre)state S restricted to n-depth, k-degree believed worlds is

$$\mathcal{B}^{k,n}_G(\mathsf{S}) = \mathsf{S} \upharpoonright \mathcal{B}^{k,n}_{G,\mathsf{S}}(W_d)$$

where $\mathsf{S} \upharpoonright X = (M \upharpoonright X, W_d \upharpoonright X)$. Writing $\mathcal{B}^k_G(\mathsf{S})$ means $\mathcal{B}^{k,n}_{G,\mathsf{S}}(W_d)$, where n is such that $\mathcal{B}^{k,n+1}_{G,\mathsf{S}}(W_d) = \mathcal{B}^{k,n}_{G,\mathsf{S}}(W_d)$. Writing $\mathcal{B}_G(\mathsf{S})$ means $\mathcal{B}^0_G(\mathsf{S})$ and writing $\mathcal{B}(\mathsf{S})$ means $\mathcal{B}^0_A(\mathsf{S})$.

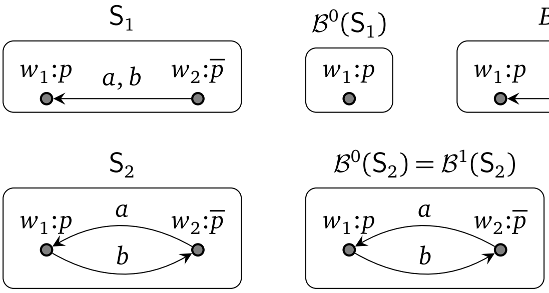

Figure 5.9: Belief restrictions for the states S_1 and S_2.

Example 5.26. Figure 5.9 shows two states S_1 and S_2, along with belief restrictions of degree 0 and 1. We have that $\mathcal{B}^0_{\{a,b\},\mathsf{S}_1}(\{w_1,w_2\}) = Min_a\{w_1,w_2\} \cup Min_b\{w_1,w_2\} = \{w_1\}$, so $\mathcal{B}^0(\mathsf{S}_1) = \mathsf{S}_1 \upharpoonright \{w_1\}$. Similarly, $\mathcal{B}^1(\mathsf{S}_1) = \mathsf{S}_1 \upharpoonright (Min^1_a\{w_1,w_2\} \cup Min^1_b\{w_1,w_2\}) = \mathsf{S}_1 \upharpoonright \{w_1,w_2\}$. Finally, we have $\mathcal{B}^0(\mathsf{S}_2) = \mathsf{S}_2 \upharpoonright (Min^0_a\{w_1,w_2\} \cup Min^0_b\{w_1,w_2\}) = \mathsf{S}_2 \upharpoonright \{w_1,w_2\} = \mathcal{B}^1(\mathsf{S}_2)$. ■

Example 5.27. Figure 5.10 shows possible ways that the agent's model can evolve under the $\mathcal{B}^0$ and $\mathcal{B}^1$ restrictions of S_0. Whether the agent plans with $\mathcal{B}^0(\mathsf{S}_0)$ or $\mathcal{B}^1(\mathsf{S}_0)$ is her choice, as is the choice of first doing Card (which is, as we saw earlier, the first action in a strong plausibility plan for achieving p). She does not choose whether the environment responds with p or $\neg p$. Suppose that she chooses the $\mathcal{B}^0$ restriction and then does her planning thing, coming up with the solution of just doing Card. If, as her plausibilities indicate, the actual world is w_1, the environment responds with p. She achieves her goal and still has a consistent model of the world (i.e. $\mathcal{B}^0(\mathsf{S}_0) \otimes \mathsf{Card}|p \neq \emptyset$). Everything is fine. On the other hand, if the actual world is *not* w_1 she ends up with $\mathcal{B}^0(\mathsf{S}_0) \otimes \mathsf{Card}|\neg p = \emptyset$. If she has no way of revising her model she is lost! Luckily she does, provided she keeps track of the past. ■

Definition 5.28 (History). A history $\mathsf{H} = [(\mathsf{A}_1,\rho_1),\dots,(\mathsf{A}_n,\rho_n)]$ on $L(P,A)$ and AL is a sequence, where $\mathsf{A}_i \in AL$ and $\rho_i \in L(P,A)$. A history is to be interpreted such that A_i is the action chosen at time i and ρ_i is the percept received immediately after. A history may be empty. A pair (S,H), where S is a state for a on P,A, and H is a history on $L(P,A)$ and AL containing only actions for a, is a history-model.

Definition 5.29 (Restriction update). With the prestate S and action A, the restriction-update is defined as $\mathsf{S} \otimes^n \mathsf{A}$ as $\mathcal{B}^n(\mathsf{S} \otimes \mathsf{A})$.

Definition 5.30 (Restriction sequence). A restriction sequence for a history-model (S,H) is a sequence $N = n_0,\dots,n_{|\mathsf{H}|} \in \mathbb{N}^{|\mathsf{H}|+1}$. We say that N is a consistent restriction sequence of degree n for (S,H) if

$$\mathcal{B}^{n+n_0}(\mathsf{S}) \otimes^{n+n_1} \mathsf{A}_1|\rho_1 \otimes^{n+n_2} \mathsf{A}_2|\rho_2 \cdots \otimes^{n+n_{|\mathsf{H}|}} \mathsf{A}_{|\mathsf{H}|}|\rho_{|\mathsf{H}|} \neq \emptyset$$

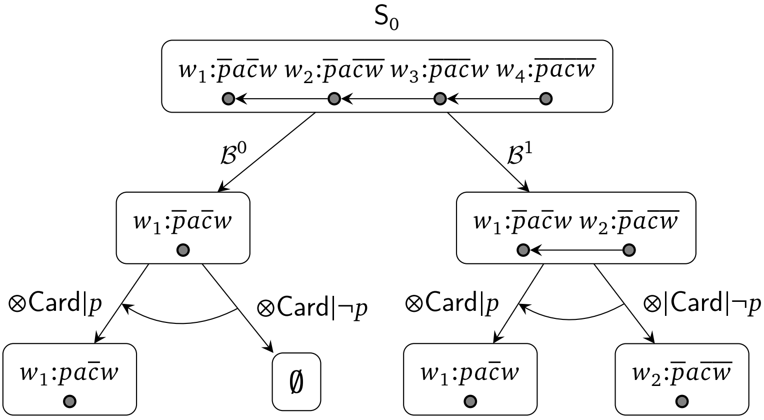
S_0
$w_1{:}\overline{p}a\overline{c}w$ $w_2{:}\overline{p}a\overline{cw}$ $w_3{:}\overline{pac}w$ $w_4{:}\overline{pacw}$
$\mathcal{B}^0$
$\mathcal{B}^1$
$w_1{:}\overline{p}a\overline{c}w$
$w_1{:}\overline{p}a\overline{c}w$ $w_2{:}\overline{p}a\overline{cw}$
$\otimes$Card$|p$
$\otimes$Card$|\neg p$
$\otimes$Card$|p$
$\otimes|$Card$|\neg p$
$w_1{:}pa\overline{c}w$
$\emptyset$
$w_1{:}pa\overline{c}w$
$w_2{:}\overline{p}a\overline{cw}$

Figure 5.10: Possible ways the agent's model can evolve depending on restrictions on the initial state and what is observed after doing Card.

A restriction sequence of degree 0 will just be called a restriction sequence.

When the feedback from the environment invalidates the current state, like in the case of $\mathcal{B}^0(\mathsf{S}_0) \otimes \mathsf{Card}|\neg p = \emptyset$, the agent must come up with a new model of the world. This model must be consistent with her past choice of actions and percepts from the environment. The question is: When a restriction sequence induces the empty model, which restriction sequence should be the next in line?

Definition 5.31 (Restriction sequence ordering)**.** For two restriction sequences $N = n_0, \ldots, n_k$ and $N' = n'_0, \ldots, n'_k$ for (S, H), we let $N > N'$ if $\exists i$ s.t. $\forall j > i : n_j = n'_j$ and $n_i > n'_i$, i.e. they agree on the last $|N| - i$ numbers and ith number of N' is smaller than the ith number of N.

The minimal consistent restriction sequence for a given (S, H) can then be defined as that consistent restriction sequence N' for which all other consistent restriction sequences N for the same history-model have $N > N'$. If $N \not> N'$ and $N \not< N'$ then $N = N'$.

Example 5.32. With the four restriction sequences $[0, 0, 0]$, $[0, 0, 1]$, $[0, 1, 0]$ and $[1, 0, 0]$ we have

$$[0, 0, 1] > [0, 1, 0] > [1, 0, 0] > [0, 0, 0]$$

This is easier to see if we read the sequences converted into strings from right to left

$$100 > 010 > 001 > 000$$

■

I now define the progression update, analogous to the product update, but for history-models and restriction sequences.

Definition 5.33 (Progression update)**.** Given a history-model (S,H), we let the *progression update* (of degree n) of S with H be

$$\mathsf{S} \rhd \mathsf{H} = \mathcal{B}^{n+n_0}(\mathsf{S}) \otimes^{n+n_1} \mathsf{A}_1|\rho_1 \otimes \cdots \otimes^{n+n_k} \mathsf{A}_k|\rho_k$$

, where $N = n_0, \ldots, n_k$ is the minimal consistent restriction sequence of for S,H.

Overloading the product update operator, we let the *knowledge progression* be

$$\mathsf{S} \otimes \mathsf{H} = \mathsf{S} \otimes \mathsf{A}_1|\rho_1 \otimes \cdots \otimes \mathsf{A}_k|\rho_k$$

I will write $S,\mathsf{H} \rhd (\mathsf{A},\rho)$ for $S \rhd (\mathsf{H}||(\mathsf{A},\rho))$. The same notation will be used for $\otimes$.

In the examples to follow, progression updates will only be of degree 0.

Example 5.34. Consider the history-model S_0, [ ], representing the initial situation before the agent does Card. As nothing has happened yet, the history is simply empty. We then have $\mathsf{S}_0 \rhd [\ ] = \mathcal{B}^0(\mathsf{S}_0)$. If the environment responds with p when the agent does Card, the $\rhd$-updated model is $\mathsf{S}_0 \rhd (\mathsf{Card}, p) = pa\bar{c}w$ (the singleton model in the bottom left of Figure 5.10), as the restriction sequence $[0,0]$ is minimal and consistent for $\mathsf{S}, [(\mathsf{Card}, p)]$.

Suppose instead that the environment responds with $\neg p$. As $[0,0]$ is not consistent for $(\mathsf{S}_0, [(\mathsf{Card}, \neg p)]$ we must look for a new restriction sequence. The second smallest restriction sequence is $[1,0]$, which is consistent and gives $\mathsf{S}_0 \rhd (\mathsf{Card}, p) = \mathcal{B}^1(\mathsf{S}_0) \otimes^0 \mathsf{Card}|\neg p = \overline{p}a\overline{cw}$ (the bottom right singleton model of Figure 5.10). ■

Example 5.35. In Figure 5.11 we can see the model that an agent thinking with progressions would maintain for the various possible histories that could come about when attempting to buy records. Initially, the agent really thinks that $\mathcal{B}^0(\mathsf{S}_0) = \overline{p}a\bar{c}w$ is the state of the world. The nodes show the result of the progression of S_0 with the history accumulated by following the edges in the "revision tree". For instance, the bottom left model is $\mathsf{S}_0 \rhd [(\mathsf{Card}, \neg p), (\mathsf{ATM}, c), (\mathsf{Cash}, p)]$, whereas the bottom right model is $\mathsf{S}_0 \rhd [(\mathsf{Card}, \neg p), (\mathsf{ATM}, \neg c)]$. ■

A fair question is how $\rhd$- and $\otimes$-progressions differ. Let us illustrate the difference with a new example.

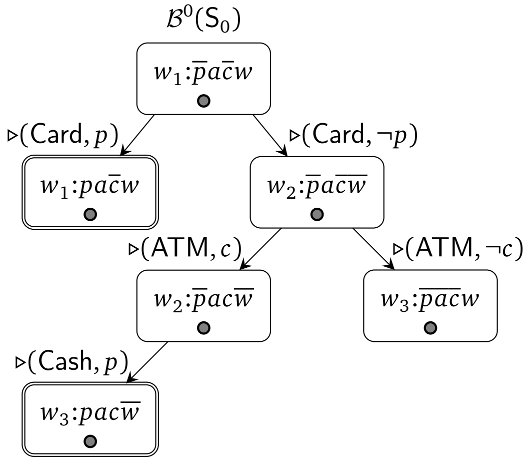
$\mathcal{B}^0(\mathsf{S}_0)$
$w_1{:}\overline{p}a\overline{c}w$
$\triangleright(\mathsf{Card}, p)$
$\triangleright(\mathsf{Card}, \neg p)$
$w_1{:}pa\overline{c}w$
$w_2{:}\overline{p}a\overline{c}\overline{w}$
$\triangleright(\mathsf{ATM}, c)$
$\triangleright(\mathsf{ATM}, \neg c)$
$w_2{:}\overline{p}ac\overline{w}$
$w_3{:}\overline{pac}w$
$\triangleright(\mathsf{Cash}, p)$
$w_3{:}pac\overline{w}$

Figure 5.11: Tree of progressions with $\triangleright$. Doubly drawn nodes indicate nodes where p has been achieved. The secondary effects of belief revision can be seen in bottom left node. Prior to $(\mathsf{ATM}, \neg c)$, the agent believes that the card reader doesn't work $(\overline{w})$. After $(\mathsf{ATM}, \neg c)$ she believes that it does work (w), because the most plausible cause of $\neg p$ in response to Cash is that there is no money on her account $(\overline{a})$.

Example 5.36 (Soft-updates and hidden cointosses)**.** Consider the hidden, biased coin example again. The coin shows either heads (h) or tails $(\neg h)$, and is biased towards landing heads after Toss. The agent can also Check which side landed up. An initial state where the coin is known to show heads is shown in Figure 5.12.

The agent does Toss, after which $\top$ is perceived. For any formula φ holding in $\mathsf{S}_0 \triangleright (\mathsf{Toss}, \top) = \mathcal{B}^0(\mathsf{S}_0) \otimes^0 \mathsf{Toss}|\top$, we have $\mathsf{S}_0 \otimes \mathsf{Toss}|\top \models B\varphi$. Further, for $\mathsf{H} = [(\mathsf{Toss}, \top), (\mathsf{Check}, h)]$, we have $\mathsf{S}_0 \triangleright \mathsf{H} \models \varphi$ iff $\mathsf{S}_0 \otimes H \models \varphi$.[7] The same holds for $\mathsf{H} = [(\mathsf{Toss}, \top), (\mathsf{Check}, \neg h)]$.

Consider instead the history $\mathsf{H} = [(\mathsf{Toss}, \top), (\mathsf{BelieveTails}, \top)]$. BelieveTails is a type of action known in the literature as a soft update [Baltag and Smets, 2008b]. It does not change ontic or epistemic facts, but does change beliefs. Now we have a discrepancy between $\triangleright$ and $\otimes$. Because $\mathcal{B}^0(\mathsf{S}_0) \otimes^0 \mathsf{Toss} = h$, the soft update does nothing (there is no $\neg h$ world to make more plausible). Thus we have $\mathcal{B}^0(\mathsf{S}_0) \otimes^0 \mathsf{Toss}|\top \otimes^0 \mathsf{BelieveTails}|\top \models Bh$, while $\mathsf{S}_0 \otimes [(\mathsf{Toss}, \top), (\mathsf{BelieveTails}, \top)] \models B\neg h$. ■

Note to the reader: *A possible way to remedy this is to define the restriction update*

[7] $\mathsf{S}_0 \triangleright \mathsf{H}s = \mathcal{B}^0(\mathsf{S}_0) \otimes^0 \mathsf{Toss}|\top \otimes^0 \mathsf{Check}|h$

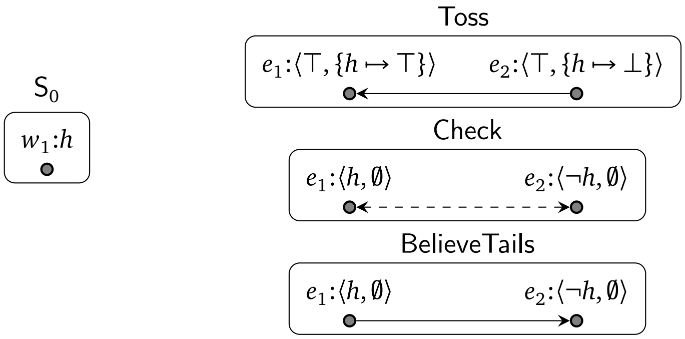
Toss
S_0
$e_1{:}\langle\top,\{h\mapsto\top\}\rangle$
$e_2{:}\langle\top,\{h\mapsto\bot\}\rangle$
$w_1{:}h$
Check
$e_1{:}\langle h,\emptyset\rangle$
$e_2{:}\langle\neg h,\emptyset\rangle$
BelieveTails
$e_1{:}\langle h,\emptyset\rangle$
$e_2{:}\langle\neg h,\emptyset\rangle$

Figure 5.12: The hidden, biased coin toss example, with a soft update for believing $\neg h$.

as $\mathsf{S}\otimes^n\mathsf{A}=\mathcal{B}^n(\mathsf{S}\otimes\mathcal{B}^n(\mathsf{A}))$ *or* $\mathsf{S}\otimes^n\mathsf{A}=\mathsf{S}\otimes\mathcal{B}^n(\mathsf{A})$*. Neither solution is particularly satisfactory. The former because of the unappealing nesting, the latter because it doesn't guarantee that* $\mathsf{S}\otimes^n\mathsf{A}$ *only contains worlds up to degree* n *(*$\mathcal{B}^n(S)\otimes\mathcal{B}^n(\mathsf{A})$ *can have worlds up to degree* n^2*).*

To examine the trade-offs between computational cost and logical fidelity of the two progression types, we revisit the agent algorithm from Chapter 1 with modifications (see Algorithm 2), defining two types of agents by varying the implementation of *update* and *revise*.

Algorithm 2 AgentLoop

```
$\mathsf{H} \leftarrow [\,]$
$\mathsf{S}_C \leftarrow revise(\mathsf{S}_0, \mathsf{H})$
while true do
    $\mathsf{A} \leftarrow nextAction(\mathsf{S}_C)$
    $\rho \leftarrow do(\mathsf{A})$
    Append $(\mathsf{A}, \rho)$ to $\mathsf{H}$
    $\mathsf{S}_N \leftarrow update(\mathsf{S}_C, \mathsf{A}, \rho)$
    if $\mathsf{S}_N = \emptyset$ then
        $\mathsf{S}_N \leftarrow revise(\mathsf{S}_0, \mathsf{H})$
    end if
    $\mathsf{S}_C \leftarrow \mathsf{S}_N$
end while
```

Two obvious agent types can be defined:

Type K With $update(\mathsf{S}, \mathsf{A}, \rho) = \mathsf{S}\otimes\mathsf{A}|\rho$, $\mathsf{S}_N = \emptyset$ never happens, so $revise(\mathsf{S}, \mathsf{H})$

can be anything. For $\mathsf{S}_C \leftarrow revise(\mathsf{S}_0, \mathsf{H})$, just let $revise(\mathsf{S}_0, \mathsf{H}) = \mathsf{S}_0$.

Type B With $update(\mathsf{S}, \mathsf{A}, \rho) = \mathsf{S} \otimes^0 \mathsf{A}|\rho$ and $revise(\mathsf{S}_0, \mathsf{H}) = \mathsf{S}_0 \rhd \mathsf{H}$.

A Type K agent will always take all epistemic possibilities into account when planning (via $nextAction$) and updating the current model. This makes $nextAction$ and $update$ the most expensive, but eliminates the need for belief revision.

Type B uses the restriction update which is (generally) computationally cheaper than for Type K agents, because both S_C and A will (generally) be smaller than for Type K, at the cost of a more expensive $revise$.

Whether a Type K or Type B agent is the better choice *must* be scenario-dependent. If the agent is *very* unlucky, then it may indeed be best to just keep track of all possibilities. With the reasonable assumption that the modelled plausibilities are true to the dynamics of the world, the progression update more often than not gives the true state. In this case, a Type B agent is the way to go.

5.3.2 Synthesising Default Plans

Note to the reader: *In the following, I'm moving into more speculative territory. This is noticeable particularly in the absence of a plan language and translations for default solutions and (a little later) solutions for default planning with multiple acting agents. While it should not be too difficult to define n-degree strong default and weak default solutions, it is not entirely clear that there are neat proofs for showing an analogue of Lemma 5.21. I agree entirely with those readers who think that this is important to get a handle on.*

I now define a default planning method analogous to plausibility planning with multi-agent models, making use of belief restrictions. Planning trees are defined as for plausibility planning with multi-agent models and the tree expansion rule has been amended to use $\otimes^n$.

Default Planning trees (of degree n) for $\mathcal{P} = (\mathsf{S}_0, AL, \varphi_g)$ are constructed as one would expect: Let the initial planning tree T_0 consist of just one OR-node $root(T_0)$ with $\mathsf{S}(root(T_0)) = \mathcal{B}^n(\mathsf{S}_0)$, where S_0 is a state for the planning agent a. The action library AL contains only actions for a. A planning tree for $\mathcal{P}$ is a tree that can be constructed from T_0 by repeated applications of the following tree expansion rule.

Definition 5.37 (Default Planning Tree Expansion Rule). Let T be a default planning tree (of degree n) for a planning problem $\mathcal{P} = (\mathsf{S}_0, AL, \varphi_g)$ with a being the planning agent.

The tree expansion rule is: Pick an OR-node m in T and an event model $\mathsf{A} \in AL$ applicable in $\mathsf{S}(m)$ not already labelling existing an outgoing edge from m. Then:

1. Add a new AND-node m to T with $\mathsf{S}(m') = \mathsf{S}(m) \otimes^n \mathsf{A}$, and add an edge (m, m') with $\mathsf{A}(m, m') = \mathsf{A}$.
2. For each S' in $\mathcal{K}_a(\mathsf{S}(m'))$, add an OR-node m'' with $\mathsf{S}(m'') = \mathsf{S}'$ and add the edge (m', m'').

A couple of things are important to note here: Firstly, we use this construction because the agent truly expects that the evolution of the system modelled by the default planning tree is how things will play out. It may turn out to be wrong when the actions are actually being carried out, but for now she is only imagining the future. We do not label edges from AND-nodes to OR-nodes by percepts. These are only required for revision. The agent just assumes that percepts corresponding to the states will be produced by the environment. Secondly, we are using applicability $\langle \mathsf{A} \rangle \top$ on models which may have had the actual world discarded (if it was not believed in). This can prove a problem for belief revision.

What if there is a pickpocket about the market who steals our happy shopper's money as she is making her way back from the ATM? To model this, assume that ATM includes the possibility of making c false ($e_3 : \langle a, \{c \mapsto \bot\}\rangle$), indistinguishable from $e_1 : \langle a, \{c \mapsto \top\}\rangle$ and less plausible than both e_1 and e_2. The response from the environment is a. There is indeed money on her account. When she does Cash, the environment will respond with something like $\neg p$, $\neg c$ or $\neg p \wedge \neg c$. As Cash does not provide for $\neg c$, there will be no consistent restriction sequence for $\mathsf{S}_0, [(\mathsf{ATM}, a), (\mathsf{Cash}, \neg p)]$. The problem is that $\mathsf{S} \rhd [(\mathsf{ATM}, a)] \models \langle \mathsf{Cash} \rangle \top$, but $\mathsf{S} \otimes [(\mathsf{ATM}, a)] \not\models \langle \mathsf{Cash} \rangle \top$.

We can deal with this by insisting that the preconditions of the events of any particular action covers the entire logical space – as Card and ATM already do. Instead of Cash including only an event with preconditions c, it should include one for $\neg c$ also; the event where paying with Cash fails. We can in fact view both Card and ATM as actions amended to include events for when these actions fail. This need not place an extra burden on the designer of the action library, as the process of adding these fail-events can be easily automated.

Definition 5.38 ($\mathcal{E}$-or fail). Let $\mathsf{A} = ((E, \sim, \geq, pre, post), E_d)$ be an action-model on $L(P,A)$, and let A-or fail $= ((E', \sim', \geq, pre', post'), E'_d)$, where

- $E' = E \cup \{e_{fail}\}$,
- $\sim'_i = \sim_i \cup \{e_{fail}, e_{fail}\}$

- $\geq_i' = \geq_i \cup \{(e_{fail}, e) | e \in E\}$
- $E_d' = \{e_{fail}\}$.
- pre' and $post'$ are unchanged for $e \in E$
- $pre'(e_{fail}) = \neg \bigwedge_{e \in E} pre(e)$ and $post'(e)(p) = p$.

As both Card and ATM already cover the entire logical space, *-or fail*ing them just add events whose preconditions amount to $\perp$. They can therefore be left unchanged. For Cash we get that Cash-or fail has a new event $e_{fail} : \langle \neg c, \emptyset \rangle$ distinguishable from and less plausible than e_1 for all agents. It the *or-fail* versions we use for progressions.

We still wish to use the unmodified versions when (default) planning, as for any action A and state S, S $\models \langle$A-or fail$\rangle\top$. If using the *or fail*-versions for planning, the entire action library becomes applicable everywhere, increasing the number of choices the agent has at each OR-node. This gives a different understanding of applicability. An action has a primary meaning encoded by A, such as paying with cash, whereas the A-*or fail* encodes the primary meaning + possible failures. Redefining the progression operators to use A_1-*or fail*, $\ldots \mathsf{A}_k$-*or fail* instead of $\mathsf{A}_1, \ldots \mathsf{A}_k$, we get $\mathsf{S}_0 \rhd [(\mathsf{ATM}, a)] \models c$ and $\mathsf{S}_0 \rhd [(\mathsf{ATM}, a), (\mathsf{Cash}, \neg p)] \models \neg c$.

With the amended progression operator, the first full procedure for an agent wishing to plan with epistemic and doxastic concepts, but unable to (computationally) afford plausibility planning is straight forward to define. We do assume though, that most of the time, the agent's plausibilities and expectations are correct. If they were wrong more often than right, then the computational savings from doing default planning would be eclipsed by the cost of belief revision.

As for plausibility planning with multi-agent models, we unfortunately still have that default planning is at best semi-decidable (if the tree is expanded breadth-first).

5.4 Discussion

In this chapter we have seen how to define two new types of epistemic planning. One is a generalisation of plausibility planning to multi-agent models, while the other is a specialisation aimed at making plausibility planning less computationally expensive. The latter formalism, named default planning, requires the introduction of a belief revision mechanism. A number of weaknesses have been pointed

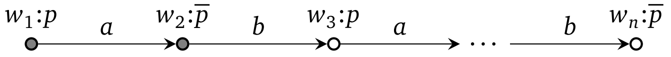

Figure 5.13: Arbitrarily long, but finite chain in a state for a.

out in places where essential proofs or definitions are missing. One of the first steps towards getting the ideas presented in this chapter up to scratch is showing soundness for these new formalisms (as we indeed cannot have completeness).

The underlying assumption that must hold for default planning to be a good idea, is that the plausibilities in states and actions reflect the reality that is being modelled. If beliefs are reliable and things go as expected *most of the time*, then for a resource-bounded agent default planning is a better choice than plausibility planning.

Additionally, there are two particularly interesting lines of inquiry suggested by the ideas in this chapter:

The first concerns decidability. As for plausibility planning with multi-agent models, we unfortunately still have that default planning is at best semi-decidable (if the tree is expanded breadth-first). The undecidability proof from [Bolander and Andersen, 2011] (based on constructing a planning problem $\mathcal{P}_{TM}$ for an arbitrary Turing machine TM, such that there is a solution to $\mathcal{P}_{TM}$ iff TM halts) still works for default planning, even degree 0, simply by making all worlds and events in $\mathcal{P}_{TM}$ equiplausible. Essentially, the proof relies on there being no upper no upper bound on the size of models that the planning procedure might have to go through in order to get to a state in which the goal formula holds (even though there is an upper bound on the size of goal states). That proof, and those in [Aucher and Bolander, 2013] (based on the halting problem for 2-counter machines), require incontractible models, meaning that the planning procedure must be able to generate arbitrarily long chains of alternating $\sim_i$- and $\sim_j$-edges. But consider this:

Figure 5.13 shows a state for a. Because of the requirements for planning, this wouldn't have been a state if W_d contained more than w_1 and w_2. As any state S has $\forall w \in W_d : W_d = [w]_a$, restricting states to those worlds no more than k steps away from a world in W_d (like done in a k-depth belief restriction) means that that we get an upper bound on the size of W_d. If planning for a goal formula of modal depth k, we can restrict states to those worlds which are at most k steps from W_d, because the goal formula does not say anything worlds further away than k. This argument goes along the same lines as the proofs for decidability for the fragments of multi-agent planning in [Yu et al., 2013]. An essential difference is that for decidability of "k-restricted" multi-agent plausibility planning, we need only make assumptions about the events which are not designated, instead of all events.

The second line of inquiry is into a formalism for allowing an agent to model other *acting* agents. With inspiration taken from [Markey and Vester, 2014], a first step would be defining 2-player planning games, akin to game trees for classic games like chess, checkers or Tic-Tac-Toe. With alternating turns, a 2-player planning (game) tree would simulate the moves of the non-planning agent by planning with the roles of planning and non-planning agent reversed: If a is the planning agent, b is the other agent and S is a prestate for a, then a can hypothese about b's expected moves in S by trying to plan as b in each of $Min_a\mathcal{K}_b\mathsf{S}$. The first actions in the plan(s) for b's goal (assuming that goals are common knowledge) would then be the actions that a would plan for responding to.

As a simple example, let's suppose that the vendor selling records is modelled as an explicitly acting agent. He has two actions, skip which changes nothing, and sell which consists of just a single event e with $pre(e) = p$ and $post(r) \to \top$. In the initial state, r is false everywhere, indicating that our vinyl happy shopper has not yet received the records. Instead of p, she wants to achieve r. Having the records is what is at stake for her, not (necessarily) paying. The vendor has the goal $p \to r$ – only when having received the payment will he want to hand over the records. Until the shopper successfully pays and makes p true, he will do nothing. His response to every action taken by the shopper will be skip, because his goal is already achieved. Once the shopper makes p true, $p \to r$ is no longer true. Thus when the shopper imagines to be the vendor and plans for $p \to r$ in his stead, the plan is going to consist of sell. This shifts our understanding of the planning being done by the shopper from being about making p true, to being about reaching a state in which the vendor wants to bring about her goal. If she cannot achieve her goals on her own, or if it leads to her goal being achieved in fewer actions, she can change the state so that the other agent(s) assist her. That would truly be planning with a Theory of Mind.

Chapter 6

Conclusion and Future Work

The main results of thesis are new ways of doing automated planning with Dynamic Epistemic Logic, and model theoretic results for plausibility models. Two novel approaches, conditional epistemic planning and plausibility planning, have been defined, along with terminating, sound and complete algorithms for synthesising plans of various strengths. For multi-agent plausibility models, we have shown a new approach to bisimulations, reestablishing the correspondence between bisimulation and modal equivalence for three canonical logics. We also investigated the expressivity of these three logics (and combinations thereof). In spite of our earlier conjectures, we surprised ourselves by showing that these logics are indeed not equally expressive. These results have a direct influence on further developments in epistemic planning, as properties of both models and logics have consequences for decidability and specification of such planning problems. Finally, I presented preliminary work on plausibility planning with multi-agent models and a restricted type of planning incorporating a notion of belief revision, intended for agents on the computational cheap.

There is a multitude of open problems, but here I mention what I believe are the most promising and pressing issues for further investigation.

True multi-agent planning Having left off just before getting into the first shot at true multi-agent planning, this is the direction I find most interesting. A test bed as in [Brenner and Nebel, 2009], requiring working agents that can be empirically tested, should be the benchmark here.

Formal verification of games The logics of ATL and ATL* and the concurrent game structures they are interpreted on, seem to be the most promising places to find already developed tools for analysing true multi-agent planning. Decidability and complexity results, model checking algorithms, symmetries (akin to restricted bisimulations [Markey and Vester, 2014]) and classification of strategies seem to be ripe for application in epistemic planning, particularly the very interesting multi-agent case. It is, in fact, not difficult to imagine that automated planning is just model checking a concurrent game structure that is generated with actions, rather than given as input.

Specification languages If we criticise the theory of games for being indifferent to where the transition system comes from, then we must subject ourselves to as similar critique. While building event models for encoding actions is not as impossible as building a complete transition system by hand, there is still a considerable amount of craftsmanship involved. Particularly when there are many agents present, observability can be very problematic to model by hand—who is observing that others are observing that still others are observing? Suitable specification languages are needed if epistemic planning methods are to be adopted by non-experts. While the language proposed by [Baral et al., 2012] is a good start, this does not seem to be the end of it. Particularly higher-order observability is not covered by that otherwise fine approach.

While the road ahead is still long, we've come far enough to say with some confidence that there is a lot of promise in techniques like those presented in this thesis. As we surround ourselves with more and more technology, maybe it is time to start demanding that we interact with it on our own terms, rather than the technology demanding that we adapt to it. Perhaps our computers and smartphones (for a start) will someday not just be autistic sources of frustration, but accommodating and understanding pieces of technology. I think this will one day be the case, and I think that artificial agents with a (possibly limited) Theory of Mind are going to be a huge step in this direction. While we may have to wait a long time for the day when I'm proven right (if it ever comes), I can at least take comfort in the knowledge that it will be nearly impossible to prove me wrong.

Bibliography

[Alur et al., 2002] Alur, R., Henzinger, T. A., and Kupferman, O. (2002). Alternating-time temporal logic. *J. ACM*, 49(5):672–713.

[Alur et al., 1998] Alur, R., Henzinger, T. A., Kupferman, O., and Vardi, M. Y. (1998). Alternating refinement relations. In *CONCUR'98 Concurrency Theory*, pages 163–178. Springer.

[Andersen et al., 2012] Andersen, M., Bolander, T., and Jensen, M. (2012). Conditional epistemic planning. In *Proc. of 13th JELIA*, LNCS 7519, pages 94–106. Springer.

[Andersen et al., 2014] Andersen, M., Bolander, T., and Jensen, M. (2014). Don't plan for the unexpected: Planning based on plausibility models. *Logique et Analyse: Special Issue on Dynamics in Logic*.

[Andersen et al., 2013] Andersen, M. B., Bolander, T., van Ditmarsch, H. P., and Jensen, M. H. (2013). Bisimulation for single-agent plausibility models. In Cranefield, S. and Nayak, A., editors, *Australasian Conference on Artificial Intelligence*, volume 8272 of *Lecture Notes in Computer Science*, pages 277–288. Springer.

[Aucher, 2005] Aucher, G. (2005). A combined system for update logic and belief revision. In *Proc. of 7th PRIMA*, pages 1–17. Springer. LNAI 3371.

[Aucher, 2010] Aucher, G. (2010). An internal version of epistemic logic. *Studia Logica*, 94(1):1–22.

[Aucher, 2012] Aucher, G. (2012). Del-sequents for regression and epistemic planning. *Journal of Applied Non-Classical Logics*, 22(4):337–367.

[Aucher and Bolander, 2013] Aucher, G. and Bolander, T. (2013). Undecidability in Epistemic Planning. Rapport de recherche.

[Aumann and Hart, 1992] Aumann, R. and Hart, S., editors (1992). *Handbook of Game Theory with Economic Applications*. Elsevier.

[Baltag and Moss, 2004] Baltag, A. and Moss, L. S. (2004). Logics for Epistemic Programs. *Synthese*, 139:165–224.

[Baltag et al., 1998] Baltag, A., Moss, L. S., and Solecki, S. (1998). The logic of public announcements, common knowledge, and private suspicions. In *Proceedings of the 7th conference on Theoretical aspects of rationality and knowledge*, TARK '98, pages 43–56, San Francisco, CA, USA. Morgan Kaufmann Publishers Inc.

[Baltag and Smets, 2006] Baltag, A. and Smets, S. (2006). Dynamic belief revision over multi-agent plausibility models.

[Baltag and Smets, 2008a] Baltag, A. and Smets, S. (2008a). The logic of conditional doxastic actions. In *New Perspectives on Games and Interaction*, Texts in Logic and Games 4, pages 9–31. Amsterdam University Press.

[Baltag and Smets, 2008b] Baltag, A. and Smets, S. (2008b). A qualitative theory of dynamic interactive belief revision. In Bonanno, G., van der Hoek, W., and Wooldridge, M., editors, *Logic and the Foundations of Game and Decision Theory (LOFT7)*, volume 3 of *Texts in Logic and Games*, pages 13–60. Amsterdam University Press.

[Baral et al., 2012] Baral, C., Gelfond, G., Pontelli, E., and Son, T. C. (2012). An action language for reasoning about beliefs in multi-agent domains. In *Proceedings of the 14th International Workshop on Non-Monotonic Reasoning*.

[Baron-Cohen et al., 1985] Baron-Cohen, S., Leslie, A., and Frith, U. (1985). Does the autistic child have a 'theory of mind'? *Cognition*, 21:37–46.

[Barwise and Moss, 1996] Barwise, J. and Moss, L. (1996). *Vicious circles*. CSLI Publications.

[Bertoli et al., 2003] Bertoli, P., Cimatti, A., Dal Lago, U., and Pistore, M. (2003). Extending pddl to nondeterminism, limited sensing and iterative conditional plans. In *Proceedings of ICAPS'03 Workshop on PDDL*. Citeseer.

[Bicks, 2010] Bicks, M. (2010). Artificial Intelligence Pioneer. `http://www.pbs.org/wgbh/nova/tech/pioneer-artificial-intelligence.html`. [Online; accessed 02-October-2013].

[Blackburn et al., 2001] Blackburn, P., de Rijke, M., and Venema, Y. (2001). *Modal Logic*. Cambridge University Press, Cambridge. Cambridge Tracts in Theoretical Computer Science 53.

[Blackburn and van Benthem, 2006] Blackburn, P. and van Benthem, J. (2006). Modal logic: A semantic perspective. In *Handbook of Modal Logic*. Elsevier.

[Bloom and German, 2000] Bloom, P. and German, T. (2000). Two reasons to abandon the false belief task as a test of theory of mind. *Cognition*, 77(1):25–31.

[Bolander and Andersen, 2011] Bolander, T. and Andersen, M. B. (2011). Epistemic planning for single and multi-agent systems. *Journal of Applied Non-Classical Logics*, 21(1):9–34.

[Brenner and Nebel, 2009] Brenner, M. and Nebel, B. (2009). Continual planning and acting in dynamic multiagent environments. *Autonomous Agents and Multi-Agent Systems*, 19(3):297–331.

[Britz and Varzinczak, 2013] Britz, K. and Varzinczak, I. (2013). Defeasible modalities. In *Proc. of the 14th TARK*.

[Bulling et al., 2010] Bulling, N., Dix, J., and Jamroga, W. (2010). Model checking logics of strategic ability: Complexity*. In *Specification and verification of multi-agent systems*, pages 125–159. Springer.

[Colton, 2012] Colton, S. (2012). *The Painting Fool: Stories from Building an Automated Painter,* chapter 1, page 3–38. Springer, Berlin; Heidelberg.

[de Lima, 2007] de Lima, T. (2007). *Optimal Methods for Reasoning about Actions and Plans in Multi-Agents Systems*. PhD thesis, IRIT, University of Toulouse 3, France.

[de Weerd et al., 2013] de Weerd, H., Verbrugge, R., and Verheij, B. (2013). Higher-order theory of mind in negotiations under incomplete information. In *PRIMA 2013: Principles and Practice of Multi-Agent Systems*, pages 101–116. Springer.

[Demey, 2011] Demey, L. (2011). Some remarks on the model theory of epistemic plausibility models. *Journal of Applied Non-Classical Logics*, 21(3-4):375–395.

[Dennis et al., 2007] Dennis, L. A., Farwer, B., Bordini, R. H., Fisher, M., and Wooldridge, M. (2007). A common semantic basis for BDI languages. In *Programming Multi-Agent Systems, 5th International Workshop, ProMAS 2007, Honolulu, HI, USA, May 15, 2007, Revised and Invited Papers*, pages 124–139.

[Ditmarsch et al., 2007] Ditmarsch, H. v., van der Hoek, W., and Kooi, B. (2007). *Dynamic Epistemic Logic*. Springer.

[Fagin et al., 1995] Fagin, R., Halpern, J., Moses, Y., and Vardi, M. (1995). *Reasoning about Knowledge*. MIT Press, Cambridge MA.

[Fagundes et al., 2009] Fagundes, M. S., Vicari, R. M., and Coelho, H. (2009). Agent computing and multi-agent systems. chapter Deliberation Process in a BDI Model with Bayesian Networks, pages 207–218. Springer-Verlag, Berlin, Heidelberg.

[Ferrucci, 2012] Ferrucci, D. A. (2012). Introduction to "this is watson". *IBM Journal of Research and Development*, 56(3):1.

[Fikes and Nilsson, 1971] Fikes, R. E. and Nilsson, N. J. (1971). Strips: A new approach to the application of theorem proving to problem solving. In *Proceedings of the 2Nd International Joint Conference on Artificial Intelligence*, IJCAI'71, pages 608–620, San Francisco, CA, USA. Morgan Kaufmann Publishers Inc.

[Foster and Petrick, 2014] Foster, M. E. and Petrick, R. P. A. (2014). Planning for social interaction with sensor uncertainty. In *Proceedings of the ICAPS 2014 Scheduling and Planning Applications Workshop (SPARK)*, pages 19–20, Portsmouth, New Hampshire, USA.

[Friedman and Halpern, 1994] Friedman, N. and Halpern, J. (1994). A knowledge-based framework for belief change - part i: Foundations. In *Proc. of 5th TARK*, pages 44–64. Morgan Kaufmann.

[Ghallab et al., 1998] Ghallab, M., Howe, A., Knoblock, C., Mcdermott, D., Ram, A., Veloso, M., Weld, D., and Wilkins, D. (1998). PDDL—The Planning Domain Definition Language.

[Ghallab et al., 2004] Ghallab, M., Nau, D. S., and Traverso, P. (2004). *Automated Planning: Theory and Practice*. Morgan Kaufmann.

[Goranko and Jamroga, 2004] Goranko, V. and Jamroga, W. (2004). Comparing semantics of logics for multi-agent systems. *SYNTHESE*, 139(2):241–280.

[Gorankoa and Vester, 2014] Gorankoa, V. and Vester, S. (2014). Optimal decision procedures for satisfiability in fragments of alternating-time temporal logics. *Advances in Modal Logic 2014*.

[Gorniak and Roy, 2005] Gorniak, P. and Roy, D. (2005). Speaking with your sidekick: Understanding situated speech in computer role playing games. In *AIIDE*, pages 57–62.

[Gorniak, 2005] Gorniak, P. J. (2005). *The affordance-based concept*. PhD thesis, Massachusetts Institute of Technology.

[Grove, 1988] Grove, A. (1988). Two modellings for theory change. *Journal of Philosophical Logic*, 17:157–170.

[Halpern, 2003] Halpern, J. (2003). *Reasoning about Uncertainty*. MIT Press, Cambridge MA.

[Helmert, 2004] Helmert, M. (2004). A planning heuristic based on causal graph analysis. In *ICAPS*, volume 16, pages 161–170.

[Helmert and Geffner, 2008] Helmert, M. and Geffner, H. (2008). Unifying the causal graph and additive heuristics. In *ICAPS*, pages 140–147.

[Hepple et al., 2007] Hepple, A., Dennis, L. A., and Fisher, M. (2007). A common basis for agent organisation in BDI languages. In *Languages, Methodologies and Development Tools for Multi-Agent Systems, First International Workshop, LADS 2007, Durham, UK, September 4-6, 2007. Revised Selected Papers*, pages 71–88.

[Herzig et al., 2003] Herzig, A., Lang, J., and Marquis, P. (2003). Action representation and partially observable planning using epistemic logic. In Gottlob, G. and Walsh, T., editors, *IJCAI*, pages 1067–1072. Morgan Kaufmann.

[Hintikka, 1962] Hintikka, J. (1962). *Knowledge and Belief*. Cornell University Press, Ithaca, NY.

[Horrocks et al., 2006] Horrocks, I., Hustadt, U., Sattler, U., and Schmidt, R. (2006). Computational modal logic. In *Handbook of Modal Logic*. Elsevier.

[Jamroga and Ågotnes, 2007] Jamroga, W. and Ågotnes, T. (2007). Constructive knowledge: what agents can achieve under imperfect information. *Journal of Applied Non-Classical Logics*, 17(4):423–475.

[Jamroga and van der Hoek, 2004] Jamroga, W. and van der Hoek, W. (2004). Agents that know how to play. *Fundamenta Informaticae*, 63:185–219.

[Jensen, 2013a] Jensen, M. H. (2013a). The computational complexity of single agent epistemic planning (manuscript).

[Jensen, 2013b] Jensen, M. H. (2013b). Planning using dynamic epistemic logic: Correspondence and complexity. In Grossi, D., Roy, O., and Huang, H., editors, *LORI*, volume 8196 of *Lecture Notes in Computer Science*, pages 316–320. Springer.

[Kahneman, 2011] Kahneman, D. (2011). *Thinking, fast and slow*. Farrar, Straus and Giroux, New York.

[Kraus and Lehmann, 1988] Kraus, S. and Lehmann, D. (1988). Knowledge, belief and time. *Theoretical Computer Science*, 58:155–174.

[Kraus et al., 1990] Kraus, S., Lehmann, D., and Magidor, M. (1990). Nonmonotonic reasoning, preferential models and cumulative logics. *Artificial Intelligence*, 44:167–207.

[Laverny, 2006] Laverny, N. (2006). *Révision, mises à jour et planification en logique doxastique graduelle*. PhD thesis, Institut de Recherche en Informatique de Toulouse (IRIT), Toulouse, France.

[Lenzen, 1978] Lenzen, W. (1978). Recent work in epistemic logic. *Acta Philosophica Fennica*, 30:1–219.

[Lenzen, 2003] Lenzen, W. (2003). Knowledge, belief, and subjective probability: outlines of a unified system of epistemic/doxastic logic. In Hendricks, V., Jorgensen, K., and Pedersen, S., editors, *Knowledge Contributors*, pages 17–31, Dordrecht. Kluwer Academic Publishers. Synthese Library Volume 322.

[Levesque, 2014] Levesque, H. J. (2014). On our best behaviour. *Artif. Intell.*, 212:27–35.

[Lewis, 1973] Lewis, D. (1973). *Counterfactuals*. Harvard University Press, Cambridge (MA).

[Löwe et al., 2011a] Löwe, B., Pacuit, E., and Witzel, A. (2011a). Del planning and some tractable cases. In *Proceedings of the Third International Conference on Logic, Rationality, and Interaction*, LORI'11, pages 179–192, Berlin, Heidelberg. Springer-Verlag.

[Löwe et al., 2011b] Löwe, B., Pacuit, E., and Witzel, A. (2011b). DEL planning and some tractable cases. In *Proc. of LORI 3*, pages 179–192. Springer.

[Markey and Vester, 2014] Markey, N. and Vester, S. (2014). Symmetry reduction in infinite games with finite branching. *In Proc. of International Symposium on Automated Technology for Verification and Analysis*.

[Meyer et al., 2000] Meyer, T., Labuschagne, W., and Heidema, J. (2000). Refined epistemic entrenchment. *Journal of Logic, Language, and Information*, 9:237–259.

[Michlmayr, 2002] Michlmayr, M. (2002). Simulation theory versus theory theory: Theories concerning the ability to read minds. Master's thesis, University of Innsbruck.

[Osborne and Rubinstein, 1994] Osborne, M. J. and Rubinstein, A. (1994). *A course in game theory*. The MIT Press, Cambridge, USA. electronic edition.

[Pauly, 2002] Pauly, M. (2002). A modal logic for coalitional power in games. *Journal of Logic and Computation*, 12(1):149–166.

[Petrick and Bacchus, 2002] Petrick, R. P. A. and Bacchus, F. (2002). A knowledge-based approach to planning with incomplete information and sensing. In Ghallab, M., Hertzberg, J., and Traverso, P., editors, *AIPS*, pages 212–222. AAAI.

[Petrick and Bacchus, 2004] Petrick, R. P. A. and Bacchus, F. (2004). Extending the knowledge-based approach to planning with incomplete information and sensing. In [Zilberstein et al., 2004], pages 2–11.

[Premack and Woodruff, 1978] Premack, D. and Woodruff, G. (1978). Does the chimpanzee have a 'theory of mind'? *Behavioral and Brain Sciences*, 4:515–526.

[Rintanen, 2004] Rintanen, J. (2004). Complexity of planning with partial observability. In [Zilberstein et al., 2004], pages 345–354.

[Segerberg, 1998] Segerberg, K. (1998). Irrevocable belief revision in dynamic doxastic logic. *Notre Dame Journal of Formal Logic*, 39(3):287–306.

[Spohn, 1988] Spohn, W. (1988). Ordinal conditional functions: a dynamic theory of epistemic states. In Harper, W. and Skyrms, B., editors, *Causation in Decision, Belief Change, and Statistics*, volume II, pages 105–134.

[Stalnaker, 1996] Stalnaker, R. (1996). Knowledge, belief and counterfactual reasoning in games. *Economics and Philosophy*, 12:133–163.

[van Benthem, 1998] van Benthem, J. (1998). Dynamic odds and ends. Technical Report ML-1998-08, University of Amsterdam.

[van Benthem, 2001] van Benthem, J. (2001). Games in dynamic-epistemic logic. *Bulletin of Economic Research*, 53(4):219–48.

[van Benthem, 2006] van Benthem, J. (2006). One is a lonely number: on the logic of communication. In *Logic colloquium 2002. Lecture Notes in Logic, Vol. 27*, pages 96–129. A.K. Peters.

[van Benthem, 2007] van Benthem, J. (2007). Dynamic logic of belief revision. *Journal of Applied Non-Classical Logics*, 17(2):129–155.

[Van Benthem, 2011] Van Benthem, J. (2011). Exploring a theory of play. In *Proceedings of the 13th Conference on Theoretical Aspects of Rationality and Knowledge*, pages 12–16. ACM.

[van Benthem, 2011] van Benthem, J. (2011). *Logical Dynamics of Information and Interaction*. Cambridge University Press.

[van Benthem et al., 2007] van Benthem, J., Gerbrandy, J., and Pacuit, E. (2007). Merging frameworks for interaction: Del and etl. In *Proceedings of the 11th conference on Theoretical aspects of rationality and knowledge*, TARK '07, pages 72–81, New York, NY, USA. ACM.

[van der Hoek and Wooldridge, 2003] van der Hoek and Wooldridge (2003). Cooperation, knowledge, and time: alternating-time temporal epistemic logic and its applications. *Studia Logica*, 75(1):125–157.

[van der Hoek, 1992] van der Hoek, W. (1992). On the semantics of graded modalities. *Journal of Applied Non-Classical Logics*, 2(1).

[van der Hoek, 1993] van der Hoek, W. (1993). Systems for knowledge and beliefs. *Journal of Logic and Computation*, 3(2):173–195.

[van der Hoek et al., 2006] van der Hoek, W., Lomuscio, A., and Wooldridge, M. (2006). On the complexity of practical atl model checking. In *Proc. of the Fifth International Joint Conference on Autonomous Agents & Multiagent Systems (AAMAS 2006)*, pages 201–208. ACM.

[van der Hoek and Wooldridge, 2002] van der Hoek, W. and Wooldridge, M. (2002). Tractable multiagent planning for epistemic goals. In *Proc. of the First International Joint Conference on Autonomous Agents & Multiagent Systems (AAMAS 2002)*, pages 1167–1174. ACM.

[van Ditmarsch, 2005] van Ditmarsch, H. (2005). Prolegomena to dynamic logic for belief revision. *Synthese (Knowledge, Rationality & Action)*, 147:229–275.

[van Ditmarsch, 2008] van Ditmarsch, H. (2008). Comments on 'The logic of conditional doxastic actions'. In *New Perspectives on Games and Interaction*, Texts in Logic and Games 4, pages 33–44. Amsterdam University Press.

[van Ditmarsch et al., 2012] van Ditmarsch, H., Fernández-Duque, D., and van der Hoek, W. (2012). On the definability of simulation and bisimulation in epistemic logic. *Journal of Logic and Computation*. doi:10.1093/logcom/exs058.

[van Ditmarsch and Kooi, 2008] van Ditmarsch, H. and Kooi, B. (2008). Semantic results for ontic and epistemic change. In Bonanno, G., van der Hoek, W., and Wooldridge, M., editors, *Logic and the Foundations of Game and Decision Theory (LOFT 7)*, Texts in Logic and Games, pages 87–117. Amsterdam University Press.

[van Ditmarsch and Labuschagne, 2007] van Ditmarsch, H. and Labuschagne, W. (2007). My beliefs about your beliefs – a case study in theory of mind and epistemic logic. *Synthese*, 155:191–209.

[van Ditmarsch et al., 2007] van Ditmarsch, H., van der Hoek, W., and Kooi, B. (2007). *Dynamic Epistemic Logic*, volume 337 of *Synthese Library*. Springer.

[Vester, 2013] Vester, S. (2013). Alternating-time temporal logic with finite-memory strategies. *arXiv preprint arXiv:1307.4476*.

[Vogeley et al., 2001] Vogeley, K., Bussfeld, P., Newen, A., Herrmann, S., Happé, F., Falkai, P., Maier, W., Shah, N. J., Fink, G. R., and Zilles, K. (2001). Mind reading: neural mechanisms of theory of mind and self-perspective. *Neuroimage*, 14(1):170–181.

[von Wright, 1951] von Wright, G. (1951). *An Essay in Modal Logic*. North Holland, Amsterdam.

[Wittgenstein, 1953] Wittgenstein, L. (1953). *Philosophical Investigations*. Basil Blackwell, Oxford.

[Wooldridge, 1996] Wooldridge, M. (1996). A logic of bdi agents with procedural knowledge.

[Wooldridge, 2000] Wooldridge, M. (2000). *Reasoning about Rational Agents*. MIT Press.

[Yu et al., 2013] Yu, Q., Wen, X., and Liu, Y. (2013). Multi-agent epistemic explanatory diagnosis via reasoning about actions. In Rossi, F., editor, *IJCAI*. IJCAI/AAAI.

[Zilberstein et al., 2004] Zilberstein, S., Koehler, J., and Koenig, S., editors (2004). *Proceedings of the Fourteenth International Conference on Automated Planning and Scheduling (ICAPS 2004), June 3-7 2004, Whistler, British Columbia, Canada*. AAAI.