BELIEF CHANGE IN THE PRESENCE OF ACTIONS AND OBSERVATIONS: A TRANSITION SYSTEM APPROACH

by

Aaron Hunter

B.Sc. Pure Mathematics, University of Calgary, 1996

M.Sc. Mathematics, Simon Fraser University, 1998

A THESIS SUBMITTED IN PARTIAL FULFILLMENT
OF THE REQUIREMENTS FOR THE DEGREE OF
DOCTOR OF PHILOSOPHY
in the School
of
Computing Science

SIMON FRASER UNIVERSITY
Summer 2006

APPROVAL

Name:	Aaron Hunter
Degree:	Doctor of Philosophy
Title of thesis:	Belief Change in the Presence of Actions and Observations: A Transition System Approach
Examining Committee:	Dr. Binay Bhattacharya, *Chair*

Dr. James P. Delgrande, *Senior Supervisor*
Professor, Computing Science, SFU

Dr. Eugenia Ternovska, *Supervisor*
Assistant Professor, Computing Science, SFU

Dr. Oliver Schulte, *Supervisor*
Associate Professor, Computing Science, SFU

Dr. Jeffry Pelletier, *SFU Examiner*
Professor, Philosophy and Linguistics, SFU

Dr. Maurice Pagnucco, *External Examiner*
Senior Lecturer, Computer Science and Engineering
University of New South Wales

Date Approved: July 26, 2006

Abstract

The beliefs of an agent should change due to actions and observations. In particular, actions cause an agent to perform belief update, and observations cause an agent to perform belief revision. However, the interaction between actions and observations can be non-elementary; there are simple examples where it is clear that simply updating and revising in succession does not lead to plausible results. In this dissertation, we consider the belief change that occurs due to an iterated sequence of actions and observations.

We assume that the effects of actions are given by a transition system, which can be used to define a belief update operator. We introduce a set of basic postulates that should intuitively be satisfied when a revision is followed by an update, and we use these postulates to define a new belief change operator. The new operator provides a plausible model of iterated belief change in the presence of actions, and it can be characterized by a modified class of systems of spheres. One limitation of this approach to iterated belief change is that it assumes the agent has perfect knowledge of the history of actions executed. We relax this assumption by using ranking functions to represent uncertainty about the actions executed at each point in time. The resulting formalism is able to represent fallible knowledge, erroneous perception, exogenous actions, and failed actions.

Our work is distinguished from related work in that we explicitly consider the manner in which action histories affect the interpretation of observations. Our formal tools are useful not only for the representation of simple action domains, but also for the evaluation of related formalisms involving iterated belief change due to action.

"Why, sometimes I've believed as many as six impossible things before breakfast."

— *The White Queen,* THROUGH THE LOOKING GLASS

Acknowledgments

I would like to sincerely thank my senior supervisor, Jim Delgrande. Jim has been an excellent mentor, and I feel very fortunate to have had the opportunity to work with him. His knowledge and insight have been invaluable in helping shape the overall direction of this research project. Moreover, he has been extremely flexible and accommodating with my schedule, which has allowed me to balance my research time and my family time in an optimal manner. For this, I continue to be very grateful.

In addition to his academic support, I would like to thank Jim for providing a great deal of financial support, both in terms of research assistantships and also in terms of travel support for conferences. I would also like to acknowledge the financial support of David Mitchell, the Dean of Graduate Studies, the Faculty of Applied Sciences, the Advanced Systems Institute of British Columbia, and the BRICS institute at the University of Århus.

On a personal level, I would like to thank my wife Susanne for a great many things – far too many to list here. For the purpose of these acknowledgments, I would like to thank her for providing me with an enormous amount of emotional and practical support throughout the process of writing this thesis. I would also like to thank my parents for their unwaivering and unconditional support; in the context of this thesis, this support has included everything from extended babysitting sessions to valuable academic advice. Finally, I would like to thank my children, Erika and Ryan, for always giving me something better to do at the end of the day.

Contents

Chapter 1

Introduction

1.1 Motivation

1.1.1 The Basic Problem

We are interested in modeling the belief change that is caused by executing a sequence of actions. Broadly speaking, our work falls under the umbrella of logical Artificial Intelligence(AI): we use the formal tools of mathematical logic to model reasoning. We begin with a description of the basic problem at an informal level.

An agent typically does not have complete knowledge about the state of the world. Instead, an agent's beliefs may be both incorrect and incomplete. Informally, we can think of an agent's beliefs as a partial description of some hypothetical world. All other things being equal, an agent will behave as though this hypothetical model provides an accurate model of the real world. As new information about the world is acquired, the hypothetical model is changed in order to minimize the difference between the actual state of the world and the believed state of world.

There are two natural situations when an agent's beliefs should change:

1. The agent receives new information about a change in the state of the world.

2. The agent receives new information about an unchanged world.

The belief change that occurs is different in each case. In case (1), an agent basically wants the believed state of the world to keep up with the change that has occurred. As such, the change that occurred in the real world must also be made in the hypothetical model

of the world. In case (2), the hypothetical model needs to be revised and replaced with a new model. The new model should intuitively describe the most plausible world that is consistent with the new information. We describe a simple example where the distinction between (1) and (2) is clear.

Example John and Mary always arrive together when there is a party, because they come in the same car. However, John tends to leave earlier than Mary because he works mornings. Now suppose that it is early in the evening at one particular party, and Bill believes that everyone is already there. When looking for John, consider two things that Bill could be told:

1. "John just left."
2. "John is not here yet."

In both cases, Bill will come to believe that John is not at the party. Bill's beliefs about Mary, however, are different in each case. In case (1), Bill has become aware of a change in the world. In this case, Bill should change his beliefs to incorporate that fact that John has left; none of Bill's beliefs about Mary should change. As such, Bill should still believe that Mary is at the party. In case (2), on the other hand, Bill has become aware of a mistaken belief: Bill mistakenly believed that John had already arrived. After Bill is told that John hasn't arrived, this belief should be retracted. Since Mary and John normally arrive together, the most plausible thing for Bill to believe is that Mary has not yet arrived either. Therefore, Bill's final beliefs in case (2) will not be the same as Bill's final beliefs in case (1).

The belief change that occurs in response to a change in the world is called *belief update.* The belief change that occurs in response to new information about a static world is called *belief revision.* The party example illustrates that belief update and belief revision are distinct phenomena. Later in this chapter, we illustrate the standard formal approaches to update and revision; for now, we continue at an informal level.

Thus far, we have discussed belief change in an idealized setting using the undefined notion of "acquiring information." In practice, agents acquire information by performing actions. We consider two distinct kinds of actions: *ontic actions* and *epistemic actions.* Ontic actions are actions that change the state of the world, such as moving a block or

turning on a lamp. Epistemic actions are actions that only affect the beliefs of an agent, such as looking out a window or listening to the radio. The prototypical example of an epistemic action is a *sensing action*, or an *observation.*

In terms of belief change, ontic actions cause an agent to perform belief update and epistemic actions cause an agent to perform belief revision. However, in order to update and revise appropriately, an agent must reason about the effects of actions. Reasoning about the effects of actions is one of the oldest fields of logical AI, with formal approaches dating back to McCarthy's early work [71]. Many different formalisms have been proposed for the representation of action effects, because there are several fundamental problems that are difficult to solve in a completely satisfactory manner. Informally, the problem is the following. We typically think of action effects in terms of statements of this form:

"Performing a *jump* action causes my feet to leave the ground."

Note that some important features of jumping are left unstated in this simple description. For example, jumping causes my shoes to leave the ground but it does not cause my house to leave the ground. Although these inferences are easy for a human reasoner, it is non-trivial to efficiently formalize exactly which properties of the world change when an action is executed.

Without introducing a specific formal representation of actions, we remark simply that actions normally have conditional effects. For example, toggling a lamp switch causes the lamp to turn on just in case the lamp is initially off. So reasoning about the belief change caused by an action involves two steps. First, an agent uses the current belief state to predict the effect of an action that has been executed. Second, the agent updates or revises the current belief state based on this predicted effect. Note that, if the current belief state is incorrect, then the predicted effect of an action may also be incorrect.

As stated initially, we are interested in the belief change associated with sequences of actions and observations. New problems arise in this context that can not simply be dealt with by iteratively modifying an agent's beliefs as actions are executed. Suppose, for example, that John turns off the headlights on his car when he parks at work, but then someone later tells him that his headlights are still on. In this case, either John failed to turn the lights off or he has received faulty information; we cannot determine the appropriate belief change based on a purely iterative approach.

The basic problem that we address in this dissertation can be summarized as follows.

Given some initial beliefs followed by a sequence of ontic and epistemic actions, what should an agent believe? At an informal level, this problem arises in a wide range of situations. For example, this kind of reasoning is at the heart of the typical scientific experiment: an agent has some hypothesis (initial beliefs), runs an experiment (ontic action), then observes the results (epistemic action). The interpretation of the results depends on the preceding ontic action. Clearly it would be valuable to have an accurate formal model of this kind of reasoning. We will see that existing approaches to belief change have not explicitly addressed problems of this form.

To simplify the exposition in subsequent sections, we restrict the use of the term *action* to refer to only to ontic actions. We will use the terms *epistemic action* and *observation* interchangeably.

1.1.2 Motivating Example

In this section, we introduce a problem that will serve as a running example throughout the rest of the dissertation. In particular, we introduce Moore's litmus paper problem [74]. This problem is of interest not only because it involves ontic actions and observations, but also because it presents a challenge for the standard approach to belief update. In particular, Boutilier[11] suggests that the usual approach to belief update does not provide an appropriate model of the belief change that occurs in this problem. We give an informal description of the problem.

In the litmus paper problem, there is a beaker containing either an acid or a base, and there is an agent holding a piece of litmus paper that can be dipped into the beaker to determine the contents. The litmus paper will turn red if it is placed in an acid and it will turn blue if it is placed in a base. The problem is to provide a formal model of the belief change that occurs when an agent uses the litmus paper to test the contents of the beaker. We assume that the agent correctly believes that they are holding litmus paper, the agent correctly believes that the beaker contains either an acid or a base, and the agent only makes correct observations.

Note that the litmus paper problem involves an ontic action with conditional effects followed by an observation. The agent dips the paper in the beaker, and this causes a change in the state the of the world. After dipping, the agent looks at the paper to see what colour it is. We remark that, in isolation, looking at the paper should not indicate the contents of the beaker. Looking at the paper only gives the agent an indication of the

contents of the beaker because the agent is aware that the paper was previously dipped in the beaker.

Intuitively, the litmus paper problem seems to require an agent to revise the initial beliefs in response to an observation at a later point in time. For example, suppose that the agent dips the paper and then sees that the paper turns red. This observation not only causes the agent to believe the beaker contains an acid *now*, but it also causes the agent to believe that the beaker contained an acid *before dipping*. We refer to this process as a *prior revision*, since the agent appears to revise their beliefs at a prior point in time. This kind of phenomenon is not explicitly discussed in many formalisms for reasoning about belief change caused by action.

We will return to this problem periodically as we introduce our formal approach to belief change.

1.1.3 Relation to Existing Work

Our work is distinguished from existing work in that we explicitly formalize the high-level interaction between actions and observations in iterated belief change. The prototypical problem that we address has the following form.

$$(\mathit{InitialBeliefs}) \cdot (\mathit{Action}) \cdot (\mathit{Observation}) \cdots (\mathit{Action}) \cdot (\mathit{Observation}) \tag{1.1}$$

We are interested in determining how a rational agent's beliefs should change in response to such action sequences. The standard approach to representing this kind of problem is to start with a formalism for reasoning about action, and then add some formal approach to belief revision. We suggest that such an approach is a reasonable start, but it does not provide a complete representation.

We say that action effects are *Markovian* if the outcome of an action depends only on the current state of the world. The fundamental observation underlying our work is that actions with Markovian effects may plausibly give rise to non-Markovian belief change. This is the case, for example, in the litmus paper problem: the result of dipping depends only on the current contents of the beaker, and the result of looking depends only on the current colour of the paper. In this case, the naive approach to solving a problem of the form (1.1) is to treat iterated sequences of actions and observations by successively determining the effects of each event. However, this iterative approach does not permit the agent in the litmus paper problem to revise their initial beliefs after dipping. This is not just a superficial

conflict with our intuitions; we will see that successively applying updates and revisions can lead to results that are plainly incorrect.

This brief discussion illustrates that there is something missing from existing approaches to belief change caused by action. In order to understand problems of the form (1.1), we need to specify how iterated belief change can be compositionally defined in terms of the effects of individual actions and observations. That is the main problem that is addressed in this dissertation, and it is not explicitly addressed in existing work. In the most general case, our work is also distinguished from existing work in that we explicitly consider the reliability of an agent's perceived action history and perceived state of the world.

Our results can be understood either prescriptively or descriptively. Prescriptively, our work can be seen as providing a recipe for combining an action formalism with a belief revision operator. Descriptively, we can use our results to evaluate existing formalisms for the representation of epistemic action effects.

In the following sections, we present a summary of related work. Reasoning about epistemic action effects combines two established areas of enquiry: reasoning about action and belief change. As such, after presenting some logical preliminaries, we need to briefly introduce some of the most influential formal approaches to each of these areas. We then present some existing formalisms for reasoning about the epistemic effects of action. We conclude this chapter with an outline of the rest of the dissertation.

1.2 Logical Preliminaries

1.2.1 Propositional Logic and Predicate Logic

We assume the reader is familiar with classical propositional logic and predicate logic. In this section, we quickly outline some of the terminology and notation that we will adopt.

We assume an infinite set of *atomic propositional symbols*, and we define a *propositional signature* to be a set of atomic propositional symbols. We use the primitive propositional connectives $\{\neg, \rightarrow\}$, where $\neg$ denotes classical negation and $\rightarrow$ denotes implication. Conjunction, disjunction and equivalence are defined in the usual manner, and they are denoted by $\wedge$, $\vee$ and $\equiv$, respectively. A *formula* is a propositional combination of atomic symbols. A *literal* is either an atomic propositional symbol, or an atomic propositional symbol preceded by the negation symbol. Let *Lits* denote the set of all literals over a fixed signature.

An *interpretation* of a propositional signature $\mathbf{P}$ is a function that assigns every atomic

symbol a truth value. We will normally identify an interpretation I with the subset of atomic symbols that are true in I. The set of all interpretations over $\mathbf{P}$ is denoted by $2^{\mathbf{P}}$. The satisfaction relation $I \models \phi$ is defined for formulas ϕ by the usual recursive definition. For any formula ϕ, we define $|\phi|$ to be the set of all interpretations I such that $I \models \phi$, and we say that ϕ is satisfiable if and only if $|\phi| \neq \emptyset$.

1.2.2 Modal Logic

Propositional modal logic extends propositional logic by introducing an operator $\Box$ on formulas. We restrict attention to modal logics with a single unary modal operator $\Box$. Formulas are defined recursively just as they are defined in propositional logic, with the additional clause that $\Box\phi$ is a formula whenever ϕ is a formula. Intuitively, we read $\Box\phi$ as "ϕ is necessarily true." For a detailed introduction to modal logic, the reader is referred to [17]. In this section, we sketch the key definitions.

The usual semantics of modal logic is defined with respect to Kripke structures. A Kripke structure is a triple $\mathcal{M} = \langle M, R, \pi \rangle$, where M is a non-empty set of states (or worlds), R is a binary accessibility relation on M and π associates a subset of M with every atomic formula. We remark that π defines a propositional interpretation at each state m. We will let π_m denote the interpretation that π defines at m. The satisfaction relation $\models$ for modal logic indicates if a formula ϕ is true at a state m in a Kripke structure $\mathcal{M}$, written $\mathcal{M}, m \models \phi$. The relation is defined recursively, with modal formulas handled by a clause stating that $\mathcal{M}, m \models \Box\phi$ if and only if $\mathcal{M}, m' \models \phi$ for each m' such that Rmm'. In other words, $\Box\phi$ holds at m if and only if ϕ holds at every world accessible from m. We omit the mention of $\mathcal{M}$ if it is clear from the context.

A *system of modal logic* is a set of modal formulas that is closed under propositional consequence. Many important systems of modal logic can be defined by placing natural restrictions on the accessibility relation. In particular, *standard epistemic logic* is defined by restricting attention to Kripke structures where the accessibility relation is an equivalence relation. This logic is known as $KT5$. Intuitively, we think of the accessibility relation as an indistinguishability relation; two states are indistinguishable if the underlying agent is unable to tell them apart. A formula is known to be true just in case it is true in every state that is indistinguishable from the actual state. The fact that the accessibility relation is reflexive ensures that everything that is known must be true in the actual state. So standard epistemic logic has the property that the formula $\Box\phi \rightarrow \phi$ holds in every state, for every

formula ϕ. This axiom is known as T, and it captures the difference between logic and belief in modal logic.

The modal approach to belief is captured by *standard doxastic logic*, which is the modal logic $KD45$ defined by restricting attention to Kripke structures where the accessibility relation is serial, transitive, and euclidean. Since the accessibility relation need not be reflexive, the formulas that are believed need not be true in the actual state.

We remark that the term *doxastic* is not often used in the AI literature; it is common to use the term *epistemic* to refer to both knowledge and belief. It is still useful to use the terms to distinguish between the modal logics described above. However, we will use terms like *epistemic change* or *epistemic effects* to refer to changes in both knowledge and belief.

1.3 Reasoning about Action

1.3.1 Overview

Broadly speaking, reasoning about action is the branch of logic-based AI which is concerned with modeling how agents draw conclusions after executing actions that modify the state of the world. Many formalisms have been proposed for representing and reasoning about different kinds of action effects. One of the main applications of an action formalism is to solve *planning problems*. A planning problem is a problem in which an agent is given some goal, and then the agent is asked to find a sequence of actions that will achieve the goal. In order to solve planning problems, an action formalism must efficiently specify which aspects of the world change when an action is executed. A large portion of work in reasoning about action has been guided by three classic problems: the frame problem, the ramification problem, and the qualification problem. We briefly describe each problem.

Typically, an action will affect a specific property of the world, and it will leave everything else unchanged. For example, walking to the store in Vancouver does not change the weather in Halifax. The frame problem is the problem of providing an efficient mechanism for specifying all of the properties of the world that do not change when an action is executed.

The most salient effects of an action are the direct effects. For example, walking to the store causes an agent to be located at the store; this is the direct effect of the act of walking. However, there are also numerous indirect effects associated with such an action. For example, walking to the store will also cause an agent's pants to be located at the store. The ramification problem is concerned with the efficient specification of the indirect effects

of actions.

The qualification problem is the problem of specifying the preconditions for action execution. Returning to the same example, in order to walk to the store, an agent must assure that the door is open and that the store is nearby. Giving a complete specification of all qualifications in an efficient manner can be difficult.

Note that all of the classic problems are concerned with efficiently specifying some aspect of the effects of actions. Formalisms for reasoning about action vary greatly in the treatment of the classic problems. In the following sections, we briefly outline two popular approaches to reasoning about action.

1.3.2 Action Description Languages and Transition Systems

Action description languages are simple formal languages that are used to describe transition systems. In this section, we introduce action description languages and we look at a prototypical example. First, we introduce some notation from [34].

An *action signature* is a triple $\langle \mathbf{A}, \mathbf{F}, \mathbf{V} \rangle$ where $\mathbf{A}, \mathbf{F}, \mathbf{V}$ are non-empty sets of symbols. We call $\mathbf{A}$ the set of *action names*, we call $\mathbf{F}$ the set of *fluent names* and we call $\mathbf{V}$ the set of *values*. Informally, the fluent symbols in $\mathbf{F}$ denote properties of the world that assume the values in $\mathbf{V}$. The action symbols in $\mathbf{A}$ denote the actions that an agent may perform.

An action description language specifies how the state of the world changes when an action is executed. Formally, the semantics of an action description language relies on the notion of a transition system.

Definition 1 *A transition system T for an action signature $\sigma = \langle \mathbf{A}, \mathbf{F}, \mathbf{V} \rangle$ is a triple $\langle S, V, R \rangle$ where*

1. *S is a non-empty set,*
2. *$V : \mathbf{F} \times S \rightarrow \mathbf{V}$, and*
3. *$R \subseteq S \times \mathbf{A} \times S$.*

The set S is called the set of *states*, V is called the *valuation function* and R is the *transition relation*. If $F \in \mathbf{F}$ and $s \in S$, then $V(F, s)$ is the value of the fluent F in the state s. If $(s, A, s') \in R$, then we think of the state s' as a possible resulting state that could occur if the action A is executed in state s.

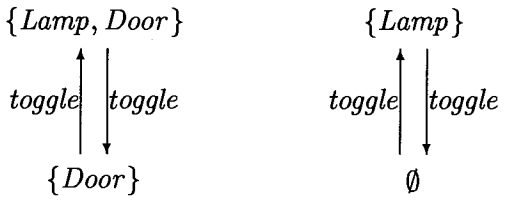

Figure 1.1: A Transition System

Action signatures where $\mathbf{V}$ is the set of propositional truth values $\{t, f\}$ are called *propositional action signatures.* Throughout this dissertation, we will only be concerned with propositional action signatures. As such, we will specify an action signature by a pair $\langle \mathbf{A}, \mathbf{F} \rangle$, and we leave the set of values $\{t, f\}$ implicit. Moreover, we refer to fluent symbols as being *true* or *false* if they are assigned the values t or f, respectively.

Note that the effects of actions in a transition system are explicitly Markovian. As such, transition systems can be visualized as directed graphs, where each node is labeled with a state and each edge is labeled with an element of R. We illustrate with an example.

Example Let $\sigma_L = \langle \{toggle\}, \{Lamp, Door\} \rangle$. Intuitively, the action signature σ_L is intended to describe a world with a lamp and a door, where the only available action is to toggle the switch on the lamp. The fluent symbol *Lamp* is true just in case the lamp is on, and the fluent symbol *Door* is true just in case the door is open. We will create a transition system for σ_L that represents the effects of actions in this simple world.

Let S be the set of propositional interpretations of $\{Lamp, Door\}$. We identify each element of S with the set of propositions that are true in S. Given $s \in S$ and $F \in \{Lamp, Door\}$, let $s(F)$ be the truth value that s assigns to F. Define $V : \mathbf{F} \times S \rightarrow \mathbf{V}$ such that $V(F, s) = s(F)$. Finally, define R as follows:

$$(s, toggle, s') \in R \iff s(Lamp) \neq s'(Lamp).$$

Hence, whenever the light switch is toggled, the value of the fluent symbol *Lamp* is changed. The graph representation of the transition system $T = \langle S, V, R \rangle$ is given in Figure 1.1.

One important example of an action description language is the action description language $\mathcal{A}$ [33, 34], which we define presently. Let $\sigma = \langle \mathbf{A}, \mathbf{F} \rangle$ be a propositional action

signature.

Definition 2 *A proposition of the language $\mathcal{A}$ is an expression of the form:*

$$A \textbf{ causes } L \textbf{ if } F_1 \wedge \cdots \wedge F_n$$

where $A \in \mathbf{A}$, $L \in Lits$, and each $F_i \in Lits$. A set of propositions is called an action description.

Definition 2 gives the syntax of the action description language $\mathcal{A}$. The semantics is defined by associating a transition system with every action description.

Definition 3 *Let AD be an action description in $\mathcal{A}$. The transition system $\langle S, V, R\rangle$ defined by AD is given by the following conditions:*

1. *S is the set of propositional interpretations of the symbols in $\mathbf{F}$*
2. *$V(F, s)$ is the value assigned to F by the interpretation s*
3. *R is defined as follows:*
   - *let $E(A, s)$ be the set of literals such that $L \in E(A, s)$ if and only if*

     $$(A \textbf{ causes } L \textbf{ if } F_1 \wedge \cdots \wedge F_n) \in AD$$

     and every literal F_i holds in s
   - *R is the set of all triples (s, A, s') with*

     $$E(A, s) \subseteq s' \subseteq E(A, s) \cup s.$$

Intuitively, R is the set of triples (s, A, s') such that s' is obtained from s by applying all of the relevant propositions in AD, and making no other changes to the fluent values.

Example Let σ_L be the action signature from the previous example. Let AD be the following action description:

$$\{toggle \textbf{ causes } Lamp \textbf{ if } \neg Lamp,$$
$$toggle \textbf{ causes } \neg Lamp \textbf{ if } Lamp\}$$

The transition system associated with AD is the transition system given in Figure 1.1.

The action description language $\mathcal{A}$ is a useful formalism for reasoning about the direct effects of actions. The syntax is extremely simple, and it allows us to describe any transition system where concurrent actions are not permitted. The frame problem for $\mathcal{A}$ is solved in the semantics by specifying that no fluents change value when an action is executed, except for those that are specifically mentioned in the proposition describing the action effect. The ramification problem is not addressed in $\mathcal{A}$; actions can only have direct effects. The qualification problem is trivially solved by allowing all actions to be executable in all states.

Some additional action description languages are provided in [34] and [64]. One action description language that has been particularly influential is $\mathcal{C}$. There are several ways in which $\mathcal{C}$ is more expressive than $\mathcal{A}$. For example, in $\mathcal{C}$ it is possible to represent concurrent action executions and static causal dependencies. More recently, the action language $\mathcal{K}$ has been introduced to represent action domains where we do not have complete information about the state of the world [24]. The action language $\mathcal{K}$ is reminiscent of $\mathcal{C}$ with the addition of negation as failure.

In all cases, an action description language provides a formal definition of a proposition, and it associates a transition system with every set of propositions. We remark that, although many action description languages employ syntax that is superficially similar, there is no unifying formal structure. Each action description language defines a new set of propositions and a new semantics from scratch.

The main advantage of an action description language like $\mathcal{A}$ is that it provides compact, transparent descriptions of action effects. Moreover, $\mathcal{A}$ can be used to solve simple planning problems. A planning problem in $\mathcal{A}$ can be given by a triple $\langle AD, \Gamma_{init}, \Gamma_{goal} \rangle$ where AD is an action description and $\Gamma_{init}, \Gamma_{goal}$ are consistent conjunctions of literals. A solution to this planning problem is a sequence of actions $A_1, \ldots, A_n$ such that, for every state s such that $s \models \Gamma_{init}$, the state s' that results from starting at s and executing $A_1, \ldots, A_n$ has the property that $s' \models \Gamma_{goal}$. Solutions to planning problems in $\mathcal{A}$ can be found by translating action descriptions into *extended logic programs*. We briefly sketch the idea.

An extended logic program is a collection of *rules*, which are expressions of the form

$$F \leftarrow G_1, \ldots, G_p, not\ N_1, \ldots, not\ N_q$$

where F is a literal, each G_i is a literal and each N_i is a literal. The symbol *not* represents negation as failure, so *not* G is true just in case G can not be proved. The answer set

semantics for logic programming associates a collection of *answer sets* with every extended logic program [32]. Informally, an answer set for a program P is a minimal set of atoms that is closed and grounded with respect to P. Action descriptions in $\mathcal{A}$ can be translated into extended logic programs where the answer sets correspond to paths in the described transition system [33]. By using existing answer set solvers, such as smodels [77] or dlv[24], one can automate the planning process. We remark that $\mathcal{A}$ is not the only action description language where action descriptions can be translated into extended logic programs. Many action description languages have been defined primarily as intermediary formalisms to facilitate the representation of planning domains for answer set planning [65].

1.3.3 Situation Calculus

For comparison, we introduce the Situation Calculus(SitCalc), another popular formalism for reasoning about action. This is a much more expressive formalism based on predicate logic. For a complete overview of the SitCalc, we refer the reader to [61].

The language of SitCalc is second-order logic with equality. It is a many sorted language with three sorts: *action*, *situation* and *object*. There is a distinguished constant symbol S_0 that denotes the *initial situation*. There is a distinguished function symbol *do* that takes an action and a situation as arguments, and returns a new situation. The situation

$$do(A_n, do(A_{n-1} \dots (do(A_1, S_0) \dots)$$

represents a world history in which the actions $A_1, \dots, A_n$ are executed, starting from the initial situation S_0. As a shorthand notation, we will also use the notation

$$do([A_1, \dots, A_n], S_0)$$

to represent the same situation.

A *fluent* is a predicate that takes a situation as one of its arguments. We remark that two distinct situations may satisfy the exact same set of fluents. However, two situations are equal just in case they represent the same sequence of actions. In this sense, there is an important distinction between the states in a transition system and the situations in a SitCalc representation.

We describe two important kinds of formulas that are used in a SitCalc representation of an action domain: successor state axioms and action precondition axioms. A successor

state axiom for a fluent symbol F is a formula of the form

$$F(\bar{x}, do(A, s)) \equiv \psi(\bar{x}, A, s)$$

where $\bar{x}$ is a tuple of free variables and ψ is a formula with free variables among $\bar{x}$, A, s. A successor state axiom gives a necessary and sufficient condition for a fluent symbol to be true after executing the action A in state s. An action precondition axiom is a formula of the form

$$Poss(A(\bar{x}), s) \equiv \psi(\bar{x}, s)$$

where $\bar{x}$ is again a tuple of free variables and ψ is a formula with free variables among $\bar{x}$ and s. The predicate $Poss$ is a distinguished predicate that is intended to express the conditions under which an action is executable. So an action precondition axiom gives necessary and sufficient conditions for an action to be executable.

A *basic action theory* in the SitCalc is axiomatized by the set of foundational axioms for situations, the set of unique names axioms for actions, a successor state axiom for each fluent symbol, and an action precondition axiom for each action symbol. We do not list the foundational axioms here, but we remark that it includes a second-order induction axiom which states that that any property that holds at S_0 and every *do*-successor of S_0 must hold at every situation. Given a basic action theory T, we can prove various properties of action effects by reasoning in predicate logic.

Note that it is easy to represent transition systems with basic action theories in the SitCalc, but the converse is not true. In particular, action effects in SitCalc are non-Markovian in the general case; the effects of an action depend on the current situation, which encodes the entire action history.

1.4 Belief Change

1.4.1 Belief Revision

As noted previously, the term belief revision refers to the process in which an agent receives new information about the world, and must incorporate this information with some prior beliefs. The underlying assumption is that the world does not change, and the new information is "better" than the initial beliefs. In this section, we briefly sketch one of the most influential approaches to belief revision: the AGM approach of Alchourrón, Gärdenfors and Makinson [3].

Let $\mathbf{F}$ be a propositional signature. A belief set is a deductively closed set of formulas over $\mathbf{F}$. In the AGM approach, belief revision is framed as the following problem. Given an initial belief set K along with a formula ϕ representing new information about the world, how should the new belief set be determined? The main intuition is that the new information given by ϕ must be incorporated, along with "as much of K" as consistently possible. Clearly, if ϕ is consistent with K, then the new belief set should be the deductive closure of $K \cup \{\phi\}$. If ϕ is not consistent with K, the problem is more difficult.

Define a *belief change operator* to be any function that maps a belief set and a formula to a new belief set. We would like to determine which belief change operators capture our intuitions about the process of belief revision. First, we remark that most belief change operators are not suitable. For example, let $+$ denote the so-called *belief expansion operator* $+$, which is defined by setting $K + \phi$ to be the deductive closure of $K \cup \{\phi\}$. In the case where K is consistent with ϕ, it seems that $+$ provides a reasonable account of belief revision. However, if either K or ϕ is inconsistent, then $K + \phi$ is inconsistent.

The AGM approach to belief revision does not provide a specific recipe for revising a belief set. Instead, a set of postulates is given, and any belief change operator that satisfies the postulates is called an AGM belief revision operator. Let $\mathcal{L}$ denote the set of all formulas over $\mathbf{F}$ and let $*$ denote a belief change operator. We say that $*$ is an AGM belief revision operator if it satisfies the following postulates, for every K and ϕ.

[AGM1] $K * \phi$ is deductively closed
[AGM2] $\phi \in K * \phi$
[AGM3] $K * \phi \subseteq K + \phi$
[AGM4] If $\neg\phi \notin K$, then $K + \phi \subseteq K * \phi$
[AGM5] $K * \phi = \mathcal{L}$ iff $\models \neg\phi$
[AGM6] If $\models \phi \equiv \psi$, then $K * \phi = K * \psi$
[AGM7] $K * (\phi \wedge \psi) \subseteq (K * \phi) + \psi$
[AGM8] If $\neg\psi \notin K * \phi$, then $(K * \phi) + \psi \subseteq K * (\phi \wedge \psi)$

The AGM postulates provide a simple set of conditions that are intuitively plausible as restrictions on belief revision operators. Moreover, the postulates completely determine a specific semantics for revision in terms of Grove's systems of spheres [39]. We now sketch Grove's characterization of AGM revision.

In the following definition, $M_{\mathcal{L}}$ denotes the set of consistent, complete theories over $\mathcal{L}$.

Definition 4 *A set of subsets $\mathcal{S}$ of $M_{\mathcal{L}}$ is a system of spheres centered on X where $X \subseteq M_{\mathcal{L}}$, if it satisfies the conditions:*

S1. $\mathcal{S}$ *is totally ordered by* $\subseteq$

S2. X *is the minimum of* $\mathcal{S}$ *under* $\subseteq$

S3. $M_{\mathcal{L}} \in \mathcal{S}$

S4. For any formula ϕ *with* $|\phi| \neq \emptyset$, *there is a least sphere* $c(\phi)$ *such that* $c(\phi) \cap |\phi| \neq \emptyset$ *and* $U \cap |\phi| \neq \emptyset$ *implies that* $c(\phi) \subseteq U$ *for every* $U \in \mathcal{S}$

We picture a system of spheres as a series of concentric circles, with innermost circle X.

A system of spheres provides a representation of the plausibility of theories over $\mathcal{L}$. The innermost theories are the most plausible, and they become successively less plausible as we move outwards. Under this ordering, the most natural way to incorporate a new formula ϕ is to try and determine the most plausible theory in which ϕ holds. This process associates a belief change operator $*$ with every system of spheres $\mathcal{S}$ as follows:

$$K *_{\mathcal{S}} \phi = \cap\{x \mid x \in |\phi| \cap c(\phi)\}.$$

Grove proves that this function is actually an AGM revision operator.

Proposition 1 (Grove) *If $\mathcal{S}$ is a system of spheres, then $*_{\mathcal{S}}$ satisfies the AGM postulates.*

The converse also holds. For fixed K, every AGM revision operator is determined by some system of spheres.

Proposition 2 (Grove) *Let $*$ be an AGM revision operator. For any belief set K, there is a system of spheres centered on K such that, for any ϕ,*

$$K * \phi = K *_{\mathcal{S}} \phi.$$

This result gives us an alternative perspective on AGM revision. In particular, it makes it clear that the process of AGM revision is always the same: every AGM revision operator relies on finding the most plausible theories satisfying the new information. Note that complete theories are essentially equivalent to propositional interpretations, so we can think of a

system of spheres as a total pre-order over interpretations. Informally, an agent performing AGM revision has some underlying notion of the plausibility of every possible world. When new information is encountered, the agent looks for the most plausible worlds in which the new information is true.

Grove's results demonstrate that AGM revision operators implicitly rely on a total pre-order over interpretations in the input. However, the belief set that results from AGM revision does not come with an attached ordering. Without a new ordering over states, it is not clear how the results of a second revision should be determined. Hence, AGM revision says nothing about *iterated revision*. In order to address iterated revision, Darwiche and Pearl propose a reformulation of AGM revision in which the orderings are explicit [19]. We briefly summarize this reformulated approach, which we call DP revision.

In DP revision, the beliefs of an agent are represented by an *epistemic state*. An epistemic state E consists of a total pre-order $\preceq_E$ over interpretations and a belief set $B(E)$ with the property that $|B(E)| = min(\preceq_E)$. A belief change operator in this context is a function that maps an epistemic state and a formula to a new epistemic state. Darwiche and Pearl reformulate the AGM postulates as follows.

[AGM1*] $B(E * \phi)$ is the deductively closed
[AGM2*] $\phi \in B(E * \phi)$
[AGM3*] $B(E * \phi) \subseteq B(E) + \phi$
[AGM4*] If $\neg\phi \notin B(E)$, then $B(E) + \phi \subseteq B(E * \phi)$
[AGM5*] $\bot \in B(E * \phi)$ iff $\models \neg\phi$
[AGM6*] If $\models \phi \equiv \psi$, then $E * \phi = E * \psi$
[AGM7*] $B(E * (\phi \wedge \psi)) \subseteq B(E * \phi) + \psi$
[AGM8*] If $\neg\psi \notin B(E * \phi)$, then $B(E * \phi) + \psi \subseteq B(E * (\phi \wedge \psi))$

For the most part, the reformulated postulates are obtained by replacing K with $B(E)$. The most important change occurs in AGM6*, which asserts that revision by equivalent formulas results not only in the same belief set, but in the same epistemic state. Darwiche and Pearl prove that belief change operators satisfying the reformulated postulates have the property that $|B(E * \phi)|$ must be equal to the set of $\preceq_E$-minimal models of ϕ. However, the reformulated AGM postulates do not fix the ordering over the non-minimal interpretations in $E*\phi$. For this purpose, Darwiche and Pearl introduce four new postulates.

[DP1] If $\phi \models \psi$, then $B(E * \psi * \phi) = B(E * \phi)$
[DP2] If $\phi \models \neg\psi$, then $B(E * \psi * \phi) = B(E * \phi)$
[DP3] If $\psi \in B(E * \phi)$, then $\psi \in B(E * \psi * \phi)$
[DP4] If $\neg\psi \notin B(E * \phi)$, then $\neg\psi \notin B(E * \psi * \phi)$

A DP revision operator is a belief change operator that satisfies AGM1*–AGM8* as well as DP1–DP4.

We remark that DP revision is just one proposal for the treatment of iterated revision. We have presented the DP approach here as an illustrative example, primarily because it builds directly on the AGM model. An alternative approach, based on a different set of postulates, is presented by Lehmann[59]. For our purposes, we will primarily be interested in action domains that have been supplemented with an AGM revision operator. However, we will be interested in ensuring that the postulates for iterated revision are satisfied where appropriate.

1.4.2 Belief Update

Belief update is the belief change that occurs when new information is acquired regarding a change in the state of the world. It is clear that our intuitions regarding belief change in this context differ from our intuitions in the context of belief revision. Informally, an agent performing belief revision is looking for a new belief set that describes the most plausible worlds supporting some new information. By contrast, an agent performing belief update is looking for a new belief set that captures the most plausible manner in which the world may have changed.

The standard approach to belief update is given by Katsuno and Mendelzon [51]. Following the AGM approach, Katsuno and Mendelzon give a set of postulates characterizing belief update. A belief change operator that satisfies all of the postulates is called a KM belief update operator. We remark that, in this case, we are assuming that ϕ represents some new information about the world that has just become true as the result of a change. The following reformulation of the KM postulates originally appeared in [85].

[KM1.] $K \diamond \phi$ is deductively closed
[KM2.] $\phi \in K \diamond \phi$

[KM3.] If $\phi \in K$, then $K \diamond \phi = K$

[KM4.] $K \diamond \phi = \mathcal{L}$ iff $K \models \bot$ or $\phi \models \bot$

[KM5.] If $\models \phi \equiv \psi$, then $K \diamond \phi = K \diamond \psi$

[KM6.] $K \diamond (\phi \wedge \psi) \subseteq (K \diamond \phi) + \psi$

[KM7.] If K is complete and $\neg\psi \notin K \diamond \phi$, then $(K \diamond \phi) + \psi \subseteq K \diamond (\phi \wedge \psi)$

[KM8.] If $|K| \neq \emptyset$, then $K \diamond \phi = \cap_{w \in |K|} w \diamond \phi$

The key postulate is KM8, which states that the updated belief set can equivalently be obtained by updating each model of K individually. Hence, an agent presented with the information ϕ proceeds by minimally modifying every possible initial state of the world to ensure that ϕ is true.

We have suggested that belief update and belief revision are distinct operations, but we have only provided informal arguments to support the claim. A formal distinction is established in [81], where it is shown that neither operation can be captured by the other. It is possible, however, to define a more general framework that incorporates both revision and update. One general framework that subsumes revision and update is Boutilier's *generalized belief update* framework [11]. Boutilier argues that a realistic characterization of belief change must combine elements of both revision and update. Briefly, the generalized update framework assumes that there is some fixed set E of *events*, and each event e is associated with a ranked set of possible outcomes. Moreover, there is a function μ that ranks the likelihood of each event occurring in any given state. When an agent finds out that some formula ϕ is true, then the agent tries to find a plausible state s, where there is a plausible event e with a plausible outcome o that explains ϕ. Boutilier proves that KM update and AGM revision can both be modeled in the generalized update framework. Although some of our methods will be informed by Boutilier's work, we will maintain an explicit distinction between revision and update in the remainder of this dissertation.

1.4.3 Belief Extrapolation

Dupin de Saint-Cyr and Lang argue that KM belief update is primarily suitable for applications where an agent is able to predict the manner in which the world changes over time. However, they suggest that KM update is not suitable for applications where the world changes in an unpredictable manner [23]. To capture belief change in domains where there can be unpredictable changes in the world, they introduce *belief extrapolation operators.*

In this section, we give a very brief introduction to belief extrapolation. We remark that belief extrapolation is a comparatively recent development, and it is not a standard tool like AGM revision or KM update. However, belief extrapolation provides a useful alternative approach to iterated belief change that is relevant to the fundamental problems addressed in this dissertation.

A belief extrapolation operator $\updownarrow$ takes a sequence of formulas, called a *scenario* as input, and it outputs another scenario. The intuition is that the output gives the most general sequence of formulas that can possibly be true, given the input and the assumption that fluents tend to be inertial. We give the basic construction.

A *trajectory* is a sequence τ of interpretations over some fixed signature. Let $\tau(i)$ denote the i^{th} interpretation in the trajectory τ. Given a scenario Σ, let $Traj(\Sigma)$ denote the set of trajectories that satisfy each formula in Σ on a point-by-point basis. Every ordering $\preceq$ on the class of trajectories defines an extrapolation operator $\updownarrow$ as follows:

$$|(\Sigma \updownarrow (t))| = \{\tau(t) \mid \tau \in Min(\preceq, Traj(\Sigma))\}$$

Hence, $\Sigma \updownarrow$ picks out the minimal trajectories satisfying Σ.

The most interesting belief extrapolation operators are given by orderings that are inertial. Informally, an ordering is inertial if static trajectories are always strictly less than non-static trajectories. Natural examples of inertial orderings can be given by ordering trajectories based on the number of fluents that change, or the number of elementary switches in fluent values.

1.5 Combining Reasoning about Action and Belief Change

We have introduced reasoning about action and belief change as two separate topics. However, it is clear that reasoning about the epistemic effects of actions involves a combination of both. There are four basic features that must be incorporated in a formalism for reasoning about the epistemic effects of actions [44].

1. Agents can perform ontic actions and epistemic actions.

2. Agents can perform belief update.

3. Agents can perform belief revision.

4. Perception and beliefs may be incorrect.

Existing approaches to modeling belief change often fall short with respect to one or more of these points. For example, [95] does not consider erroneous perception; [4] does not consider epistemic actions; and [35] does not consider belief revision. The problem is even more difficult if we consider action domains involving multiple agents, as done in both [95] and [44]. However, throughout this dissertation, we will restrict attention to the beliefs of a single agent.

As noted previously, one common approach to the representation of epistemic action effects is to start with an existing action formalism, and then add some mechanism for performing belief revision. In this section, we illustrate this approach by presenting epistemic extensions of $\mathcal{A}$ and the SitCalc.

1.5.1 Extensions of $\mathcal{A}$

Two epistemic extensions of $\mathcal{A}$ have been proposed in the literature [68, 86]. Originally, each of them was named $\mathcal{A}_K$. In order to reduce ambiguity, we refer to the extension of [68] as $\mathcal{A}_L$ and we refer to the extension of [86] as $\mathcal{A}_B$.

Assume that the action symbols in $\mathbf{A}$ are partitioned into sensing actions and non-sensing actions. In $\mathcal{A}_L$, there are two kinds of propositions. First, if A is a non-sensing action, $L \in Lits$ and each $F_i \in Lits$, then standard $\mathcal{A}$ propositions of the form

$$A \textbf{ causes } L \textbf{ if } F_1 \wedge \cdots \wedge F_n$$

are propositions of $\mathcal{A}_L$. If A is a sensing action, then

$$A \textbf{ causes to know } L \textbf{ if } F_1 \wedge \cdots \wedge F_n$$

is a proposition. Non-deterministic action effects are also introduced through a third proposition, but we will not concern ourselves with non-deterministic effects for the moment.

The semantics of $\mathcal{A}_L$ is defined with respect to *situations*, which are sets of states. The truth of a fluent symbol F with respect to a situation Σ is defined as follows.

- F is true in Σ if $s \models F$ for every $s \in \Sigma$.
- F is false in Σ if $s \models \neg F$ for every $s \in \Sigma$.
- F is unknown otherwise.

Truth or falsity in $\mathcal{A}_L$ is understood to reflect the knowledge of an agent, and knowledge is understood to be correct but not necessarily complete.

The semantics of $\mathcal{A}_L$ associates a transition relation Φ_{AD} with every action description AD. We give an informal definition. Let Σ be a situation and let A be an action symbol. The triple (Σ, A, Σ^*) is in Φ_{AD} if and only if Σ^* can be obtained from Σ as follows.

1. If A is non-sensing, then update each world in Σ in accordance with the $\mathcal{A}$ propositions in AD.

2. If A is sensing and L is unknown, then for each rule of the form

$$A \textbf{ causes to know } L \textbf{ if } F_1 \wedge \cdots \wedge F_n$$

Σ^* satisfies one of the following three conditions

(a) Σ^* is the set of situations in Σ where $F_1 \wedge \cdots \wedge F_n$ and L hold

(b) Σ^* is the set of situations in Σ where $F_1 \wedge \cdots \wedge F_n$ and $\neg L$ hold

(c) Σ^* is the set of situations in Σ where $\neg(F_1 \wedge \cdots \wedge F_n)$ holds

We illustrate the intuition behind the the effects of sensing actions with an example. Consider the proposition

$$\textit{Listen} \textbf{ causes to know } \textit{MusicOn} \textbf{ if } \neg \textit{EarPlugs}.$$

If an agent executes the action *Listen*, there are 3 possible outcomes: 1) the agent could learn that the music is on, 2) the agent could learn that the music is not on, or 3) the agent could learn neither. We assume that the only way the third possibility could arise is if the agent is wearing ear plugs. Hence, if the agent listens and still does not know if the music is on, then the agent still learns something. In particular, the agent learns the fact that the agent is wearing earplugs.

We remark that, given a pair (Σ, A), there will generally be several possible successor situations. A set of situations is called an *epistemic state*. Hence, the semantics of $\mathcal{A}_L$ actually maps an epistemic state and an action to a new epistemic state.

Our summary of $\mathcal{A}_B$ will be brief, due to the similarity with $\mathcal{A}_L$. The syntax of $\mathcal{A}_B$ introduces a new set of propositions of the form

$$A \textbf{ determines } F$$

where $A \in \mathbf{A}$ and $F \in \mathbf{F}$. The intended interpretation of such a proposition is that an agent will know the value of F after executing A.

The semantics of $\mathcal{A}_B$ is based on pairs $\langle s, \Sigma \rangle$, where s is a state and Σ is a set of states containing s. The state s represents the actual world, and Σ represents those worlds that are believed to be possible. Let AD be an action description. If A is a non-sensing action, $\Phi_{AD}(\langle s, \Sigma \rangle, A)$ is obtained by updating each world in Σ in accordance with the semantics of A. If A is a sensing action, and

$$A \textbf{ determines } F$$

is in AD, then $\Phi_{AD}(\langle s, \Sigma \rangle, A)$ is obtained by removing from Σ each world that differs from s in the interpretation of F.

1.5.2 The Situation Calculus Approach

The SitCalc has been extended to reason about knowledge by adding an accessibility relation K over situations [82]. The relation K is actually added to the syntax of the SitCalc, as opposed to being introduced as a modal operator in the logic. We think of K as an indistinguishability relation, so the underlying agent considers a situation to be possible just in case it is K-accessible from the actual situation. By convention, we read $Ks's$ as "s' is accessible from s." Note that this is the reverse of the usual reading in modal logic. For an atomic fluent formula $F(\bar{x}, s)$, we write $Knows(F(\bar{x}, s))$ as a shorthand for the formula

$$\forall s' Ks's \rightarrow F(\bar{x}, s').$$

Incomplete initial information is represented by introducing hypothetical alternatives to the initial situation. These alternatives can never occur, but they may be considered possible by an agent. The relation K is specified explicitly on the set of initial situations. However, in order to reason about belief change, we need to define a successor state axiom for K.

The epistemic SitCalc features two kinds of actions: sensing and non-sensing. Following [60], sensing actions are binary-valued, with effects given by a predicate $SF(A, s)$. To simplify the exposition, we assume that every sensing action A has a unique corresponding sensed fluent F_A. The effects of the sensing action A are given by an axiom of the form

$$SF(A, s) \equiv F_A(s).$$

For non-sensing actions, the axiom is always $SF(A,s) \equiv \top$. Given this new machinery, the successor state axiom for K is defined as follows.

$$\begin{aligned} K(s'', do(A,s)) \equiv\ & (\exists s'\ s'' = do(A,s') \\ & \wedge K(s',s) \wedge Poss(A,s') \\ & SF(A,s) \equiv SF(A,s')) \end{aligned}$$

This axiom states that s'' is accessible from $do(A,s)$ if and only if there is some state s' that is indistinguishable from s, and the sensing effects of A are identical in s and s'. If A is non-sensing, this means that K is copied from s to $do(A,s)$. If A is sensing, this means that performing A allows an agent to distinguish between states that differ in the sensing result associated with A.

The epistemic extension of the SitCalc that we have presented thus far is intended for action domains where the initial beliefs are necessarily correct. In the terminology of belief change, the semantics of epistemic actions is given by belief expansion rather than belief revision. However, the epistemic extension has been further modified to allow fallible beliefs and belief revision [85]. In the remainder of this chapter, we sketch the reformulated extension.

The main addition to the formalism is a function pl that assigns a natural number to each situation; we refer to the value $pl(s)$ as the *plausibility* of the situation s. The lower the plausibility assigned to a situation, the more plausible that situation is considered by the agent. The beliefs of an agent are represented by the set of maximally plausible situations among those that are accessible from the current situation. More precisely, belief is defined as follows:

$$Bel(\phi, s) \iff \forall s'[K(s',s) \wedge (\forall s'' K(s'',s) \rightarrow pl(s') \leq pl(s''))] \rightarrow \phi[s'].$$

The set of formulas believed in situation s is defined to be the set of formulas ϕ for which $Bel(\phi, s)$. Plausibility is defined for the set of initial situations, and then plausibility is restricted to persist after every action. As such, the plausibility of any situation is simply the plausibility of the corresponding initial situation.

Given a situation s, let κ_s denote the set of formulas that are true at every maximally plausible world accessible from s. Skipping over the details, if A is a sensing action that causes an agent to believe ϕ, we can define an operator $*_A$ as follows:

$$\kappa_s *_A \phi = \kappa_{do(A,s)}.$$

In the original presentation of the extension [85], this is called a revision operator, and it is proved that it satisfies 5 of the AGM postulates.

The introduction of plausibility values makes it possible for an agent to have erroneous beliefs about fluent values. One limitation of this variation of the SitCalc is the fact that agents still have perfect knowledge of the actions that have been executed. We remark that this limitation has been lifted in a more recent reformulation of the SitCalc in which exogenous actions may occur [84].

1.6 Our Approach

1.6.1 Overview

We are interested in iterated belief change that is caused by an alternating sequence of ontic actions and epistemic actions. As noted previously, the prototypical problem of interest has the following form.

$$(\mathit{InitialBeliefs}) \cdot (\mathit{Action}) \cdot (\mathit{Observation}) \cdots (\mathit{Action}) \cdot (\mathit{Observation})$$

In the standard approach to belief change, this corresponds to an expression of the form

$$K \diamond \phi_1 * \psi_1 \diamond \cdots \diamond \phi_n * \psi_n \tag{1.2}$$

where K is a belief set , $\diamond$ is an update operator and $*$ is a revision operator. From the perspective of reasoning about action, this representation is too simple; actions may have conditional effects that are not easily representable by a formula. It is straightforward to address conditional effects by reformulating the definition of belief update with respect to a particular action formalism. There is another problem with this approach, however, that has not been addressed to date. In particular, the interaction between ontic actions and epistemic actions is not explicitly considered; one is left to assume that each operator is applied successively. Our goal in this dissertation is to explicitly formalize, at a high level, the manner in which alternating sequences of actions and observations should be interpreted by a rational agent.

We make two underlying assumptions. First, we assume that the state of the world can be captured by a propositional interpretation over some fixed set $\mathbf{F}$ of fluent symbols. Second, we assume that the effects of the ontic actions $\mathbf{A}$ are given by a transition system. In this general setting, we consider iterated belief change caused by actions. We remark that

many existing action formalisms could be used to give the effects of ontic actions; we use transition systems primarily because they are conceptually simple, and they can easily be represented in a wide range of formalisms. Hence, using a transition system framework will make it possible to evaluate the treatment of iterated belief change in related formalisms.

We reformulate (1.2) in a form that is more appropriate for reasoning about actions in a transition system framework. In particular, we adopt the following conventions.

1. The beliefs of an agent are represented by a set of states κ.
2. The update operator $\diamond$ is a function $\diamond : 2^S \times \mathbf{A} \rightarrow 2^S$.
3. The revision operator $*$ is a function $* : 2^S \times 2^S \rightarrow 2^S$.

Note that new information for revision is represented by a set of states rather than a formula; it is easy to reformulate the AGM postulates to deal with sets of states. Note also that update is defined with respect to *actions* that have conditional effects given by the underlying transition system. Our prototypical problem can now be written as follows.

$$\kappa \diamond A_1 * \alpha_1 \diamond \cdots \diamond A_n * \alpha_n. \tag{1.3}$$

where each A_i is an action and each α_i is a set of states. There are natural examples of this form where it is clear that successively applying the updates and revisions leads to unintuitive results.

As an aside, we remark that an alternative to our approach would be to represent actions by their direct effects. In this manner, actions could be associated with sets of states just as we associate observations with sets of states. We have not taken this approach, because we want to explicitly reason about actions with conditional effects. For this purpose, a transition system provides a more natural representation.

In the body of this dissertation, we introduce a set of so-called *interaction properties* that should intuitively hold whenever an update is followed by a revision. We state the properties in the style of the AGM postulates, and we argue that the properties should hold in any action domain involving a single agent with perfect knowledge of the actions executed. We then define a new class of belief change operators, called *belief evolution* operators. A belief evolution operator takes two arguments: a set of states and an alternating sequence of actions and observations. Each belief evolution operator $\circ$ is defined with respect to a fixed

update operator $\diamond$ and a fixed AGM revision operator $*$. Informally, we have the following correspondence

$$\kappa \circ \langle A_1, \alpha_1, \ldots, A_n, \alpha_n \rangle \approx \kappa \diamond A_1 * \alpha_1 \diamond \cdots \alpha A_n * \alpha_n.$$

Under this interpretation, we will see that belief evolution satisfies our interaction properties. Moreover, we will prove that belief evolution can be formally characterized by a natural "shifting" on the underlying AGM revision operator. Relationships with existing formalisms will be discussed, and we will use belief evolution to define a new epistemic extension of $\mathcal{A}$.

We consider belief evolution in some detail, because it provides a straightforward mechanism for combining a given update operator with a given revision operator. The problem with belief evolution is that it requires the relatively strong assumption that the history of ontic actions is infallible. If an agent can be mistaken about the actions that have occurred, then our interaction postulates no longer hold. In order to address action histories that may be incorrect, we suggest that we need some additional information regarding the relative likelihood of each event in an alternating sequence of actions and observations.

Suppose that, at each point in time, the underlying agent has some beliefs about the action that has occurred, but these beliefs may be incorrect. In this case, we propose that Spohn-style ranking functions can be used to represent the agent's beliefs about the state of the world, as well as the agent's beliefs about the actions that occur at each point in time. The transition system T still gives the effects of the ontic actions, but a ranking function is used to capture the plausibility of a given action occurring at a given point in time. The most plausible world histories are determined by an aggregate function that combines the agent's beliefs about the actions that occur at each point in time. The resulting formalism provably subsumes belief evolution, and it is suitable for representing action domains where actions, beliefs, and observations are all fallible.

Note that, throughout the entire dissertation, we use an underlying transition system to give the effects of actions. The transition system could be replaced by some alternative description of action effects; the key point is that we need an action formalism to provide the set of admissible world histories. However, the action formalism does not play any role in determining which histories are the most plausible. The plausibility of a history is determined by formal machinery that is independent of the action formalism.

1.6.2 Contributions to Existing Research

As noted previously, there are plausible examples in which agents appear to revise a prior belief state in response to a new observation. In order to represent such problems, we need to explicitly consider belief change in the context of iterated actions and observations. To the best of our knowledge, our work is the first attempt at formally specifying any high-level interaction between belief update and belief revision caused by actions. In this section, we briefly discuss the contributions that we make in the general area of belief change caused by actions.

1. **Specifying Properties of Iterated Belief Change**
   We specify precise properties that should hold whenever an ontic action is followed by an epistemic action. The properties that we specify are a natural generalization of the AGM postulates; as such, they can easily be justified for action domains involving an a priori AGM operator. However, even if one disagrees with the given properties, explicitly considering iterated belief change is still valuable. It is clear in examples like the litmus paper problem that the action history can play a role in the interpretation of an observation. Therefore, it is useful to formalize the role of the action history in determining the appropriate belief change. However, this problem has not been explicitly addressed in related work.

2. **Evaluating Existing Formalisms**
   We evaluate the performance of some existing epistemic action formalisms with regards to iterated belief change. In particular, we consider the existing epistemic extensions of $\mathcal{A}$ as well as the epistemic extension of the SitCalc. It is easy to see that the extensions of $\mathcal{A}$ fail to satisfy our interaction properties, and they do not provide an accurate model of reasoning in litmus-type problems. On the other hand, we prove that the epistemic SitCalc satisfies our interaction properties. This result illustrates that it is possible for a formalism to handle iterated action effects appropriately, without explicitly considering our properties.

   In general, we illustrate that simply extending an existing action formalism with an AGM revision operator is not sufficient. We illustrate that such formalisms are either lacking the formal machinery required for reasoning about iterated belief change, or they are making substantive assumptions implicitly. Hence, we provide a tool that can

be used to determine if a given epistemic action formalism is suitable for reasoning about litmus-type problems.

Our work explicitly illustrates the role that an action formalism plays in the representation of belief change. The precise role of the transition system is clear and it is also clear which additional assumptions must be made in order to reason about iterated belief change. By making the role of the action formalism explicit, we can better evaluate the suitability of existing formalisms for particular applications.

3. **Combining Update and Revision**

   We give a specific methodology for combining an update operator and a revision operator in a single formalism. Informally, the idea is simply to translate all observations into conditions on the initial beliefs. In this manner, we can define an iterated belief change operator that respects our interaction properties and handles litmus-paper problems appropriately. Formally, our so-called belief evolution operators are appropriate because they can easily be characterized in terms of systems of spheres. In this sense, we can view iterated belief change as a modified form of revision.

   The belief evolution methodology is presented for action domains given by a transition system and an a priori AGM revision operator. To facilitate the introduction of examples, we present an extended class of transition systems with a distance function on states that defines an AGM revision operator. From a more general perspective, the belief evolution methodology is useful for combining any action formalism with an AGM revision operator. Hence, we can view belief evolution as an improved methodology for adding a revision operator to an action formalism. By using this methodology, we can avoid the problems that arise when ontic action effects and epistemic action effects are naively determined in succession.

4. **Applications**

   As an application of belief evolution, we define a new extension of $\mathcal{A}$. The new extension provably subsumes the existing epistemic extensions of $\mathcal{A}$ when we consider a single action, but it also satisfies our properties for iterated action effects. The extension itself improves upon existing epistemic action languages. It is particularly notable in that we illustrate how to implement a solver.

   In order to implement a solver, we define a belief revision operator based on path length

in a transition system. This operator is useful for reasoning about belief change in action domains where there is no underlying similarity relation on states. Under this revision operator, we can solve belief evolution problems by finding shortest paths. We illustrate that shortest paths can be found by translating action descriptions into extended logic programs, then finding answer sets. This approach is of interest because it can be implemented using existing answer set solvers, and there are relatively few existing implementations of belief change formalisms.

One new application that we consider is the use of our formal tools for the verification of cryptographic protocols. It is our hope that this application could benefit researchers in the belief change community by providing an interesting class of examples, and it could also benefit researchers in the security community by providing a more accurate model of belief change in cryptographic protocols.

5. **Reasoning with Fallible Action Histories**
   Existing epistemic action formalisms specify the effects of actions, but there is often no mechanism for dealing with uncertainty about the actions that have occurred. By using ranking functions to represent actions and observations, we illustrate a plausible representation of this kind of uncertainty. The resulting formalism is suitable for the representation of fallible beliefs, erroneous perception, exogenous actions and failed actions. Moreover, we explicitly address the manner in which prior action occurrences can be postulated or retracted in response to new observations. We are not aware of another action formalism that is able to simultaneously represent all of these phenomena.

   The use of ranking functions leads to a formalism that is superficially very different from belief evolution. However, we show that belief evolution can actually be seen as a special case of the more general formalism. The unifying feature of both approaches is a transition system giving the effects of ontic actions. This correspondence illustrates once again the role that is played by an action formalism in belief change; external notions of plausibility can be manipulated independently while keeping a fixed representation of action effects.

1.6.3 Outline

In chapter 2, we introduce the basic foundations of our formalism, and we fix our notation and terminology. We define a belief update operator based on transition systems, and demonstrate by example that this approach to belief update is superior to KM update for some interesting action domains. We also introduce an extended class of transition systems which is supplemented with a distance function on states. It is easy to define a belief revision operator given such a function; so this extended class of transition systems provides a simple formal tool in which both belief update and belief revision are possible. This provides a starting point for considering the interaction between belief revision and belief update.

In chapter 3, we look in detail at the interaction between revision and update. We restrict attention to action domains involving a single agent with perfect knowledge of the actions that have been executed, and we illustrate that there are action domains in which revising and updating iteratively does not provide intuitive results. The interaction between revision and update is discussed in the context of several desirable properties for iterated belief change. A new *belief evolution* operator is introduced based on some new rationality postulates, along with a representation result based on translated systems of spheres.

In chapter 4, we consider some applications of our formal tools for representing belief change. First, we introduce a flexible modal extension of the action language $\mathcal{A}$ with a semantics based on belief evolution. We demonstrate that existing epistemic extensions of $\mathcal{A}$ are subsumed by this approach. Second, we give some preliminary considerations on the implementation of a belief evolution solver. We illustrate that, in principle, existing answer set solvers may be used to determine the result of belief evolution under a fixed revision operator. The third application that we consider is the verification of cryptographic protocols. The idea is to encode the goals of the protocol as epistemic formulas, and prove that they must be true if certain conditions hold. We suggest that our formal approach is able to capture some subtle problems in protocol verification in a straightforward manner.

In chapter 5, we look at the more general problem in which an agent does not have perfect knowledge about the action history. This is the case, for example, in action domains where exogenous actions may occur. We present a new formalism that is based on Spohn-style ranking functions over actions and states. The new formalism requires ranking functions over actions and observations at each point in time, and it is expressive enough to represent fallible beliefs, erroneous perception, exogenous actions and failed actions. We prove that

the new formalism actually subsumes belief revision and belief evolution. A comparison with belief extrapolation concludes that each formalism has advantages and disadvantages.

In chapter 6, we offer some concluding remarks and ideas for future work.

Chapter 2

Transition Systems for Belief Change

In this chapter, we present a transition system framework for reasoning about belief change in the presence of actions. We focus on the effects of a single action or a single observation, leaving iterated actions for consideration in subsequent chapters. The effects of actions are given by a transition system that is known by the agent. We address two general problems. Given a transition system, how should the beliefs of an agent change following a single action? Similarly, how should the beliefs of an agent change following a single observation? As noted previously, the first problem basically involves belief update and the second problem involves belief revision. We define update and revision operators in a transition system framework.

First, we illustrate that every transition system defines a natural belief update operator where an agent's beliefs are updated by an action with conditional effects. Next, we define an *epistemic transition system* to be a transition system together with a similarity relation on states that defines an AGM revision operator. Hence, every epistemic transition system defines a belief update operator and a belief revision operator. We present a natural class of epistemic transition systems defined by extending a standard transition system with a distance function on states. The result is a single, graphically-motivated formalism for reasoning about epistemic action effects.

We remark that our goal in this chapter is not to present a sophisticated new formalism for reasoning about belief change. The update and revision operators presented are not

new, nor are they intended to capture a wide range of phenomena beyond that captured by traditional approaches to belief change. Instead, we are primarily interested in laying the groundwork for subsequent chapters. An epistemic transition system is a simple formalism for representing belief change in an action domain allowing both ontic actions and epistemic actions. We will use this basic framework to illustrate the problems that can arise if iterated sequences of actions and observations are handled in a naive manner.

2.1 Preliminaries

2.1.1 Notation and Terminology

Let $\langle \mathbf{A}, \mathbf{F} \rangle$ be an action signature. We assume that the action symbols in $\mathbf{A}$ represent ontic actions, with effects given by a transition system. A *belief state* is a set of states over $\mathbf{F}$. Informally, an agent with belief state κ believes that the actual world is represented by one of the interpretations in κ. We can think of a belief state as expressing a proposition. In addition to the ontic actions in $\mathbf{A}$, we also allow an agent to make *observations*. An observation is an epistemic action that provides the agent with new information about a static world. Formally, we define an observation to be a set of states. The intuition is that the observation α provides evidence that the actual world is in the set α. In terms of notation, the uppercase letter A, possibly with subscripts, will range over actions. The Greek letter α will range over observations and the Greek letter κ will range over belief states, again with possible subscripts in each case. We use the notation $\bar{A}$ to denote a finite sequence of action symbols of indeterminate length.

Note that we have defined the beliefs of an agent to be a set of states, rather than a set of formulas. Similarly, observed information is represented in terms of sets of states. As such, the revision operators that we have in mind differ superficially from AGM or DP revision operators. However, it is easy to translate the AGM and DP postulates into equivalent conditions on sets of states, and we will provide such translations when required. We remark that we will diverge further from the KM approach to belief update in that we will update belief states by *actions* rather than *formulas*. This is *not* an equivalent approach, but we will illustrate that it is still an acceptable approach from the perspective of the KM postulates.

2.1.2 Background Assumptions

We make two explicit assumptions about the action domains of interest.

1. The transition system describing action effects is both correct and complete.

2. The world is unchanging except for the changes caused by actions.

In addition, we place a closure condition on the transition systems that we will consider.

Definition 5 *A closed transition system is a transition system in which, for all* $s \in S$, $A \in \mathbf{A}$, *there exists some* $s' \in S$ *such that* $(s, A, s') \in R$.

Intuitively, closed transition systems represent action domains in which every action is always executable.

Given an arbitrary transition system $T = \langle S, R \rangle$, we define the closure $T' = \langle S, R' \rangle$ where

$$R' = R \cup \{(s, A, s) \mid (s, A, s') \notin R \text{ for any } s'\}.$$

Hence, the closure of T is obtained by replacing all non-executable actions with actions that do nothing. Graphically, for each action, this amounts to adding a self-loop at every node with no outgoing edge. Clearly, the closure of a transition system is a closed transition system.

Throughout this dissertation, we assume that all transition systems are closed. When presenting examples, we typically present a simplified transition system without self-loops. However, for formal results, we implicitly move to the closure. By considering only closed transition systems, we avoid the qualification problem.

2.2 Ontic Action Effects

2.2.1 Belief Update

In this section, we define belief update with respect to a transition system T. As noted earlier, we will define belief update operators that take a belief state and an action as arguments. So, to be more precise, the notion of belief update that we consider is actually a form of *action progression*. It has been argued elsewhere that the standard account of belief update can be understood to be a special case of this kind of progression [58]. The

advantage of our approach is that it provides a simple representation of the belief change that occurs following an action with conditional effects.

Intuitively, after executing an action A, an agent updates the belief state by projecting every state s to the state s' that would result if the action A was executed in the state s.

Definition 6 *Let $T = \langle S, R \rangle$ be a transition system. The update function $\diamond : 2^S \times \mathbf{A} \rightarrow 2^S$ is defined as follows*

$$\kappa \diamond A = \{s' \mid (s, A, s') \in R \text{ for some } s \in \kappa\}.$$

In order to compare our approach with the Katsuno and Mendelzon approach, we need to restrict attention to actions with constant effects. In the remainder of this section, we demonstrate that our belief update operators satisfy the Katsuno and Mendelzon postulates in this restricted case.

Let ϕ be a consistent conjunction of literals over a propositional language $\mathbf{F}$. Given a state s, there is a unique state $s(\phi)$ such that

- $s(\phi) \models \phi$
- if f is a fluent symbol not in ϕ, then $s(\phi) \models f$ iff $s \models f$.

Informally, the state $s(\phi)$ is obtained by minimally modifying s to ensure that ϕ is true.

We define a transition system T where every consistent conjunction of literals is the effect of some action. Let Γ be the set of all consistent conjunctions of literals over F, and let $\langle \Gamma, \mathbf{F} \rangle$ be the underlying action signature; note that conjunctions of literals are both formulas and action symbols in this context. Define T as follows:

- $S = 2^{\mathbf{F}}$
- $R = \{(s, \phi, s(\phi)) \mid s \in S, \phi \text{ a consistent conjunction of literals }\}$.

The transition system T is able to represent belief update by formulas, similar to the belief update operators of Katsuno and Mendelzon. The following proposition indicates that the associated update operator satisfies the appropriate rationality postulates.

Proposition 3 *Let $\mathbf{F}$ be a propositional signature, and let T be the transition system defined above. For consistent conjunctions of literals, the update operator obtained from T satisfies all of Katsuno and Mendelzon's postulates.*

Proof We rephrase the Katsuno and Mendelzon postulates in terms of belief states. Let κ denote a belief state and let ϕ, ψ denote consistent conjunctions of literals. If we translate the postulates to make the same assertions for sets of states rather than sets of formulas, we get the following.

[KM2*] $\kappa \diamond \phi \models \phi$

[KM3*] If $\kappa \models \phi$, then $\kappa \diamond \phi = \kappa$

[KM4*] $\kappa \diamond \phi = \emptyset$ iff $\kappa = \emptyset$

[KM5*] If $\models \phi \equiv \psi$, then $\kappa \diamond \phi = \kappa \diamond \psi$

[KM6*] If $\phi \wedge \psi$ is consistent, then $(\kappa \diamond \phi) \cap |\psi| \subseteq \kappa \diamond (\phi \wedge \psi)$

[KM8*] If $\kappa \neq \emptyset$, then $\kappa \diamond \phi = \cap_{w \in \kappa} w \diamond \phi$

Note that no translation of [KM1] is required, since every belief state corresponds to a deductively closed set of formulas. Moreover, since we are restricting attention to consistent formulas, we have collapsed [KM6] and [KM7] into a single postulate. Let $\diamond$ denote the update operator obtained from T. We demonstrate that $\diamond$ satisfies each postulate.

[KM2*]. Every state in $\kappa \diamond \phi$ is of the form $s(\phi)$ for some ϕ, and $s(\phi) \models \phi$ by definition.

[KM3*]. If $\kappa \models \phi$, then $s(\phi) = s$ for all $s \in \kappa$. In this case, it follows that $\kappa \diamond \phi = \kappa$.

[KM4*]. If $\kappa = \emptyset$, then $\kappa \diamond \phi = \emptyset$ by definition. If $\kappa \neq \emptyset$, then there exists some $s \in \kappa$ and it follows that $s(\phi) \in \kappa \diamond \phi$.

[KM5*]. For conjunctions of literals, $\models \phi \equiv \psi$ holds just in case ϕ and ψ contain the same positive and negative literals. It follows that $s(\phi) = s(\psi)$ for every state s, and hence $\kappa \diamond \phi = \kappa \diamond \psi$.

[KM6*]. Suppose $s' \in (\kappa \diamond \phi) \cap |\psi|$. So $s' = s(\phi)$ for some $s \in \kappa$ and $s(\phi) \models \psi$. But then $s(\phi) = s(\phi \wedge \psi)$, so $s' \in \kappa \diamond (\phi \wedge \psi)$.

[KM8*]. Follows immediately from Definition 6. □

Hence, if we restrict attention to non-conditional updates, then our notion of belief update defines a Katsuno-Mendelzon operator.

2.2.2 The Litmus Paper Problem Revisited

Defining update with respect to a transition system allows us to provide a better representation of the litmus paper problem. Before illustrating our approach, we briefly discuss Boutilier's objection to the Katsuno-Mendelzon representation of the litmus paper problem [11].

Figure 2.1: Litmus Test

Suppose that dipping the litmus paper indicates that the beaker actually contains an acid. The Katsuno-Mendelzon approach requires us to update each possible state of the world to reflect the change; this is essentially the content of postulate [KM8]. Note that the agent's initial belief state contains worlds where the liquid is a base. In such worlds, the contents of the beaker appear to be magically altered after dipping the litmus paper. Note that Boutilier's objection is not a formal objection to the end result obtained by a Katsuno-Mendelzon update operator; the objection is that the process in which each state is updated by a single formula is not appropriate.

Boutilier's objection is avoided if we define belief update with respect to a transition system and we separate the dipping action from the observation of the colour change. The litmus paper problem can be represented with the action signature

$$\langle \{dip\}, \{Red, Blue, Acid\} \rangle.$$

Intuitively, the fluent symbols *Red* and *Blue* represent the colour of the litmus paper, and the fluent symbol *Acid* indicates whether the beaker contains an acid or not. The only action available is to dip the litmus paper in the beaker; the effects of dipping are given by the transition system in Figure 2.1.

Initially, the agent believes the litmus paper is white, but it is not known whether the beaker contains an acid or a base. Hence, the initial belief state κ is the following:

$$\kappa = \{\emptyset, \{Acid\}\}.$$

After executing the *dip* action, the belief state is updated as follows:

$$\kappa \diamond dip = \{\{Blue\}, \{Red, Acid\}\}.$$

The set $\kappa \diamond dip$ consists of all possible outcomes of the *dip* action. To determine which outcome has occurred, the agent must perform an epistemic action. In particular, the agent must look at the paper to see what colour it is. We will consider the effects of epistemic actions in the next section.

Note that Boutilier's objection no longer applies: the contents of the beaker do not magically change after dipping the litmus paper. Instead, belief update simply projects every world forward, suitably modified by the effects of the dipping action. The change that occurs in each state is conditional on the fluents that are true. There is no uniform change that is made on a state by state basis following the dipping action. In fact, after dipping, the agent still does not know the contents of the beaker.

2.3 Epistemic Action Effects

2.3.1 Belief Revision

A standard transition system does not indicate how an agent should incorporate observed information. In some action domains, it may be reasonable to use a form of abductive reasoning in which an agent tries to find a sequence of exogenous actions that provides a justification for a new observation. However, this is not always appropriate: consider an observation which is not the effect of any action. In general, we need a similarity relation on states in order to reason about the effects of epistemic actions.

In subsequent chapters, we will normally assume that we have a fixed transition system T and a fixed AGM revision operator $*$. We remark, however, that it is not easy to give a compact representation of $*$ in terms of a similarity relation on states. In the general case, we need the following definition.

Definition 7 *An epistemic transition system T is a triple $\langle S, R, G\rangle$ where*

1. $S \subseteq 2^{\mathbf{F}}$
2. $R \subseteq S \times \mathbf{A} \times S$
3. *G is a function that maps every $X \subseteq S$ to a system of spheres centered on X*

By Grove's representation results, epistemic transition systems are precisely what we need if we want to capture every AGM revision operator in an extended transition system framework. However, we remark that this definition is not particularly insightful; it basically says that an epistemic transition system is a transition system together with an AGM revision operator.

One interesting class of epistemic transition systems can be defined by simply extending standard transition systems with a distance function on states. The notion of distance has a natural graphical interpretation and it provides a compact encoding of the function G. In the remainder of this section, we introduce distance-based revision in the context of transition systems.

A *metric* over $\mathbf{F}$ is a function that maps each pair of $\mathbf{F}$-states to a non-negative real number, and satisfies the following properties:

1. $d(w_1, w_2) = 0$ iff $w_1 = w_2$
2. $d(w_1, w_2) = d(w_2, w_1)$
3. $d(w_1, w_2) + d(w_2, w_3) \geq d(w_1, w_3)$

An *integral metric* is a metric that always takes integer values. We will only be concerned with integral metrics, so from here on we will use the term metric to refer only to integer valued metrics.

A metric transition system is simply a transition system along with a metric that defines a distance on states.

Definition 8 *A metric transition system T is a triple $\langle S, R, d\rangle$ where*

1. $S \subseteq 2^{\mathbf{F}}$
2. $R \subseteq S \times \mathbf{A} \times S$
3. *d is a metric on S*

Informally, if w_1 is close to w_2, then an agent will consider w_2 to be a plausible alternative to w_1.

As indicated previously, we identify observations with sets of states. With each metric transition system T, we associate a revision function. The revision function associated with T is the distance-based revision from [20].

Definition 9 *Let $T = \langle S, R, d\rangle$ be a metric transition system. The revision function $* : 2^S \times 2^S \rightarrow 2^S$ is defined as follows*

$$\begin{aligned}\kappa * \alpha \;\; = \;\; & \{w \in \alpha \mid \exists v_1 \in \kappa \text{ such that for all } v_2 \in \alpha, v_3 \in K \\ & \text{we have } d(w, v_1) \leq d(v_2, v_3)\}.\end{aligned}$$

Hence, if an agent is in belief state κ, then $\kappa * \phi$ is the set of all worlds in α that are minimally distant from some world in κ. We remark that the function d defines a system of spheres over every subset of states, so this revision function satisfies the AGM postulates.

We give some examples of metrics.

Example The Hamming distance d_{ham} is the metric defined such that $d_{ham}(w_1, w_2)$ is the number of fluent symbols assigned different truth values by w_1 and w_2. In this case, two states are similar to the degree that they assign the same values to fluent symbols. This metric provides a natural notion of similarity for action domains in which each fluent is given equal credence in determining plausible alternative worlds. Revision based on the Hamming distance metric has been explored previously by Dalal[18]; as a result, this particular revision operator is sometimes referred to as the *Dalal operator.* We remark also that the Hamming distance has been used in connection with belief update in [29].

Example Let $W : \mathbf{F} \rightarrow \mathbf{Z}^+$. The weighted Hamming distance d^W_{ham} is defined such that $d^W_{ham}(w_1, w_2)$ is the sum of the weights $W(f)$ for all fluent symbols f assigned different truth values by w_1 and w_2. Clearly, the Hamming distance is a special case of the weighted Hamming distance. However, the weighted version is convenient for action domains in which certain fluents are seen as more significant than others. For example, alternative worlds in which the weather is different may be more plausible than alternative worlds in which pigs can fly.

Example The topological distance function d_{top} is the metric defined such that $d_{top}(w_1, w_2)$ is the length of the shortest path from w_1 to w_2 in the underlying transition system. Note that this metric is only well defined for transition systems that are connected. The main advantage of the topological distance function is that the transition system itself defines the notion of similarity; no additional distance function is required. The notion of similarity

captured by the topological distance function is appropriate for action domains in which an agent is uncertain about the actions that have been executed previously. A state w_1 is understood to be similar to w_2 if a small number of actions can lead from w_1 to w_2.

We stress that our formal methods will not be limited to distance-based revision functions. We are primarily interested in adding a distance function because it provides a concrete formalism which defines both an update operator and a revision operator, and this is useful for describing simple examples. Hence, our interest in distance-based revision is largely pragmatic. Nevertheless, one could argue that the distance-based approach is particularly appropriate for our purposes. A function G mapping belief states to systems of spheres can be cumbersome to describe, and such functions may not have a natural graphical representation. By contrast, distance functions can be described succinctly, and the notion of distance has a natural graphical interpretation. Therefore, distance functions provide sufficient information to define a revision operator and they do so in a straightforward manner at the same level of abstraction of a standard transition system.

2.3.2 The Litmus Paper Problem Concluded

In order to complete the representation of the litmus paper problem, we need to consider the observation that the agent makes following the dipping action. For the moment, we simply revise the belief state that was obtained from the belief update. We will see in the next chapter that simply applying the action effects iteratively in this manner does not always lead to a desirable result.

Let d be a metric on the states in Figure 2.1. Recall that

$$\kappa \diamond dip = \{\{Blue\}, \{Red, Acid\}\}$$

and assume that the beaker actually contains an acid, so the litmus paper turns red after dipping. Hence, looking at the litmus paper causes the agent to revise the current belief state by the set of states where *Red* is true. More precisely, we need to revise by the observation α defined as follows

$$\alpha = \{\{Red, Acid\}, \{Red\}, \{Red, Blue, Acid\}, \{Red, Blue\}\}.$$

Note that the state $\{Red, Acid\}$ is in α and it is also in the prior belief state. It follows

immediately that the revised belief state is

$$\{\{Red, Acid\}\}.$$

We make some brief remarks about this example. First of all, note that this is clearly a case of belief expansion rather than true belief revision. As a result, we have not needed to specify a definite metric. We were able to determine the outcome of the revision for any metric, because a distance of zero is never obtained between distinct states. This is a general property of AGM revision: if we revise by consistent information, then the underlying ordering over states or formulas does not come into play.

Chapter 3

Iterated Epistemic Action Effects

In this chapter, we consider iterated belief change in the context of epistemic transition systems. We illustrate that there is a problem with the naive approach in which updates and revisions are applied successively. Roughly, if an agent has access to the history of actions that have been executed, then there are examples where it is plausible for an agent to revise the initial belief state rather than the current belief state. We introduce a new belief change operator for the representation of this kind of reasoning. The new operator intuitively captures the evolution of an agent's beliefs following a sequence of ontic actions and observations.

We briefly outline the rest of the chapter. In §3.1, we give a schematic view of a typical problem and we introduce an illustrative running example where the interaction between revision and update is non-elementary. In §3.2, we formally specify a set of properties describing the belief change that occurs when an update is followed by a revision. We introduce some simple notation for the representation of action histories in §3.3, and we introduce a new belief change operator in §3.4. Our new belief change operator implicitly defines an approach to iterated revision, so in §3.5 we consider the defined approach with respect to two well-known sets of postulates. In §3.6, we compare our approach to belief change with some existing approaches. We conclude in §3.7 by demonstrating that our new belief change operator can be characterized by a pair of rationality postulates, and we prove a representation result based on systems of spheres.

A preliminary version of the material in this chapter previously appeared in [46].

3.1 Motivation

3.1.1 The Basic Problem

Let $\diamond$ be an update operator and let $*$ be a revision operator. We are interested in giving a reasonable interpretation to sequences of the form

$$\kappa \diamond A_1 * \alpha_1 \diamond \cdots \diamond A_n * \alpha_n. \tag{3.1}$$

There are intuitively plausible examples in which applying the operators iteratively results in an unsatisfactory result. The essential point is that an observation at time n can lead an agent to revise the initial belief state, rather than the current belief state. This is particularly common in single-agent action domains in which the agent has complete knowledge of the action history.

The simplest interesting case is given by an expression of the form

$$\kappa \diamond A_1 \diamond \cdots \diamond A_n * \alpha. \tag{3.2}$$

In this case, since there is only a single observation, we can focus entirely on the interaction between revision and update. The general case represented by (3.1) is complicated by the fact that it implicitly requires some form of iterated revision. Our formalism handles the general case, but our initial focus will be on problems of the form in (3.2). In the next section, we consider an example of such a problem and we illustrate that the most natural solution requires the initial state to be revised at a later point in time.

3.1.2 An Illustrative Example

We extend the litmus paper problem. The extended problem is just like the original, except that we allow for the possibility that the paper is not litmus paper; it might simply be a piece of plain white paper. In order to represent this possibility, we introduce a new fluent symbol *Litmus*. Hence, we now have the fluent symbols $\mathbf{F} = \{Red, Blue, Acid, Litmus\}$. The set of action symbols still contains the single action *dip*.

Define the metric transition system $T = \langle S, R, d \rangle$ as follows.

1. $S = 2^{\mathbf{F}}$.

2. R is obtained by taking the closure of the transition system in Figure 3.1.

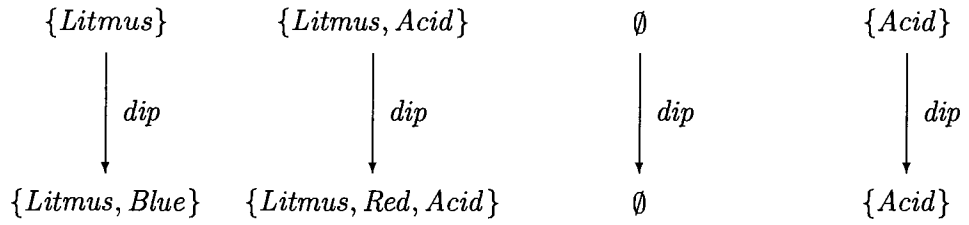

Figure 3.1: Extended Litmus Test

3. d is the Hamming distance.

This transition system implicitly makes the simplifying assumption that dipping does not do anything if the paper is already red or blue.

We describe a sequence of events informally. Initially, the agent believes that the paper is a piece of litmus paper, but the agent is unsure about the contents of the beaker. To test the contents, the agent dips the paper in the beaker. After dipping, the agent looks at the paper and observes that it is still white. We are interested in determining a plausible final belief state.

We now give a more formal representation of the problem. The initial belief state is

$$\kappa = \{\{Litmus\}, \{Litmus, Acid\}\}.$$

After dipping the paper in the beaker, we update the belief state as follows:

$$\kappa \diamond dip = \{\{Litmus, Blue\}, \{Litmus, Red, Acid\}\}.$$

At this point, the agent looks at the paper and sees that it is neither blue nor red. This observation is represented by the following set of worlds:

$$\alpha = \{\emptyset, \{Litmus\}, \{Acid\}, \{Litmus, Acid\}\}.$$

The naive suggestion is to simply revise $\kappa \diamond dip$ by α, which gives

$$\kappa' = \{\{Litmus\}, \{Litmus, Acid\}\}.$$

We claim that this is not a plausible final belief state.

Informally, if the paper is litmus paper, then it must be either red or blue after a dipping action is performed. Hence, neither $\{Litmus\}$ nor $\{Litmus, Acid\}$ is a plausible state after dipping; simply revising by the observation gives a belief state that is guaranteed to be incorrect. The final belief state should consist entirely of states that are possible consequences of dipping.

We suggest that a rational agent should reason as follows. After dipping the paper and seeing that it does not change colour, an agent should conclude that the paper was never litmus paper to begin with. The initial belief state should be modified to reflect this new belief *before* calculating the effects of the dipping action. This approach ensures that we will have a final belief state that is a possible outcome of dipping. At the end of the experiment, the agent should believe that the paper is not litmus paper and the agent should have no definite beliefs regarding the contents of the beaker. Hence, we propose that the most plausible final belief state is the set

$$\{\emptyset, \{Acid\}\}.$$

This simple example serves to illustrate the fact that it is sometimes useful for an agent to revise prior belief states in the face of new knowledge. In order to formalize this intuition in greater generality, we need to introduce some new formal machinery.

3.2 Interaction Between Revision and Update

In this section, we give a set of formal properties that we expect to hold when an update is followed by a revision. The properties are not overly restrictive and they do not provide a basis for a categorical semantics; they simply provide a point for discussion and comparison. Our underlying assumption is that ontic action histories are infallible. The most recent observation is always incorporated, provided that it is consistent with the history of actions that have been executed. Hence, the properties we discuss are only expected to hold in action domains in which there are no failed actions and no exogenous actions.

We briefly present some of our underlying intuitions. Let κ be a belief state, let $\bar{A}$ be a sequence of actions, and let α be an observation. We are interested in the situation where an agent has initial belief state κ, then $\bar{A}$ is executed, and then it is observed that the actual state must be in α. We adopt the shorthand notation $\kappa \diamond \bar{A}$ as an abbreviation for the sequential update of κ by each element of $\bar{A}$. There are three distinct cases to consider.

1. There are some α-states in $\kappa \diamond \bar{A}$.

2. There are no α-states are in $\kappa \diamond \bar{A}$, but some α-states are possible after executing $\bar{A}$.

3. No α-states are possible after executing $\bar{A}$.

We discuss each case separately.

Case (1) is the situation in which the observation α allows the agent to refine their knowledge of the world. After the observation α, the agent should believe that the most plausible states are the states in κ that are also in α. In other words, we propose that the agent should adopt the belief state $(\kappa \diamond \bar{A}) \cap \alpha$.

In case (2), the agent should conclude that the actual state was not initially in κ. This conclusion is based on our underlying assumption that the action sequence $\bar{A}$ cannot fail, and the additional assumption that a new observation should be incorporated whenever possible. Both of these assumptions can be satisfied by modifying the initial belief state before performing the update. Informally, we would like to modify the initial belief state minimally in a manner that ensures that α will be true after executing $\bar{A}$. This is the case that occurs in the extended litmus paper problem.

Case (3) is problematic, because it suggests that the agent has some incorrect information: either the observation α is incorrect or the sequence $\bar{A}$ is incorrect. For the moment, we are assuming that action histories are infallible, so the agent must abandon the observation α to remain consistent with $\bar{A}$.

Assume a fixed finite propositional signature F. Let κ and α be sets of worlds, let $\bar{A}$ be a sequence of actions, let $\diamond$ be an update operator, and let $*$ be a revision operator. We formalize our intuitions by suggesting that the following conditions should be satisfied when an update is followed by a revision.

Interaction Properties

P1. If $(2^F \diamond \bar{A}) \cap \alpha \neq \emptyset$, then $\kappa \diamond \bar{A} * \alpha \subseteq \alpha$

P2. If $(2^F \diamond \bar{A}) \cap \alpha = \emptyset$, then $\kappa \diamond \bar{A} * \alpha = \kappa \diamond \bar{A}$

P3. $(\kappa \diamond \bar{A}) \cap \alpha \subseteq \kappa \diamond \bar{A} * \alpha$

P4. If $(\kappa \diamond \bar{A}) \cap \alpha \neq \emptyset$, then $\kappa \diamond \bar{A} * \alpha \subseteq (\kappa \diamond \bar{A}) \cap \alpha$

P5. $\kappa \diamond \bar{A} * \alpha \subseteq 2^F \diamond \bar{A}$

We give some motivation for each property. P1 is a straightforward AGM-type assertion that α must hold after revising by α, provided α is possible after executing $\bar{A}$. P2 handles the situation where it is impossible to be in an α-world after executing $\bar{A}$. In this case, we simply discard the observation α. Together, P1 and P2 formalize the underlying assumption that there are no failed actions.

P3 and P4 assert that revising by α is equivalent to taking the intersection with α, provided the intersection is non-empty. These are similar to the AGM postulates asserting that revisions correspond to expansions, provided the observation is consistent with the knowledge base.

P5 provides the justification for revising prior belief states in the face of new knowledge. It asserts that, after revising by α, we must still have a belief state that is a possible consequence of executing $\bar{A}$. In some cases, the only way to ensure that α holds after executing $\bar{A}$ is to modify the initial belief state. We remark that P5 does not indicate how the initial belief state should be modified.

3.3 Representing Histories

Transition systems are only suitable for representing Markovian action effects. However, in the extended litmus paper problem, we saw that the outcome of a sensing action may depend on prior belief states. Even if ontic action effects are Markovian, it does not follow that changes in belief are Markovian. As such, we need to introduce some formal machinery for representing histories. We will be interested in the historical evolution of an agent's beliefs, along with all of the actions executed. Representing belief histories is straightforward.

Definition 10 *A belief trajectory of length n is an n-tuple*

$$\langle \kappa_0, \ldots, \kappa_{n-1} \rangle$$

of belief states.

Intuitively, a belief trajectory is an agent's subjective view of how the world has changed. We remark that a belief trajectory represents the agent's current beliefs about the world history, not a historical account of what an agent believed at each point in time. For example, in the extended litmus paper problem, at the end of the experiment the agent believes that they were never holding a piece of litmus paper. The fact that the agent once

believed that they were holding litmus paper is a different issue, one that is not represented in our formal conception of a belief trajectory.

We will also be interested in observation trajectories and action trajectories, each of which is simply another n-tuple.

Definition 11 *An observation trajectory of length n is an n-tuple $\bar{\alpha} = \langle \alpha_1, \ldots, \alpha_n \rangle$ where each $\alpha_i \in 2^S$.*

Each set α_i is interpreted to be evidence that the actual world is in α_i at time i.

Definition 12 *An action trajectory of length n is an n-tuple $\bar{A} = \langle A_1, \ldots, A_n \rangle$ where each $A_i \in \mathbf{A}$.*

An action trajectory is a history of the actions an agent has executed. Note that, as a matter of convention, we start the indices at 0 for belief trajectories and we start the indices at 1 for observation and action trajectories. The rationale for this convention will be clear later. We also adopt the convention hinted at in the definitions, whereby the i^{th} component of an observation trajectory $\bar{\alpha}$ will be denoted by α_i, and the i^{th} component of an action trajectory $\bar{A}$ will be denoted by A_i.

We define a notion of consistency between observation trajectories and action trajectories. The intuition is that an observation trajectory $\bar{\alpha}$ is consistent with an action trajectory $\bar{A}$ if and only if each observation α_i is possible, given that the actions $(A_j)_{j \leq i}$ have been executed.

Definition 13 *Let $\bar{\alpha} = \langle \alpha_1, \ldots, \alpha_n \rangle$ be an observation trajectory and let $\bar{A} = \langle A_1, \ldots, A_n \rangle$ be an action trajectory. We say that $\bar{A}$ is consistent with $\bar{\alpha}$ if and only if there is a belief trajectory $\langle \kappa_0, \ldots, \kappa_n \rangle$ such that, for all i with $1 \leq i \leq n$,*

1. $\kappa_i \subseteq \alpha_i$
2. $\kappa_i = \kappa_{i-1} \diamond A_{i-1}$.

If $\bar{A}$ is consistent with $\bar{\alpha}$, we write $\bar{A} || \bar{\alpha}$.

A pair consisting of an action trajectory and an observation trajectory gives a complete picture of an agent's view of the history of the world. As such, it is useful to introduce some terminology.

Definition 14 *A world view of length n is a pair $W = \langle \bar{A}, \bar{\alpha} \rangle$, where $\bar{\alpha}$ is an observation trajectory and $\bar{A}$ is an action trajectory, each of length n. We say W is consistent if $\bar{A} || \bar{\alpha}$.*

3.4 Belief Evolution

3.4.1 A New Belief Change Operator

We introduce a new operator $\circ$ that takes two arguments: a belief state and a world view. Roughly speaking, we would like $\kappa \circ \langle \bar{A}, \bar{\alpha} \rangle$ to be the belief trajectory that results from the initial belief state κ and the alternating action-observation sequence

$$\kappa \diamond A_1 * \alpha_1 \diamond \cdots \diamond A_n * \alpha_n.$$

We call $\circ$ a *belief evolution* operator because it takes a sequence of actions and returns the most plausible evolution of the world.

The formal definition of $\circ$ is presented in the following sections. The definition relies on a fixed revision operator $*$ and a fixed update operator $\diamond$. As such, it might be more accurate to adopt notation of the form $\circ_{*,\diamond}$, but we opt for the less cumbersome $\circ$ and assume that the underlying operators are clear from the context. It is worth noting that the definition of $\circ$ does not rely on any specific approach to revision. Given a transition system, any AGM revision operator can be used to define a belief evolution operator. Metric transition systems provide an important class of examples, because every finite metric transition system generates a unique belief evolution operator. However, in the interest of generality, we do not tie ourselves to a specific approach. We remark that we could actually define $\circ$ with respect to an arbitrary binary function on belief states; but for the applications that we consider, it is better to restrict $*$ to be an AGM revision operator.

The action domains of interest for belief evolution will be those in which it is reasonable to assume that action trajectories are correct and actions are successful. This is intuitively plausible in a single agent environment, because it simply amounts to assuming that an agent has complete knowledge about the actions that have been executed. Hence, in the definition of $\circ$, the belief trajectory returned will always be consistent with the actions that have been executed.

3.4.2 Infallible Observations

In this section, we assume that observations are always correct. Formally, this amounts to a restriction on the class of admissible world views. In particular, we need not consider inconsistent world views. It is easy to see that an inconsistent world view is not possible under the assumption that action histories and observations are both infallible.

We need to introduce some notation. In particular, let $s^{-1}(A)$ denote the set of all states s' such that $(s', A, s) \in R$. We call $s^{-1}(A)$ the *pre-image* of s with respect to A. The following definition generalizes this idea to give the pre-image of a set of states with respect to a sequence of actions. In the definition, given any sequence of actions $\bar{A} = \langle A_1, \ldots, A_n \rangle$, we write $s \rightsquigarrow_{\bar{A}} s'$ to indicate that there is a path from s to s' that follows the edges labeled by the actions $A_1, \ldots, A_n$.

Definition 15 *Let T be a deterministic transition system, let $\bar{A} = \langle A_1, \ldots, A_n \rangle$ and let α be an observation. Define $\alpha^{-1}(\bar{A}) = \{s \mid s \rightsquigarrow_{\bar{A}} s' \text{ for some } s' \in \alpha\}$.*

Hence, if the actual world is an element of α following the action sequence $\bar{A}$, then the initial state of the world must be in $\alpha^{-1}(\bar{A})$.

For illustrative purposes, it is useful to consider world views of length 1. Suppose we have an initial belief state κ, an ontic action A and an observation α. Without formally defining the belief evolution operator $\circ$, we can give an intuitive interpretation of an expression of the form

$$\kappa \circ \langle \langle A \rangle, \langle \alpha \rangle \rangle = \langle \kappa_0, \kappa_1 \rangle.$$

The agent knows that the actual world is in α at the final point in time, so we must have $\kappa_1 \subseteq \alpha$. Moreover, the agent should believe that κ_1 is a possible result of executing A from κ_0. In other words, we must have $\kappa_0 \subseteq \alpha^{-1}(A)$. All other things being equal, the agent would like to keep as much of κ as possible. In order to incorporate $\alpha^{-1}(A)$ while keeping as much of κ as possible, the agent should revise κ by $\alpha^{-1}(A)$. This suggests the following solution.

1. $\kappa_0 = \kappa * \alpha^{-1}(A)$,
2. $\kappa_1 = \kappa_0 \diamond A$.

This procedure can be applied to world views of length greater than 1. The idea is to trace every observation back to a precondition on the initial belief state. After revising the initial belief state by all preconditions, each subsequent belief state can be determined by a standard update operation.

We have the following formal definition for $\circ$. In the definition, if $i \leq n$ then we let $\bar{A}_i$ denote the subsequence of actions $\langle A_1, \ldots, A_i \rangle$.

Definition 16 *Let κ be a belief state, let $\bar{A}$ be an action trajectory of length n and let $\bar{\alpha}$ be an observation trajectory of length n such that $\bar{A}||\bar{\alpha}$. Define*

$$\kappa \circ \langle \bar{A}, \bar{\alpha} \rangle = \langle \kappa_0, \ldots, \kappa_n \rangle$$

where

1. $\kappa_0 = \kappa * \bigcap_i \alpha_i^{-1}(\bar{A}_i)$
2. *for $i \geq 1$, $\kappa_i = \kappa_{i-1} \diamond A_1 \diamond \cdots \diamond A_i$.*

We remark that the intersection of observation preconditions in the definition of κ_0 is non-empty, because $\bar{A}||\bar{\alpha}$.

The following propositions are immediate, and they demonstrate that for some action sequences of length 1, $\circ$ reduces to either revision or update. In each proposition, we assume that $\bar{A}||\bar{\alpha}$.

Proposition 4 *Let κ be a belief state, let $\bar{A} = \langle A \rangle$ and let $\bar{\alpha} = \langle 2^{\mathbf{F}} \rangle$. Then*

$$\kappa \circ \langle \bar{A}, \bar{\alpha} \rangle = \langle \kappa, \kappa \diamond A \rangle.$$

Proof Recall that we only allow closed transition systems, so every action is executable in every state. It follows that $(2^{\mathbf{F}})^{-1}(A) = 2^{\mathbf{F}}$. Therefore

$$\begin{aligned} \kappa \circ \langle \bar{A}, \bar{\alpha} \rangle &= \langle \kappa * 2^{\mathbf{F}}, (\kappa * 2^{\mathbf{F}}) \diamond A \rangle \\ &= \langle \kappa, \kappa \diamond A \rangle. \end{aligned}$$

□

In the following, we assume that λ is a null action that never changes the state of the world.

Proposition 5 *Let κ be a belief state, let $\bar{A} = \langle \lambda \rangle$ and let $\bar{\alpha} = \langle \alpha \rangle$. Then*

$$\kappa \circ \langle \bar{A}, \bar{\alpha} \rangle = \langle \kappa * \alpha, \kappa * \alpha \rangle.$$

Proof Since λ does not change the state, it follows that $\alpha^{-1}(\lambda) = \alpha$. Therefore

$$\begin{aligned} \kappa \circ \langle \bar{A}, \bar{\alpha} \rangle &= \langle \kappa * \alpha, (\kappa * \alpha) \diamond \lambda \rangle \\ &= \langle \kappa * \alpha, \kappa * \alpha \rangle. \end{aligned}$$

□

Hence, the original revision and update operators can be retrieved through the $\circ$ operator. As such, it is reasonable to define iterated action effects in terms of belief evolution as well. In particular, given $*$ and $\diamond$, we *define* the iterated belief change

$$\kappa \diamond A * \alpha$$

to be the final belief state in the belief trajectory

$$\kappa \circ \langle\langle A \rangle, \langle \alpha \rangle\rangle.$$

Note that, under this convention, updates and revisions are not simply applied in succession.

Proposition 6 *If $\langle A \rangle || \langle \alpha \rangle$, the iterated belief change $\kappa \diamond A * \alpha$ defined as above satisfies the interaction properties P1-P5.*

Proof Let κ be a belief state. By the convention outlined above,

$$\kappa \diamond A * \alpha = (\kappa * \alpha^{-1}(A)) \diamond A.$$

We demonstrate that this definition satisfies P1-P5.

P1. If $(2^F \diamond A) \cap \alpha \neq \emptyset$, then $(\kappa \diamond A) * \alpha \subseteq \alpha$.
Note that the antecedent is true because A and α are consistent. We have the following inclusions:

$$(\kappa * \alpha^{-1}(A)) \diamond A \ \subseteq \ \alpha^{-1}(A) \diamond A \ \subseteq \ \alpha.$$

The first inclusion holds by [AGM2], and the second holds by definition of the pre-image. Hence, the consequent is true.

P2. If $(2^F \diamond A) \cap \alpha = \emptyset$, then $(\kappa \diamond A) * \alpha = \kappa \diamond A$
The antecedent is false, since A and α are consistent.

P3. $(\kappa \diamond A) \cap \alpha \subseteq (\kappa \diamond A) * \alpha$
Suppose $s \in (\kappa \diamond A) \cap \alpha$. So $s \in \alpha$ and there is some $s' \in \kappa$ such that A maps s' to s. Hence, $s' \in \alpha^{-1}(A)$. It follows from [AGM2] that $s' \in \kappa * \alpha^{-1}(A)$. Since A maps s' to s, we have $s \in (\kappa * \alpha^{-1}(A)) \diamond A$.

P4. If $(\kappa \diamond A) \cap \alpha \neq \emptyset$, then $(\kappa \diamond A) * \alpha \subseteq (\kappa \diamond A) \cap \alpha$
Suppose that $(\kappa \diamond A) \cap \alpha \neq \emptyset$. So there is a state in κ that is mapped to α by the action A. Hence $\kappa \cap \alpha^{-1}(A) \neq \emptyset$. By [AGM3] and [AGM4], it follows that $\kappa * \alpha^{-1}(A) = \kappa \cap \alpha^{-1}(A)$. Now suppose that $s \in (\kappa * \alpha^{-1}(A)) \diamond A$. So there exists $s' \in \kappa * \alpha^{-1}(A)$ such that A maps s' to s. But then $s' \in \kappa \cap \alpha^{-1}(A)$. But this implies that $s \in \kappa \diamond A$ and $s \in \alpha$.

P5. $(\kappa \diamond A) * \alpha \subseteq 2^F \diamond A$
This is immediate, because $(\kappa * \alpha^{-1}(A)) \subseteq 2^F$ and the update operator acts on a state by state basis.
□

The three preceding propositions demonstrate the suitability of $\circ$ as a natural operator for reasoning about the interaction between revision and update. We remark that Proposition 6 can easily be generalized to the case where a sequence $\bar{A}$ of ontic actions is followed by a single observation.

3.4.3 Fallible Observations

We address fallible observations by allowing inconsistent world views. Recall that the world does not change except in response to ontic actions. As such, the goal of the underlying agent is to determine the most plausible initial belief state, given all observations that occur over time. In this section, we give a simple procedure for combining all observations in an inconsistent world view.

For consistent world views, we could simply take the intersection of the pre-image of all observations because it was guaranteed to be non-empty. For inconsistent world views, the intersection is empty; the observations and actions are not mutually satisfiable. The basic idea behind our approach is to keep the most reliable observations, and discard the least reliable observations. In order to proceed, we need to assume that we are given a reliability ordering over all observations. Given such an ordering, we can extract a maximally consistent sub-view that incorporates the most reliable observations and resolves inconsistency by discarding unreliable observations.

We start by defining belief evolution in the most general case, with respect to an arbitrary ordering. After presenting the general case, we restate the main definitions for the concrete example where reliability is determined by the recency of an observation.

In the general case, we define a belief evolution operator with respect to a total ordering $\prec$ over natural numbers. Informally, if $j \prec i$, then inconsistency caused by α_i and α_j is handled by discarding α_i.

Definition 17 *Let $W = \langle \bar{A}, \bar{\alpha} \rangle$ be a world view of length n and let $\prec$ be a total ordering over $1, \ldots, n$. Define $\tau(W, \prec) = \langle \bar{A}, \bar{\alpha}' \rangle$, where $\bar{\alpha}' = \langle \alpha'_1, \ldots, \alpha'_n \rangle$ is defined by the following recursion.*

- *If $\alpha^{-1}_{\min_\prec}(\bar{A}_{\min_\prec}) \neq \emptyset$ then $\alpha'_{\min_\prec} = \alpha_{\min_\prec}$,*
  otherwise $\alpha'_{\min_\prec} = 2^{\mathbf{F}}$.
- *For $i \neq \min_\prec$, if*

$$\bar{\alpha}_i^{-1}(\bar{A}_i) \cap \bigcap_{j \prec i} (\alpha'_j)^{-1}(\bar{A}_j) \neq \emptyset$$

  then $\alpha'_i = \alpha_i$,
  otherwise $\alpha'_i = 2^{\mathbf{F}}$.

The observations in $\tau(W, \prec)$ are determined by starting with the most reliable observation, then working progressively through the $\prec$-ordering of observations. At each point, we keep an observation if it is consistent with the observations that are more reliable; otherwise, we discard the observation as incorrect. We remark that $\prec$ is restricted to be a total ordering in order to avoid inconsistent observations that are equally reliable.

The following properties are immediate for any W and $\prec$.

- $\tau(W, \prec)$ is consistent.
- If W is a consistent, then $\tau(W, \prec) = W$.

Recall that the original definition of $\circ$ applied only to consistent world views. By passing through τ, we can extend the definition to apply to arbitrary world views. In the general case, this requires $\circ$ to be parameterized by an ordering.

Definition 18 *Let κ be a belief state, let W be a world view of length n, and let $\prec$ be a total ordering over $1, \ldots, n$. If W is inconsistent, then $\kappa \circ_\prec W = \kappa \circ \tau(W, \prec)$.*

We could equivalently have stated a single definition for $\circ_\prec$ by passing all world views through τ. We have presented the definition in two cases in order to highlight the distinct treatment of fallible observations. Note that, if there is just a single observation, then the

definition of $\circ_{\prec}$ is equivalent to the original definition and it does not rely on $\prec$. As such, Proposition 6 still holds, so the operators $*$ and $\diamond$ obtained from $\circ_{\prec}$ satisfy the interaction properties P1-P5.

Introducing an ordering over observations is cumbersome in many examples, so we would like to choose a default ordering. One natural choice is to prefer recent observations over older observations; this convention has previously been explored in [75] and [80]. In our framework, a preference for recent information is represented by taking $\prec$ to be the inverse of the usual ordering $<$ on the natural numbers. We let $<_{inv}$ denote this ordering, and we restate Definition 17 for the case in which the unstated ordering is $<_{inv}$.

Definition 19 *Let $W = \langle \bar{A}, \bar{\alpha} \rangle$ be a world view of length n. Define $\tau(W) = \langle \bar{A}, \bar{\alpha}' \rangle$, where $\bar{\alpha}' = \langle \alpha'_1, \ldots, \alpha'_n \rangle$ is defined by the following recursion.*

- *If $\alpha_n^{-1}(\bar{A}) \neq \emptyset$ then $\alpha'_n = \alpha_n$,*
  otherwise $\alpha'_n = 2^{\mathbf{F}}$.
- *For $i < n$, if*

$$\alpha_i^{-1}(\bar{A}_i) \cap \bigcap_{i<j} (\alpha'_j)^{-1}(\bar{A}_j) \neq \emptyset$$

  then $\alpha'_i = \alpha_i$,
  otherwise $\alpha'_i = 2^{\mathbf{F}}$.

Hence, the observations in $\tau(W)$ are determined by starting with the most recent observation, then working backwards through the observations from most recent until the initial observation. At each point, we keep an observation if it is consistent with the observations that followed. Throughout the remainder of this thesis, we use $\circ$ to denote the belief evolution operator $\circ_{<_{inv}}$ obtained by giving greater credence to recent observations. The following definition makes this convention precise.

Definition 20 *Let κ be a belief state, and let W be a world view of length n. If W is inconsistent, then $\kappa \circ W = \kappa \circ \tau(W)$.*

We stress that the preference for recent information is just a convention that we adopt because it simplifies the exposition. We are not making any substantive claim about the relative importance of observations. However, we will demonstrate in §3.5 that our default

ordering can be formally justified by illustrating that, if we restrict attention to null actions, then $\circ$ defines a reasonable approach to iterated revision.

We conclude this section with one final result. Thus far, applying the $\circ$ operator requires tracing action preconditions back to the initial state for revision, then applying action effects to get a complete history. If we are only concerned with the final belief state, then there are many cases in which we do not need to go to so much effort.

Proposition 7 *Let κ be a belief state, let $\bar{A}$ be an action trajectory of length n and let α be a belief state such that $\alpha \subseteq \kappa \diamond \bar{A}$. If $\bar{\alpha}$ is the observation trajectory with $n-1$ null observations followed by α, then the final belief state in $\kappa \circ \langle \bar{A}, \bar{\alpha} \rangle$ is $(\kappa \diamond \bar{A}) * \alpha$.*

Proof By definition, the final belief state of $\kappa \circ \langle \bar{A}, \bar{\alpha} \rangle$ is

$$(\kappa * \alpha^{-1}(\bar{A})) \diamond \bar{A}.$$

Since $\alpha \subseteq \kappa \diamond \bar{A}$, the intersection $\kappa \cap \alpha^{-1}(\bar{A})$ is non-empty. By [AGM3] and [AGM4], it follows that

$$\kappa * \alpha^{-1}(\bar{A}) = \kappa \cap \alpha^{-1}(\bar{A})$$

and therefore

$$(\kappa * \alpha^{-1}(\bar{A})) \diamond \bar{A} = (\kappa \cap \alpha^{-1}(\bar{A})) \diamond \bar{A}.$$

Clearly, the right hand side of this equality is equal to $(\kappa \diamond \bar{A}) \cap \alpha$. Again, since $\alpha \subseteq \kappa \diamond \bar{A}$, it follows from [AGM3] and [AGM4] that this is $(\kappa \diamond \bar{A}) * \alpha$. □

The proposition indicates that, given a single observation that is consistent with the actions that have been executed, we can simply revise the outcome of the actions and we get the correct final belief state.

3.4.4 Extended Litmus Paper Concluded

We conclude the litmus paper example by giving a plausible treatment based on a belief evolution operator. The world view $W = \langle \langle dip \rangle, \langle \alpha \rangle \rangle$ represents a dipping action followed by the observation that the paper is still white. If $\circ$ is obtained from the metric transition system defined by the Hamming distance and the transitions in Figure 3.1, the final belief state in $\kappa \circ W$ is given by

$$\begin{aligned} \kappa * \alpha^{-1}(dip) \diamond dip &= \kappa * \{\emptyset, \{Acid\}\} \diamond dip \\ &= \{\emptyset, \{Acid\}\}. \end{aligned}$$

This calculation is consistent with our original intuitions, in that the agent revises the initial belief state before updating by the *dip* action. This ensures that we will have a final belief state that is a possible outcome of dipping. Moreover, the initial belief state is revised by the pre-image of the final observation, which means it is modified as little as possible while still guaranteeing that the final observation will be feasible. Note also that the final belief state given by this calculation is intuitively plausible. It simply indicates that the contents of the beaker are still unknown, but the agent now believes the paper is not litmus paper. Hence, a belief evolution operator employs a plausible procedure and returns a desirable result.

3.5 Relationship with Iterated Revision

3.5.1 Darwiche-Pearl Revision

If the null action is the only action permitted, then belief evolution is closely related to iterated revision. In this section, we consider the suitability of belief evolution operators for reasoning about iterated revision, from the perspective of the Darwiche-Pearl approach.

In the following observation, let $\bar{\lambda}$ denote a sequence of null actions of indeterminate length.

Observation 1 *For any κ and $\bar{\alpha}$, there is a unique belief state κ' such that*

$$\kappa \circ \langle \bar{\lambda}, \bar{\alpha} \rangle = \langle \kappa', \ldots, \kappa' \rangle.$$

This observation is consistent with the view that belief evolution operators return a trajectory representing an agent's current beliefs about the evolution of the world. We remark that, in general, κ' is not obtained by successively revising by the elements of $\bar{\alpha}$. Moreover, we claim that this is appropriate because we have only assumed an underlying AGM revision operator. It is well known that many AGM revision operators do not satisfy the Darwiche-Pearl postulates for iterated revision; for example, the Hamming distance based revision operator does not satisfy the postulates. We will prove that belief evolution operators satisfy the Darwiche-Pearl postulates, even when the underlying revision operators do not.

We restate the Darwiche-Pearl postulates in terms of possible worlds. Let κ, α, and β be sets of possible worlds. Let $\tilde{\beta}$ denote the complement of β. The Darwiche-Pearl postulates are as follows.

Darwiche-Pearl Postulates

[DP1] If $\alpha \subseteq \beta$, then $(\kappa * \beta) * \alpha = \kappa * \alpha$.

[DP2] If $\alpha \subseteq \tilde{\beta}$, then $(\kappa * \beta) * \alpha = \kappa * \alpha$.

[DP3] If $\kappa * \alpha \subseteq \beta$, then $(\kappa * \beta) * \alpha \subseteq \beta$.

[DP4] If $\kappa * \alpha \not\subseteq \tilde{\beta}$, then $(\kappa * \beta) * \alpha \not\subseteq \tilde{\beta}$.

If we define the iterated revision $\kappa * \alpha_1 * \cdots * \alpha_n$ to be the unique belief state in $\kappa \circ \langle \bar{\lambda}, \bar{\alpha} \rangle$, then we get the following result.

Proposition 8 *Let $\circ$ be a belief evolution operator. For non-empty observations, the iterated revision operator obtained from $\circ$ satisfies the Darwiche-Pearl postulates.*

Proof We abuse notation and let $\kappa \circ \langle \lambda, \langle \alpha \rangle \rangle$ denote the unique belief state κ' from Observation 1. Similarly, we identify $\kappa \circ \langle \lambda, \langle \beta, \alpha \rangle \rangle$ with the corresponding belief state, rather than the complete belief trajectory. Under this convention, the iterated revision $\kappa * \beta * \alpha$ is equal to $\kappa \circ \langle \lambda, \langle \beta, \alpha \rangle \rangle$. Note that $\alpha^{-1}(\lambda) = \alpha$ and $\beta^{-1}(\lambda) = \beta$. Since $\alpha \neq \emptyset$, it follows that

$$\kappa * \beta * \alpha = \begin{cases} \kappa * (\beta \cap \alpha) & \text{if } \beta \cap \alpha \neq \emptyset \\ \kappa * \alpha & \text{otherwise} \end{cases}$$

For [DP1], suppose that $\alpha \subseteq \beta$. Since $\alpha \neq \emptyset$, it follows that $\beta \cap \alpha \neq \emptyset$ and hence $\kappa * \beta * \alpha = \kappa * (\beta \cap \alpha)$. But $\beta \cap \alpha = \alpha$, so the right hand side is equal to $\kappa * \alpha$.

For [DP2], suppose that $\alpha \subseteq \tilde{\beta}$. Hence, $\alpha \cap \beta = \emptyset$ and the desired conclusion follows immediately.

For [DP3], suppose that $\kappa * \alpha \subseteq \beta$. Since $\alpha \neq \emptyset$, it follows from [AGM5] that $\kappa * \alpha \neq \emptyset$. So there exists some $s \in \kappa * \alpha$. But then $s \in \alpha$ since $\kappa * \alpha \subseteq \alpha$, and $s \in \beta$ since $\kappa * \alpha \subseteq \beta$. Hence $\alpha \cap \beta \neq \emptyset$, and therefore

$$\kappa * \beta * \alpha = \kappa * (\beta \cap \alpha) \subseteq \beta \cap \alpha \subseteq \beta.$$

For [DP4], suppose that $\kappa * \alpha \not\subseteq \tilde{\beta}$. So there exists some $s \in \kappa * \alpha$ such that $s \in \beta$. It follows that $s \in \alpha \cap \beta$, so $\alpha \cap \beta \neq \emptyset$. Translating to possible worlds, [AGM8] says the following:

$$\text{if } \kappa * \alpha \not\subseteq \tilde{\beta}, \text{ then } (\kappa * \alpha) \cap \beta \subseteq \kappa * (\alpha \cap \beta).$$

Since $s \in (\kappa * \alpha) \cap \beta$, this implies that $s \in \kappa * (\alpha \cap \beta)$. But then, since $\alpha \cap \beta \neq \emptyset$ it follows by definition that $s \in \kappa * \beta * \alpha$. Hence $s \in \beta$ and $s \in \kappa * \beta * \alpha$. Therefore $s \in \kappa * \beta * \alpha \not\subseteq \tilde{\beta}$.

□

Hence, belief evolution always defines a Darwiche-Pearl operator, even if the underlying revision operator fails to satisfy the postulates.

It is easy to demonstrate that belief evolution satisfies the so-called *recalcitrance* postulate introduced in [76]. Rephrased in terms of possible worlds and belief evolution, recalcitrance is the following property:

(Recalcitrance) If $\beta \cap \alpha \neq \emptyset$, then $(\kappa \circ \langle \bar{\lambda}, \langle \beta, \alpha \rangle) \subseteq \beta$.

It is known that DP1, DP2, and (Recalcitrance) characterize Nayak's lexicographic iterated revision operator on epistemic states [10]. Hence, although we have defined belief evolution strictly in terms of an underlying AGM operator, it turns out that it is essentially equivalent to a well-known approach to Darwiche-Pearl revision.

Rather than assuming an underlying AGM revision operator in the definition of belief evolution, we could have assumed an underlying Darwiche-Pearl operator. Of course, Darwiche-Pearl operators are still AGM operators. The main difference for our purposes is that Darwiche and Pearl define revision with respect to epistemic states which explicitly define an ordering on the set of possible states. In the remainder of this section, we briefly summarize how to define belief evolution in this context.

Recall that Darwiche-Pearl revision is based on so-called *epistemic states.* In order to define belief evolution in terms of epistemic states, we need to first define belief update in terms of epistemic states. Let E be an epistemic state with the corresponding ordering $\preceq_E$. The epistemic state $E \diamond A$ can be obtained by defining the new ordering $\preceq_{E \diamond A}$ as follows:

$$w \diamond A \preceq_{E \diamond A} v \diamond A \iff w \preceq_E v.$$

In order to complete the definition of the new ordering, we need to consider states w such that $w \neq v \diamond A$ for any v. For such states, we simply specify that

$$u \diamond A \preceq_{E \diamond A} w$$

for every u. Extending update to epistemic states in this manner allows us to extend the definition of belief evolution for iterated revision operators. In particular, given a Darwiche-Pearl operator $*$ and an update operator $\diamond$, we can define $\circ$ as follows.

Definition 21 *(DP version) Let E be an epistemic state, let $\bar{A}$ be an action trajectory of length n and let $\bar{\alpha}$ be an observation trajectory of length n such that $\bar{A}||\bar{\alpha}$. Define*

$$E \circ \langle \bar{A}, \bar{\alpha} \rangle = \langle E_0, \ldots, E_n \rangle$$

where

1. $E_0 = E * \alpha_1^{-1}(\bar{A}_1) * \cdots * \alpha_n^{-1}(\bar{A}_n)$
2. *for* $i \geq 1$, $E_i = E_{i-1} \diamond A_1 \diamond \cdots \diamond A_i$.

This is the same definition that was used for infallible observations, rephrased in terms of epistemic states. In the new definition, no observation is discarded. Instead, the definition of $*$ will determine how conflicting observations should be treated. We remark that this new definition is only appropriate for action domains where the reliability of an observation is determined by recency. In the general case, where there is an arbitrary reliability ordering over observations, we need to perform the iterated revision in the order dictated by this ordering.

Definition 21 illustrates how belief evolution can be defined with respect to an underlying Darwiche-Pearl operator. The definition is marginally simpler than our original definition, and it allows for a more flexible approach to iterated revision. As such, we suggest that this approach is preferable for action domains where a Darwiche-Pearl operator is available. We have formulated the original definition in terms of AGM revision in the interest of generality, so that we can address iterated epistemic action effects given only an AGM operator. For example, we are interested in belief change in the context of metric transition systems, where the only available revision operator is AGM.

3.5.2 Lehmann Postulates

Another important set of postulates for iterated revision is proposed by Lehmann [59]. In this section, we consider belief evolution from the perspective of Lehmann's postulates.

We introduce some shorthand notation to simplify the statement of Lehmann's postulates. Given observation trajectories O and O', let $O \cdot O'$ denote the concatenation of the two sequences. Similarly, if α is an observation, we write $O \cdot \alpha$ as a shorthand for $O \cdot \langle \alpha \rangle$. Finally, since we are only interested in null actions, for the remainder of this section we write $\kappa \circ O$ as an abbreviation for the final belief state in $\kappa \circ \langle \bar{\lambda}, O \rangle$.

Let O, O' denote observation trajectories, let κ, α, β denote sets of states, and let $\tilde{\beta}$ denote the complement of β. Translated into our notation, the Lehmann postulates are as follows.

Lehmann Postulates

[L1] $\kappa \circ O = \emptyset$ iff $\kappa = \emptyset$.

[L2] $\kappa \circ (O \cdot \alpha) \subseteq \alpha$.

[L3] If $\kappa \circ (O \cdot \alpha) \subseteq \beta$ and $\kappa \circ O \subseteq \alpha$, then $\kappa \circ O \subseteq \beta$.

[L4] If $\kappa \circ O \subseteq \alpha$, then $\kappa \circ (O \cdot O') = \kappa \circ (O \cdot \alpha \cdot O')$.

[L5] If $\beta \subseteq \alpha$, then $\kappa \circ (O \cdot \alpha \cdot \beta \cdot O') = \kappa \circ (O \cdot \beta \cdot O')$.

[L6] If $\kappa \circ (O \cdot \alpha) \not\subseteq \tilde{\beta}$, then $\kappa \circ (O \cdot \alpha \cdot \beta \cdot O') = \kappa \circ (O \cdot \alpha \cdot \alpha \cap \beta \cdot O')$.

[L7] $\kappa \circ (O \cdot \alpha) \subseteq \kappa \circ (O \cdot \tilde{\alpha} \cdot \alpha)$.

Belief evolution does not satisfy all of the Lehmann postulates. In particular, we present a counterexample illustrating that [L4]-[L6] all fail.

Example Let s_1, s_2, s_3 be states over some action signature. Define κ, α, O and O' as follows:

$$\begin{aligned} \kappa &= \{s_1\} \\ \alpha &= \{s_2, s_3\} \\ \beta &= \{s_3\} \\ O &= \langle\{s_3\}\rangle \\ O' &= \langle\{s_1, s_2\}\rangle \end{aligned}$$

We will demonstrate that [L4]-[L6] all fail for this example.

Let $\circ$ be a belief evolution operator obtained from some update operator $\diamond$ and some AGM revision operator $*$. Note that

$$\kappa \circ O = \{s_1\} * \{s_3\} = \{s_3\} \subseteq \alpha.$$

However,

$$\begin{aligned} \kappa \circ (O \cdot O') &= \{s_1\} * \{s_1, s_2\} = \{s_1\} \\ \kappa \circ (O \cdot \alpha \cdot O') &= \{s_1\} * \{s_2\} = \{s_2\}. \end{aligned}$$

Hence $\kappa \circ (O \cdot O') \neq \kappa \circ (O \cdot \alpha \cdot O')$, which violates [L4]. Since $\beta = \kappa \circ O$, this also violates [L5].

Let $\gamma = \{s_1, s_3\}$. The following equalities refute [L6].

$$\begin{aligned} \kappa \circ (O \cdot \alpha) &= \{s_3\} \not\subseteq \tilde{\gamma} \\ \kappa \circ (O \cdot \alpha \cdot \gamma \cdot O') &= \{s_1\} \\ \kappa \circ (O \cdot \alpha \cdot \alpha \cap \gamma \cdot O') &= \{s_2\}. \end{aligned}$$

Lehmann views [L4]-[L6] as dealing with "superfluous revisions" [59]. For example, in postulate [L4], the observation α is superfluous because revising by the observations in O already leads an agent to believe that the actual state is in α. As such, observing α after O does not provide any new information. The postulate [L4] suggests that such observations may be discarded. This kind of reasoning is not supported in belief evolution, because the observation α may take on new meaning following future observations. Postulates [L5] and [L6] fail for similar reasons.

Although [L4]-[L6] do not hold, we can construct weaker versions that do hold. We have claimed that the reason these postulates fail is because future observations may affect the interpretation of observations that are initially superfluous. To avoid this problem, we modify the postulates by removing the observations that follow a superfluous observation. Weakening gives the following postulates:

Weak Lehmann Postulates

[L4*] If $\kappa \circ O \subseteq \alpha$ and $O \neq \bar{\emptyset}$, then $\kappa \circ O = \kappa \circ (O \cdot \alpha)$.

[L5*] If $\beta \subseteq \alpha$, then $\kappa \circ (O \cdot \alpha \cdot \beta) = \kappa \circ (O \cdot \beta)$.

[L6*] If $\kappa \circ (O \cdot \alpha) \not\subseteq \tilde{\beta}$, then $\kappa \circ (O \cdot \alpha \cdot \beta) = \kappa \circ (O \cdot \alpha \cdot \alpha \cap \beta)$.

If [L4]-[L6] are replaced with [L4*]-[L6*], then belief evolution satisfies the resulting set of postulates.

Proposition 9 *Let $\circ$ be a belief evolution operator, obtained from $*$ and $\diamond$. If α and β are non-empty, then $\circ$ satisfies [L1],[L2],[L3],[L4*],[L5*],[L6*], and [L7].*

Proof Clearly, if $\kappa = \emptyset$ then $\kappa \circ O = \emptyset$. For the converse, suppose that O contains only empty observations. In this case, $\kappa \circ O = \kappa$ and the result holds. If O contains some non-empty observations, then $\kappa \circ O = \kappa * \gamma$ for some non-empty intersection of observations γ. But then $\kappa * \gamma = \emptyset$ just in case κ is empty. This completes the proof of [L1].

Since $\alpha \neq \emptyset$, $\kappa \circ (O \cdot \alpha) = \kappa * \gamma$ for some $\gamma \subseteq \alpha$. Hence $\kappa \circ (O \cdot \alpha) \subseteq \gamma \subseteq \alpha$, which proves [L2].

For [L3], suppose that $\kappa \circ (O \cdot \alpha) \subseteq \beta$ and $\kappa \circ O \subseteq \alpha$. Now suppose that $s \in \kappa \circ O$. By definition, this means that $s \in \kappa * \tau(O)$. By [AGM7], we have $(\kappa * \tau(O)) \cap \alpha \subseteq \kappa * (\tau(O) \cap \alpha$. But $\kappa * (\tau(O) \cap \alpha = \kappa \circ (O \cdot \alpha)$, so it follows that $s \in \beta$. Therefore $\kappa \circ O \subseteq \beta$.

Suppose that $\kappa \circ O \subseteq \alpha$ and $O \neq \bar{\emptyset}$. By definition, $\kappa \circ O = \kappa * \bigcap \tau(O)$. Since $O \neq \bar{\emptyset}$, it follows that $\kappa \circ (O \cdot \alpha) = \kappa * (\bigcap \tau(O) \cap \alpha)$. With these equalities in mind, the following results prove that [L4*] holds:

$$\begin{aligned} \kappa * \bigcap \tau(O) \subseteq & \ (\kappa * \bigcap \tau(O)) \cap \alpha && \text{(since } \kappa \circ O \subseteq \alpha) \\ \subseteq & \ \kappa * (\bigcap \tau(O) \cap \alpha) && \text{(by [AGM8])} \\ \subseteq & \ \kappa * \bigcap \tau(O) && \text{(by [AGM7])} \end{aligned}$$

It is clear that [L5*] holds, simply because the assumption that $\beta \subseteq \alpha$ implies that $\beta = \alpha \cap \beta$.

For [L6*], suppose that $\kappa \circ (O \cdot \alpha) \not\subseteq \tilde{\beta}$. It follows that $\alpha \cap \beta \neq \emptyset$. By definition, this means that $\kappa \circ (O \cdot \alpha \cdot \beta) = \kappa \circ (O \cdot \alpha \cdot \alpha \cap \beta)$, which is the desired result.

Since $\tilde{\alpha} \cap \alpha = \emptyset$, it follows by definition that $\kappa \circ (O \cdot \alpha) = \kappa \circ (O \cdot \tilde{\alpha} \cdot \alpha)$. Clearly this entails [L7]. □

Hence, if we do not consider the influence of observations that will occur in the future, then belief evolution defines an approach to iterated revision that satisfies the Lehmann postulates.

We conclude with a brief remark about the assumption that α and β are non-empty in Propositions 8 and 9. In both cases, the postulates would not hold if empty observations were permitted. In our framework, the empty set represents an inconsistent observation and the reason we need to restrict the postulates to non-empty observations is because we treat inconsistency in a non-standard manner. The AGM approach, the Darwiche-Pearl approach, and the Lehmann approach all allow inconsistent observations to lead to inconsistent belief states. By contrast, we discard inconsistent observations and keep the

original belief state. The rationale behind our choice has little to do with any underlying assumption about inconsistent observations. Instead, the rationale behind our choice is based on the assumption that action histories are infallible. This assumption leads us to reason as follows. If a sequence of actions $\bar{A}$ has been executed, then we discard any observations that are inconsistent with the effects of $\bar{A}$. The motivation behind this approach is simply to ensure that action histories take precedence over observation histories. Since inconsistent observations are clearly inconsistent with every sequence of actions, it follows that inconsistent observations should always be discarded. If $\bar{A}$ is a sequence of null actions, we treat empty observations in the same manner. We accept this treatment of inconsistent observations, because it allows inconsistency to be treated in a uniform manner.

3.6 Comparison with Related Formalisms

3.6.1 The Scope of Belief Evolution

The litmus paper problem illustrates that agents sometimes need to revise prior belief states. We have described the problem as a belief update followed by a belief revision, and we have demonstrated that belief evolution provides a reasonable representation for this particular example. However, belief evolution is not necessarily suitable for all problems involving update and revision. In order to compare belief evolution with related formalisms, we must first delineate the class of problems that are appropriate for belief evolution.

The class of problems that are appropriate for belief evolution can be described by an ordering over action histories. Let $A_1, \alpha_1, \ldots, A_n, \alpha_n$ be an alternating sequence of actions and observations. Let $\prec$ denote a total pre-order over the elements of this sequence. Given $\prec$, we are interested in giving a natural interpretation to

$$\kappa \diamond A_1 * \alpha_1 \diamond \cdots \diamond A_n * \alpha_n.$$

The idea is to incorporate as many actions and observations as possible, giving preference to events that are ranked low in the $\prec$-ordering. Belief evolution is suitable for problems in which the underlying ordering is given as follows, for some permutation $p_1, \ldots, p_n$ of $1, \ldots, n$.

$$\left.\begin{array}{l} A_1 \\ \vdots \\ A_n \end{array}\right\} \prec \alpha_{p_1} \prec \alpha_{p_2} \prec \cdots \prec \alpha_{p_n}$$

Note that every A_i is minimal, so all ontic actions must be incorporated. This is possible since we assume all actions are always executable. After incorporating every ontic action, the observations are incorporated as much as possible.

There are two natural classes of problems that can not be represented by the ordering given above. First, there are problems in which observations should be given equal weight.

$$\left.\begin{matrix} A_1 \\ \vdots \\ A_n \end{matrix}\right\} \prec \left\{\begin{matrix} \alpha_1 \\ \vdots \\ \alpha_n \end{matrix}\right.$$

Consider, for example, an agent that is trying to determine the temperature by checking n digital thermometers. At each time i, the agent takes a reading from one thermometer. In this context, belief evolution is not appropriate; the different readings should be combined in a manner that gives equal credence to all readings. The second natural class of problems that is not representable is the class of problems where observations can be more reliable than actions. This is a plausible situation if we think of observations as infallible sensing actions. For belief evolution, it is more plausible to think of observations as reports that the agent receives from an external source. The external source may represent sensory information, but it must be understood that the sensing information is less reliable than the action history.

In Chapter 5, we consider a generalization of belief evolution that makes it possible to address both of these classes of problems. However, for the moment, we are interested in the relationship between belief evolution and related formalisms for reasoning about epistemic action effects. We focus on formalisms where the reliability of actions and observations can reasonably be described in the manner outlined above.

3.6.2 Markovian Formalisms

Given a belief update operator and a belief revision operator, we tend to associate iterated belief change operations as follows:

$$(((\kappa \diamond A_1) * \alpha_1) \diamond \cdots \diamond A_n) * \alpha_n.$$

This is the implicit approach to iterated belief change caused by actions in Markovian action formalisms, such as the epistemic extensions of $\mathcal{A}$. This approach does not capture a reasonable preference ordering for litmus-type problems. The belief evolution methodology

basically gives a more justifiable way to combine existing belief change operators for this kind of problem.

It is important to note that belief evolution should not be seen as a formalism in competition with Markovian formalisms; it should be seen as a methodology for extending Markovian formalisms to address iterated belief change. If the revision and update operators are given explicitly, the definition of the corresponding belief evolution operator is straightforward. This is true even in formalisms where the basic operators are relatively sophisticated, such as those defined in the multi-agent belief structures of Herzig, Lang and Marquis [44]. In Chapter 4, we illustrate how belief evolution can be applied in the action language framework by defining an epistemic extension of $\mathcal{A}$ that extends both [68] and [86].

3.6.3 The Situation Calculus

The SitCalc is one action formalism that considers the action history when determining the epistemic effects of actions. In this section, we demonstrate that the epistemic extension of the SitCalc implicitly defines a belief evolution operator.

Recall that the epistemic extension of the SitCalc introduces an accessibility relation K, together with a numerical function pl over situations. Belief is defined by the following formula.

$$Bel(\phi, s) \iff \forall s'[K(s', s) \wedge (\forall s'' K(s'', s) \rightarrow pl(s') \leq pl(s''))] \rightarrow \phi[s'].$$

Revision by a sensing action A is defined by the operator $*_A$, given by

$$\kappa_s *_A \phi = \kappa_{do(A,s)}.$$

We noted previously that this operator satisfies 5 of the AGM postulates [85]. The fact that $*_A$ does not satisfy all of the AGM postulates may be seen as a weakness of the approach. However, this is only a weakness if we believe that $*_A$ should be a revision operator. We suggest instead that it should not be a revision operator. The belief change that occurs in this context is the result of an observation that follows a sequence of actions; as such, the belief change should be defined by belief evolution. Although $*_A$ is called a revision operator, in the remainder of this section we demonstrate that the semantics of $*_A$ is better understood through belief evolution.

First, we demonstrate that $*_A$ is not even a function on belief states. The following example demonstrates that computing the result of the sensing action A sometimes requires

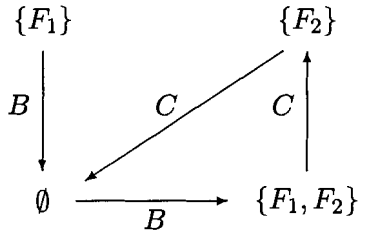

Figure 3.2: A Transition System

an agent to consider the action history.

Example Let F_1 and F_2 be the only fluent symbols in the language, let A be a sensing action for F_2, and let B, C be ontic action symbols. Suppose that we have a SitCalc theory in which action effects are given by the transition system in Figure 3.2. Suppose further that there are 4 initial situations corresponding to each interpretation of $\{F_1, F_2\}$. In particular, define s_1, s_2, s_3, s_4 as follows:

$$\begin{aligned} s_1 &\models \neg F_1 \wedge \neg F_2 \\ s_2 &\models F_1 \wedge \neg F_2 \\ s_3 &\models \neg F_1 \wedge F_2 \\ s_4 &\models F_1 \wedge F_2. \end{aligned}$$

The accessibility relation K is an equivalence relation with equivalence classes $\{s_1, s_2\}$ and $\{s_3, s_4\}$. The plausibility function pl is defined as follows:

$$pl(s) = \begin{cases} 0 & \text{if } s = s_2 \text{ or } s = s_3 \\ 1 & \text{otherwise} \end{cases}$$

We focus on the initial situations s_1 and s_4. In the situation s_1, the agent believes that the actual situation is s_2. As a result, in s_1, the agent correctly believes that F_2 is false and erroneously believes that F_1 is true. Similarly, in s_4, the agent correctly believes that F_2 is true and erroneously believes that F_1 is false.

By definition, $\kappa_{s_1} = \{F_1\}$, and $\kappa_{s_4} = \{F_2\}$. It follows that

$$\kappa_{do(B,s_1)} = \kappa_{do(C,s_4)} = \emptyset.$$

To see this, simply project the beliefs according to the effects of B and C given in the transition system. Since $\kappa_{do(B,s_1)}$ and $\kappa_{do(C,s_4)}$ denote the same belief state, given any formula ϕ and any AGM revision operator $*$, it follows that

$$\kappa_{do(B,s_1)} * \phi = \kappa_{do(C,s_4)} * \phi.$$

We will now demonstrate that the revision actions of the SitCalc do not have this property.

Consider the effects of the sensing action A. Since $(s_1, s_2) \in K$, it follows that

$$(do(B, s_1), do(B, s_2)) \in K.$$

However, note that $do(B, s_1) \models F_2$ whereas $do(B, s_2) \not\models F_2$. As a result, according to the successor state axiom for A,

$$(do([B, A], s_1), do([B, A], s_2)) \notin K.$$

The only state accessible from $do([B, A], s_1)$ is $do([B, A], s_1)$. It follows that

$$\kappa_{do(B,s_1)} *_A F_2 = \{F_1, F_2\}.$$

By a similar argument, it is easy to see that

$$\kappa_{do(C,s_4)} *_A F_2 = \{F_2\}.$$

Hence, $*_A$ is not a function on belief states; it maps the same belief state to different outcomes, depending on the situation that induces the belief state.

The preceding example illustrates that the so-called revision operators of the SitCalc are not revision operators in the AGM sense. The outcome of a sensing action depends not only on the fluents that are initially believed, but also on the history of ontic actions that have been executed. We propose that it is more natural to view belief change in the SitCalc as belief evolution.

Before proceeding, we need to restrict the class of permissible SitCalc theories. One important difference between situations and states is that two distinct situations can satisfy the exact same set of fluents. For the purpose of comparison with belief evolution, we would like to push this difference aside. In the remainder of this section, we restrict attention to SitCalc theories in which the initial situations all correspond to distinct states.

Let $\mathbf{F}$ denote a finite set of fluent symbols and let $\mathbf{A} = \mathbf{A}_O \cup \mathbf{A}_S$, where $\mathbf{A}_O$ is a set of ontic action symbols and $\mathbf{A}_S$ is a set of sensing action symbols. Each $O \in \mathbf{A}_S$ denotes a sensing action for some literal F_O. Let $\mathcal{T}$ be a SitCalc theory for $\mathbf{F}$ and $\mathbf{A}$ with the property that there are $2^{\mathbf{F}}$ initial situations, each of which satisfies a distinct interpretation of $\mathbf{F}$. It is clear that this condition can be axiomatized in $\mathcal{T}$.

Every situation s can be written in the form

$$s = do(\bar{A}, s_I)$$

where $\bar{A}$ is a sequence of actions and s_I is an initial situation. The state I_s is defined to be the interpretation satisfying

$$I_s \models F \iff F(s).$$

For any sequence of actions $\bar{A}$, define $\bar{A}'$ to be the minimal alternating sequence of ontic actions and sensing actions obtained by inserting null actions into $\bar{A}$. Observe that, for any $s = do(\bar{A}, s_I)$, we have the following property

$$I_{do(\bar{A}, s_I)} = I_{do(\bar{A}', s_I)}.$$

Recall that each sensing action O is associated with some sensed fluent F_O. It is convenient for our purposes to also allow the null sensing action λ, which is not associated with any fluent. We want to translate SitCalc sensing actions into the observations of belief evolution. Given a situation s, define the observation $O(s)$ as follows:

$$O(s) = \begin{cases} 2^{\mathbf{F}} & \text{if } O = \lambda \\ |F_O| & \text{if } s \models F_O \\ |\neg F_O| & \text{if } s \not\models F_O \end{cases}$$

The following definition associates a particular world view with the situation $do(\bar{A}, s_I)$.

Definition 22 *Let $s = do(\bar{A}, s_I)$ and let $\bar{A}'$ be the associated alternating sequence*

$$A_1, O_1, \ldots, A_n, O_n.$$

Define $wv(s)$ to be the world view $\langle \bar{A}, \bar{\alpha} \rangle$ where

1. $\bar{A} = \langle A_1, \ldots, A_n \rangle$
2. $\bar{\alpha} = \langle O_1(do(A_1, s_I)), \ldots, O_n(do(\bar{A}, s_I)) \rangle$

Intuitively, $wv(s)$ is obtained by keeping the same sequence of ontic actions and replacing each sensing action with the outcome that the action will produce given the initial state s_I and the action sequence $\bar{A}$.

For any situation s, let K_s denote the set $\{x \mid Ksx\}$. If s_I is an initial situation, then K_{s_I} is the set of all initial situations, and we let K_{init} denote the set of pl-minimal elements of K_{s_I}. Recall that we have assumed that the set of initial situations consists of 2^F distinct situations, each corresponding to a unique state. As such, we can think of K_{init} as a set of states and we can think of pl as a ranking function on initial states.

For $A \in \mathbf{A}_O$, define $\diamond$ as follows:

$$\kappa_s \diamond A = \kappa_{do(A,s)}.$$

Note that the plausibility function pl induces a system of spheres over the set of states. Given a fixed initial situation s_I, let $\mathcal{S}$ denote the system of spheres over K_{s_I} that is centered on K_{init}. Let $*$ denote the revision operator obtained from $\mathcal{S}$ and define $\circ$ to be the belief evolution operator obtained from $\diamond$ and $*$. We have the following result.

Proposition 10 *Let $\bar{A}$ be a sequence of actions, let ϕ be a formula, and let $s = do(\bar{A}, s_I)$. If $wv(s)$ is consistent, then $Bel(\phi, s)$ if and only if ϕ holds in the final element of $K_{init} \circ wv(s)$.*

Proof Let $wv(s) = \langle \bar{A}, \bar{\alpha} \rangle$. The final element of $K_{init} \circ wv(s)$ is

$$BFinal = K_{init} * \bigcap_i \alpha_i^{-1}(\bar{A}_i) \diamond \bar{A}.$$

Hence $BFinal$ is the set of states s'' such that there is some initial state s' satisfying

1. $s' \diamond \bar{A} = s''$
2. for all i, $s' \diamond \bar{A}_i \models \alpha_i$
3. $pl(s')$ is minimal among states satisfying 1 and 2.

So $BFinal \models \phi$ just in case ϕ is true in all such states s''. Note that, by the successor state axiom for K, the set $K_{do(\bar{A},s_I)}$ can be obtained from K_{s_I} by applying $\bar{A}$ on a pointwise basis, then removing all situations that disagree on some sensing result, and then keeping only the pl-minimal elements. Since pl values persist following actions, it follows that the pl-minimal elements of $R_{do(\bar{A},s_I)}$ are precisely the situations t such that $I_t \in BFinal$. □

In the following corollary, if s is a situation with $wv(s) = \langle \bar{A}, \bar{\alpha} \rangle$ and l is a literal, we let $wv(s) \cdot \langle \lambda, l \rangle$ denote the world view $\langle \bar{A} \cdot \lambda, \bar{\alpha} \cdot l \rangle$.

Corollary 1 *Let s be a situation, let ϕ be a formula, let l be a literal, and let A be a revision action for l. If $s \models l$, then*

$$\kappa_s *_A l \models \phi$$

if and only if ϕ is true in the final belief state of

$$K_{init} \circ (wv(s) \cdot \langle \lambda, l \rangle).$$

Proof Immediate. □

Corollary 1 indicates that the so-called revision operator of the SitCalc actually returns the final belief state given by a natural belief evolution operator.

Viewing $*_A$ as an evolution operator has several advantages. First of all, it makes it clear that the history of actions plays a role in the operation. Second, as a revision operator, there were plausible arguments against $*_A$ based on the fact that it did not satisfy the AGM postulates. However, when viewed as an evolution operator, this is no longer a problem. The fact that plausibility values persist following the execution of actions is equivalent to restricting belief change by always revising the initial belief state, then determining the final belief state by simply computing ontic action effects. Hence, the semantics of revision actions is based on the same underlying intuition as the semantics of belief evolution.

We conclude this section by remarking that belief evolution operators do have one expressive advantage over the epistemic extension of the SitCalc. The epistemic extension that we have considered makes three important assumptions: (1) actions and sensing are correct, (2) no exogenous actions occur, and (3) the actual initial situation is considered possible. As a result, revising by F followed by $\neg F$ leads to inconsistency. By contrast, in belief evolution, this is simply handled by keeping the more reliable observation. Hence, belief evolution is able to deal with unreliable perception in a straightforward manner that is not possible in the SitCalc. We remark however, that inconsistent observations have been treated in a later extension of the SitCalc postulating exogenous actions to account for the inconsistent sensing information [84]. In the generalization of belief evolution that we present in chapter 5, it will be possible to resolve inconsistent observations in both ways: by eliminating unreliable observations or by postulating exogenous actions.

3.7 A Representation Result

3.7.1 Interaction Postulates

We conclude this chapter with a representation result for belief evolution. In this section, we demonstrate that belief evolution operators can be characterized by a pair of rationality postulates. In the next section, we give an equivalent semantic characterization in terms of systems of spheres.

To simplify the problem, we restrict attention to world views of length 1. In this context, there is no need to provide an axiomatic treatment of conflicting observations. Instead we can focus on providing a rigorous treatment of the interaction between a single update and a single revision. First, we need to delineate a general class of belief change functions, which we will then restrict through a pair of rationality postulates.

Definition 23 *A combined belief change operator is a function*

$$\dot{\circ} : 2^S \times \langle \mathbf{A}, 2^S \rangle \to 2^S.$$

Hence, a combined belief change operator takes a belief state, an action symbol, and an observation as input and it returns a new belief state. We are interested in providing a characterization of all combined belief change operators that correspond to belief evolution operators.

In the remainder of this section, we abuse our notation as follows. Given a belief evolution operator $\circ$, we let $\kappa \circ \langle A, \alpha \rangle$ denote the belief state κ_1 where

$$\kappa \circ \langle \langle A \rangle, \langle \alpha \rangle \rangle = \langle \kappa_0, \kappa_1 \rangle.$$

Hence, we take the outcome of belief evolution to be the final belief state in the corresponding belief trajectory. Under this new notation, it is reasonable to ask if a particular combined belief change operator $\dot{\circ}$ is a belief evolution operator. This change in notation is justified by the fact that, for deterministic actions, $\circ$ is completely determined by specifying the final belief state.

Note that the definition of a combined belief change operator does not mention any particular update operator or any particular revision operator; in the general case, combined belief change operators are just arbitrary functions. However, if we are given an update operator $\diamond$ and a revision operator $*$, then it is possible to specify some simple postulates. In particular, we have the following.

I1 If $(2^F \diamond A) \cap \alpha \neq \emptyset$, then $\kappa \dot{\diamond} \langle A, \alpha \rangle = \kappa * \alpha^{-1}(A) \diamond A$.

I2 If $(2^F \diamond A) \cap \alpha = \emptyset$, then $\kappa \dot{\diamond} \langle A, \alpha \rangle = \kappa \diamond A$.

We now illustrate that these postulates characterize belief evolution on trajectories of length 1.

Proposition 11 *Let $\diamond$ be a belief update operator and let $*$ be a belief revision operator. Then $\dot{\diamond}$ is the belief evolution operator corresponding to $\diamond, *$ if and only if $\dot{\diamond}$ satisfies* **I1** *and* **I2**.

Proof Let $\dot{\diamond}$ be the belief evolution operator corresponding to $\diamond$ and $*$. If $(2^F \diamond A) \cap \alpha \neq \emptyset$, then $\kappa \dot{\diamond} \langle A, \alpha \rangle = \kappa * \alpha^{-1}(A) \diamond A$ by definition. Hence $\dot{\diamond}$ satisfies **I1**. Suppose, on the other hand, that $(2^F \diamond A) \cap \alpha = \emptyset$. In this case $\alpha^{-1}(A) = \emptyset$. Therefore, $\kappa \dot{\diamond} \langle A, \alpha \rangle = \kappa * 2^F \diamond A = \kappa \diamond A$. So $\dot{\diamond}$ satisfies **I2**.

To prove the converse, suppose that $\dot{\diamond}$ satisfies **I1** and **I2**. Let $\circ$ be the belief evolution operator defined by $\diamond$ and $*$. Suppose that $(2^F \diamond A) \cap \alpha \neq \emptyset$. If follows that

$$\begin{aligned} \kappa \circ \langle A, \alpha \langle &= \kappa * \alpha^{-1} \diamond A && \text{(since } A \text{ and } \alpha \text{ are consistent)} \\ &= \kappa \dot{\diamond} \langle A, \alpha \rangle && \text{(by } \mathbf{I1}\text{)} \end{aligned}$$

Now suppose that $(2^F \diamond A) \cap \alpha = \emptyset$.

$$\begin{aligned} \kappa \circ \langle A, \alpha \langle &= \kappa * 2^F \diamond A && \text{(since } A \text{ and } \alpha \text{ are not consistent)} \\ &= \kappa \dot{\diamond} \langle A, \alpha \rangle && \text{(by } \mathbf{I2}\text{)} \end{aligned}$$

This completes the proof. □

We are interested in fixing a particular transition system T, then characterizing all belief evolution operators for T. Proposition 11 gives one simple characterization. Let $\diamond$ be the belief update operator corresponding to T. A combined belief change operator $\dot{\diamond}$ is a belief evolution operator for $\diamond$ if and only if there is a belief revision operator $*$ such that $\dot{\diamond}$ satisfies **I1** and **I2**. In the next section, we give a semantic characterization for the same class of operators.

3.7.2 Translated Systems of Spheres

In §1.4.1, we described the Grove characterization of AGM revision operators in terms of systems of spheres. In this section, we prove a representation result for belief evolution in terms of translated systems of spheres. Throughout this section, we assume a fixed transition system T defining an update operator $\diamond$.

Definition 24 *If $\mathcal{S}$ is a system of spheres centered on κ and A is an action, then define*

$$\mathcal{S} \diamond A = \{X \diamond A \mid X \in \mathcal{S}\}.$$

Note that $\mathcal{S} \diamond A$ is not generally a system of spheres: condition $S3$ may fail because $\mathcal{S} \diamond A$ need not contain $2^{\mathbf{F}}$. However, it is easy to verify that $\mathcal{S} \diamond A$ satisfies conditions $S1$, $S2$ and $S4$ with minimum element $\kappa \diamond A$. We remark that we could modify the definition of $\mathcal{S} \diamond A$ by adding an additional sphere containing all states, thereby guaranteeing the truth of condition $S3$. We do not take this step, however, because we want $\mathcal{S} \diamond A$ to explicitly exclude states that are not possible outcomes of the action A.

Let $\mathcal{S}$ be a system of spheres centered on κ and let A be an action symbol. Define the function $f_{\mathcal{S}\diamond A}$ as follows:

$$f_{\mathcal{S}\diamond A}(\alpha) \begin{cases} \alpha \cap c_{\mathcal{S}\diamond A}(\alpha) & \text{if } (2^F \diamond \bar{A}) \cap \alpha \neq \emptyset \\ \kappa \diamond A & \text{otherwise.} \end{cases}$$

As in the case of true systems of spheres, $c_{\mathcal{S}\diamond A}(\alpha)$ is the least sphere in $\mathcal{S} \diamond A$ intersecting α. We can associate an operator $\dot{\circ}$ with $\mathcal{S}$ as follows:

$$\kappa \dot{\circ} \langle A, \alpha \rangle = f_{\mathcal{S}\diamond A}(\alpha).$$

We will prove that the class of functions definable in this manner coincides exactly with the class of belief evolution operators. The following lemma is a useful tool. Informally, it states that the least sphere intersecting α in $\mathcal{S} \diamond A$ can be determined by finding the least sphere intersecting $\alpha^{-1}(A)$ in $\mathcal{S}$, and then applying A.

Lemma 1 *Let $\mathcal{S}$ be a system of spheres centered on κ. For any action A and any observation α, if $(2^F \diamond A) \cap \alpha \neq \emptyset$ then*

$$c_{\mathcal{S}\diamond A}(\alpha) = c_{\mathcal{S}}(\alpha^{-1}(A)) \diamond A.$$

Proof Suppose that $s \in c_{\mathcal{S}\diamond A}(\alpha)$, so $s = t \diamond A$ for some state t. Towards a contradiction, suppose that $t \notin c_{\mathcal{S}}(\alpha^{-1}(A))$. Then there is some sphere $S \in \mathcal{S}$ such that $S \cap \alpha^{-1}(A) \neq \emptyset$ and $t \notin S$. But then $S \diamond A \cap \alpha \neq \emptyset$ and $t \diamond A \notin S \diamond A$. This contradicts the fact that $s \in c_{\mathcal{S}\diamond A}(\alpha)$, because $c_{\mathcal{S}\diamond A}(\alpha)$ is the least sphere in $\mathcal{S} \diamond A$ intersecting α. The converse is similar. □

We now prove that every translated system of spheres defines a belief evolution operator.

Proposition 12 *Let $\mathcal{S}$ be a system of spheres centered on κ and let $\dot{\circ}$ be the combined belief change operator defined by $\mathcal{S}$. Then there is an AGM revision operator $*$ such that $\dot{\circ}$ satisfies* **I1** *and* **I2** *for $*$ and $\diamond$.*

Proof Let A be an action and let α be an observation. We need to define a revision operator. For any observation β, define $*$ as follows

$$\begin{aligned} \kappa * \beta &= \beta \cap c_{\mathcal{S}\diamond\lambda}(\beta) \\ &= \beta \cap c_{\mathcal{S}}(\beta). \end{aligned}$$

By Grove's representation result [39], $*$ is an AGM revision operator. We will prove that $\dot{\circ}$ satisfies **I1** and **I2** with respect to $*$.

Suppose that $(2^F \diamond A) \cap \alpha \neq \emptyset$, so

$$\kappa\dot{\circ}\langle A, \alpha\rangle = \alpha \cap c_{\mathcal{S}\diamond A}(\alpha).$$

We remark that, by definition,

$$\alpha = \alpha^{-1}(A) \diamond A.$$

By Lemma 1,

$$c_{\mathcal{S}\diamond A}(\alpha) = c_{\mathcal{S}}(\alpha^{-1}(A)) \diamond A.$$

Therefore

$$\begin{aligned} \kappa\dot{\circ}\langle A, \alpha\rangle &= [\alpha^{-1}(A) \diamond A] \cap [c_{\mathcal{S}}(\alpha^{-1}(A)) \diamond A] \\ &= [\alpha^{-1}(A) \cap c_{\mathcal{S}}(\alpha^{-1}(A))] \diamond A \end{aligned}$$

By definition of $*$, we have $\kappa * \alpha^{-1}(A) = \alpha^{-1}(A) \cap c_{\mathcal{S}}(\alpha^{-1}(A))$. It follows that

$$\kappa\dot{\circ}\langle A, \alpha\rangle = (\kappa * \alpha^{-1}(A)) \diamond A$$

which proves that **I1** holds.

If $(2^F \diamond A) \cap \alpha = \emptyset$, then by definition:

$$\kappa \dot{\circ} \langle A, \alpha \rangle = \kappa \diamond A.$$

Hence **I2** is satisfied. □

We now prove the converse.

Proposition 13 *Let $\dot{\circ}$ be an operator satisfying* **I1** *and* **I2** *for some AGM revision function $*$. Then for any fixed belief state κ, there is a system of spheres $\mathcal{S}$ centered on κ such that $\kappa \dot{\circ} \langle A, \alpha \rangle = f_{\mathcal{S} \diamond A}(\alpha)$ for all A and α.*

Proof By Grove's representation result, there is a system of spheres $\mathcal{S}$ centered on κ such that

$$\kappa * \alpha = f_{\mathcal{S}}(\alpha)$$

for all α. Fix a particular α and let A be an action symbol. Suppose that $(2^F \diamond A) \cap \alpha \neq \emptyset$. So, by **I2**:

$$\kappa \dot{\circ} \langle A, \alpha \rangle = \kappa * \alpha^{-1}(A) \diamond A.$$

By definition of $*$, this is equal to

$$f_{\mathcal{S}}(\alpha^{-1}(A)) \diamond A.$$

Replacing the $f_{\mathcal{S}}$ term by its value, we get

$$[\alpha^{-1}(A) \cap c_{\mathcal{S}}(\alpha^{-1}(A))] \diamond A.$$

Distributing the update by A, we get

$$[\alpha^{-1}(A) \diamond A] \cap [c_{\mathcal{S}}(\alpha^{-1}(A)) \diamond A].$$

By simplifying and applying Lemma 1, this gives

$$\alpha \cap c_{\mathcal{S} \diamond A}(\alpha)$$

which is what we wanted to show.

Now suppose that α $(2^F \diamond A) \cap \alpha = \emptyset$ then,

$$\begin{aligned} \kappa \dot{\diamond} \langle A, \alpha \rangle &= \kappa \diamond A \qquad \text{(by } \mathbf{I1}\text{)} \\ &= f_{S \diamond A}(\alpha) \qquad \text{(by definition)} \end{aligned}$$

This completes the proof. □

Hence, the class of belief evolution operators corresponds exactly with the class of functions defined by translated systems of spheres. We remark that our representation result is essentially a corollary of Grove's representation result for AGM revision. It would be straightforward to modify the representation result to allow multiple updates followed by a single revision: the initial system of spheres would simply be translated by several actions successively.

3.7.3 Non-Deterministic Action Effects

To this point, we have restricted attention to transition systems defining actions with deterministic effects. In this section, we look briefly at non-deterministic action effects and illustrate that translated systems of spheres can inform our treatment of belief change in this setting.

Note that the definition of $\diamond$ is not restricted to deterministic transition systems. If an action A has non-deterministic effects in a transition system T, then $\kappa \diamond A$ is still the set of *all* s' such that (s, A, s') is in T. As such, the set $\kappa \diamond A$ may be strictly larger than κ. We introduce a problematic example.

Example Consider an action domain involving a single action *toggle* and two fluents *LampOn*, *FuseBroken*. The effects of the toggle action are given by the transition system in Figure 3.3. Note that the effects are non-deterministic, because toggling the switch from the empty interpretation may either cause the lamp to turn on, or it may cause the fuse to break.

Now suppose that an agent believes the lamp is off and the fuse is unbroken. After toggling the switch, the agent observes that the lamp is still off. Formally, we need to determine $\kappa \circ \langle A, \alpha \rangle$ where

- $\kappa = \{\emptyset\}$

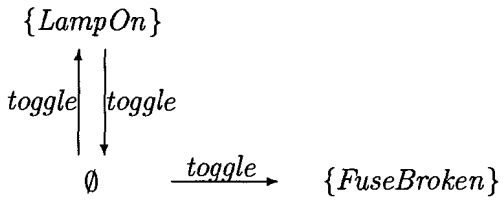

Figure 3.3: The Lamp Domain

- $A = toggle$
- $\alpha = \{\emptyset, \{FuseBroken\}\}$.

The final belief state is given by

$$\kappa * \alpha^{-1}(A) \diamond (A) = \emptyset \diamond A.$$

According to the non-deterministic effects of the *toggle* action, the final belief state is $\{\{LampOn\}, \{FuseBroken\}\}$.

Clearly the preceding example does not give the appropriate result. The agent observes that the lamp is still off, yet $\{LampOn\}$ is in the belief state. The problem is that the action effects are carried out after the revision, despite the fact that the observation α has given the agent information about which effect to choose. However, this problem only arises if we use the syntactic definition of belief evolution. If we consider the same example, and we let $\dot{\diamond}$ denote the combined belief change operator obtained from $\diamond$ and a system of spheres $\mathcal{S}$, then we get the following result:

$$\kappa \dot{\diamond} \langle A, \alpha \rangle = f_{\mathcal{S} \diamond A}(\alpha) = \{FuseBroken\}.$$

This example illustrates that the systems of spheres approach actually provides a better model of belief evolution in action domains involving non-deterministic actions. The problem with the syntactic definition of belief evolution is the following: observations are simply treated as new information about the initial belief state. If action effects are deterministic, this is a fair treatment of observations. However, if action effects are non-deterministic, then observations can serve a second purpose. In particular, observations help an agent

determine which effect has occurred after an action is executed. The systems of spheres approach to belief evolution is able to capture both interpretations of an observation.

Chapter 4

Applications of Belief Evolution

In this chapter, we look at some applications of belief evolution. First, we introduce a modal extension of the action language $\mathcal{A}$ and we define a semantics for the new action language in terms of belief evolution. We prove that the resulting action language is strictly more expressive than existing epistemic extensions of $\mathcal{A}$, and that it is able to give compact representations of action domains involving sensing actions. Since the effects of actions are given by belief evolution, the interaction between sensing actions and non-sensing actions is satisfactory from the perspective of the properties P1-P5.

The second application that we consider is the development of a solver for belief evolution based on the techniques of answer set planning. Towards this end, we introduce a topological revision operator in which revision can be reduced to path finding in a transition system. We illustrate that belief evolution under this revision operator can be computed by finding minimal length paths in a transition system. As such, if the underlying transition system is given by an action description in the action language $\mathcal{A}$, then we can use existing translations into answer set programming to compute the result of belief evolution. We describe a high-level procedure that can be used to determine the final belief state when an action is followed by an observation. This procedure can be applied to solve simple projection problems in the epistemic extension of $\mathcal{A}$.

The final application that we consider is the verification of cryptographic protocols. The general verification problem is beyond the scope of our formalism, because it involves the beliefs of multiple agents. However, we can use belief evolution operators to formalize how agents reason about so-called authentication tests, which are a common component of many authentication protocols.

4.1 A Modal Action Language

4.1.1 Motivation

The action language $\mathcal{A}$ is a simple high-level language for reasoning about the effects of actions. The basic language is suitable only for simple action domains, but it has been extended several times to address a wide range of problems [5, 6]. In this section, we illustrate that it is possible to increase the representational power of $\mathcal{A}$ without changing the action language itself. Instead, we look at extending the underlying propositional logic by adding modal operators. We consider the expressive power of the modal extension, and compare the framework with related work on epistemic extensions of $\mathcal{A}$.

As indicated in Chapter 1, there have been previous extensions of $\mathcal{A}$ that address knowledge by introducing new propositions for representing the effects of sensing actions [68, 86]. Basically these approaches focus on modeling dynamic knowledge about atomic facts. Lobo et. al. acknowledge that there are some situations in which a modal approach would be advantageous. For example, they suggest that a modal approach may provide a more natural framework for modeling situations in which introspective agents need to perform checks on the current knowledge state.

We suggest that adding a modal operator to $\mathcal{A}$ has some practical advantages over alternative approaches to the representation of knowledge or belief. In particular, by adding a modal operator, we obtain an action language that is immediately familiar and comprehensible to those with an elementary knowledge of modal logic. Moreover, using a modal operator is a natural way to represent nested beliefs in a multi-agent environment. Representing nested beliefs is important for some important application domains, such as cryptographic protocol verification [13].

For the present purpose, the most important feature of our epistemic extension of $\mathcal{A}$ is that the semantics respects the non-elementary interaction of revision and update. By contrast, existing epistemic extensions of $\mathcal{A}$ do not consider the influence that action histories may have on the epistemic effects of sensing actions.

4.1.2 Syntax

Let $\mathbf{A}$ be a fixed set of action symbols, let $\mathbf{F}$ be a fixed set of fluent symbols, and let $\mathcal{L}$ be the language of propositional modal logic over $\mathbf{F}$ with a single unary modal operator $\Box$. In

this section, we give the syntax of a new action language $\mathcal{A}_\Box$. To be precise, $\mathcal{A}_\Box$ is actually an *action description language* in the terminology of Lifschitz [64].

We want to extend $\mathcal{A}$ minimally to allow modal action effects. We remark that action effects in $\mathcal{A}$ are always literals; this restriction allows us to avoid dealing with disjunctive effects in the semantics. We will ensure that disjunctive modal effects are also prevented.

Definition 25 *A proposition of $\mathcal{A}_\Box$ is an expression of the form*

$$A \textbf{ causes } \phi \textbf{ if } F_1 \wedge \cdots \wedge F_p$$

where $A \in \mathbf{A}$, each F_i is a literal, and ϕ is either a literal or a formula of the form $\Box\psi$ for some non-modal formula ψ.

Notice that ψ need not be a literal; any non-modal formula can appear under the scope of a single $\Box$.

It is convenient to make the simplifying restriction that actions have either modal effects or non-modal effects, but not both. As such, we only allow action descriptions with the property that no action symbol occurs with both modal and non-modal effects. Restricting action descriptions in this manner simplifies the discussion without severely limiting expressive power. Since we are interested in an epistemic modality, it is natural to think of propositions of the form

$$O \textbf{ causes } \Box\phi \textbf{ if } F_1 \wedge \cdots \wedge F_p$$

as descriptions of sensing action effects. We use the terms *sensing action* and *non-sensing action* to refer to actions with modal and non-modal effects, respectively. The symbol O will range over sensing actions and the symbol A will range over non-sensing actions.

4.1.3 Epistemic Semantics

The semantics of $\mathcal{A}_\Box$ is defined by associating a transition function Φ_{AD} with every action description AD. We have previously proposed a generic modal semantics for $\mathcal{A}_\Box$ [45], but that approach is not appropriate for belief change. The problem with the generic semantics is that there is no underlying similarity relation on states, so the treatment of fallible initial beliefs is unsatisfactory. In this section, we present a new semantics for $\mathcal{A}_\Box$ that is intended specifically for belief change with respect to a fixed AGM revision operator. We define the new semantics in terms of belief evolution in order to ensure that the interaction between revision and update is handled appropriately.

Intuitively, we would like Φ_{AD} to take a belief state and an action sequence as arguments, and we would like it to return a new belief state. However, the effects of actions in $\mathcal{A}_\Box$ may have preconditions that depend on some distinguished "actual world." As such, we need to define Φ_{AD} with respect to *pointed belief states.*

Definition 26 *A pointed belief state is a pair $\langle s, \kappa \rangle$ where s is a state and κ is a non-empty belief state. If $s \in \kappa$, then $\langle s, \kappa \rangle$ is a pointed knowledge state.*

In the pointed belief state $\langle s, \kappa \rangle$, s represents the actual state of the world and κ represents the set of states that the underlying agent believes to be possible. Pointed belief states define an entailment relation for modal logic through a standard recursive definition. The key points of the definition are as follows.

- For any fluent symbol F, $\langle s, \kappa \rangle \models F \iff s \models F$.
- For any formula ϕ, $\langle s, \kappa \rangle \models \Box\phi \iff \langle s', \kappa \rangle \models \phi$ for every $s' \in \kappa$.

Negations and conjunctions are defined in the usual way. We remark that this semantics is equivalent to the modal logic $KD45$, which is the standard modal logic of belief. If we restrict attention to pointed knowledge states then we have $KT5$, which is the standard modal logic of knowledge.

We need to introduce some notation. Let O be a sensing action and let s be a state. Define $EFF(O, s)$ to be the conjunction of every formula ϕ that occurs in a proposition of the form

$$O \textbf{ causes } \Box\phi \textbf{ if } F_1 \wedge \cdots \wedge F_p$$

where $s \models F_1 \wedge \cdots \wedge F_p$. Define $O[s]$ to be the set of all models of $EFF(O, s)$. Using this notation, we can associate world views with sequences of action symbols.

Definition 27 *Let s be a state and let $Acts = \langle A_1, O_1, \ldots, A_n, O_n \rangle$ be an alternating sequence of non-sensing and sensing actions. Define $view(s, Acts) = \langle \bar{A}, \bar{\alpha} \rangle$ where*

1. $\bar{A} = \langle A_1, \ldots A_n \rangle$
2. *for each i, $\alpha_i = O_i[s \diamond A_1 \diamond \cdots \diamond A_i]$.*

So $view(s, Acts)$ is obtained by computing the sensing action effects at each time i, given that s is the initial world and the actions $A_1, \ldots, A_i$ have been executed. We remark that

$view(s, Acts)$ can be extended to non-alternating action sequences as well. Basically, if $Acts$ is a non-alternating sequence of actions, then we let $Acts'$ denote the shortest alternating sequence that can be obtained from $Acts$ by inserting null actions. Given this extended sequence, we define $view(s, Acts) = view(s, Acts')$. The details are straightforward.

We are now in a position to define the semantics of $\mathcal{A}_\Box$. Note that, for any action description AD, the non-modal portion of AD describes a transition system T which in turn defines an update operator $\diamond$. We refer to $\diamond$ as the update operator defined by AD. The following definition assumes a fixed underlying revision operator $*$.

Definition 28 *Let AD be an action description, let $\diamond$ be the update operator defined by AD and let $\circ$ be the belief evolution operator obtained from $\diamond$ and $*$. For every pointed belief state $\langle s, \kappa \rangle$ and every sequence $Acts = \langle A_1, O_1, \ldots, A_n, O_n \rangle$, define*

$$\Phi_{AD}(\langle s, \kappa \rangle, Acts) = \langle s', \kappa' \rangle$$

where

1. $s' = s \diamond A_1 \diamond \cdots \diamond A_n$
2. κ' *is the final belief state in* $\kappa \circ view(s, Acts)$.

Hence, the transition relation associated with AD returns a new pointed belief state. The new actual world is obtained by updating s by the non-sensing actions in $Acts$. The new belief state is obtained by belief evolution.

Again, the definition of Φ_{AD} can be extended to arbitrary action sequences by inserting a minimal number of null actions. Under this convention, the content of Definition 28 for action sequences of length 1 is as follows.

1. For a non-sensing action A: $\Phi_{AD}(\langle s, \kappa \rangle, A) = \langle s \diamond A, \kappa \diamond A \rangle$.
2. For a sensing action O: $\Phi_{AD}(\langle s, \kappa \rangle, O) = \langle s, \kappa * O[s] \rangle$.

The following example illustrates how to apply the definition in the context of a single action.

Example We represent a domain with a single agent inside a room with a window. Looking out the window allows the agent to determine if it is raining or not. This can be represented

by the action description AD containing the following propositions.

$$LookOutWindow \textbf{ causes } \Box(Rain) \textbf{ if } Rain$$
$$LookOutWindow \textbf{ causes } \Box(\neg Rain) \textbf{ if } \neg Rain.$$

Suppose that s is a state with $s \models Rain$ and let W_r denote the set of states in which $Rain$ is true.

By definition, $LookOutWindow[s] = W_r$. Given any belief state κ, we have

$$\Phi_{AD}(\langle s, \kappa \rangle, LookOutWindow) = \langle s, \kappa * W_r \rangle.$$

If $\kappa \subseteq W_r$, then we have belief expansion and we simply take the intersection. If $\kappa \cap W_r = \emptyset$, then the agent erroneously believed it was not raining initially, so we have belief revision.

Note that the preceding example only involves propositions of a particular form:

$$O \textbf{ causes } \Box\phi \textbf{ if } \phi.$$

Observations of this form can be understood to represent reliable observations. More generally, we have the following definition.

Definition 29 *An action description AD is reliable if $F_1 \wedge \cdots \wedge F_p \models \phi$ for every modal effect proposition in AD with the form*

$$O \textbf{ causes } \Box\phi \textbf{ if } F_1 \wedge \cdots \wedge F_p.$$

Reliable action descriptions have the following property: if a sensing action A causes an agent to believe ϕ, then ϕ must hold in the actual state.

Clearly, the action description in the rain example is reliable. By contrast, the action description would not be reliable if it contained the proposition

$$LookOutWindow \textbf{ causes } \Box(Rain).$$

This proposition asserts that looking out the window causes the agent to believe it is raining, whether or not it is actually raining.

The following proposition formalizes the fact that reliable observations lead to reasonable conclusions.

Proposition 14 *Let AD be a reliable action description and let $\langle s, \kappa \rangle$ be a pointed knowledge state. For any action sequence Acts, it follows that $\Phi_{AD}(\langle s, \kappa \rangle, Acts)$ is a pointed knowledge state.*

Proof Assume without loss of generality that $Acts = \langle A_1, O_1, \ldots, A_n, O_n \rangle$. Let

$$view(s, Acts) = \langle \bar{A}, \bar{\alpha} \rangle,$$

and let

$$\Phi_{AD}(\langle s, \kappa \rangle, Acts) = \langle s', \kappa' \rangle.$$

We need to show that $s' \in \kappa'$. Equivalently, we need to show that $s \diamond A_1 \diamond \cdots \diamond A_n$ is in the final belief state in $\kappa \circ \langle \bar{A}, \bar{\alpha} \rangle$.

Since AD is reliable, it follows that $t \in O[t]$ for any state t and any sensing action O. Hence, for all i, we have $s \diamond \bar{A}_i \in \alpha_i$. But then

$$s \in \bigcap_i \alpha_i^{-1}(\bar{A}_i).$$

Since $s \in \kappa$, it follows that

$$s \in \kappa * \bigcap_i \alpha_i^{-1}(\bar{A}_i)$$

is the initial belief state in the belief evolution $\kappa \circ \langle \bar{A}, \bar{\alpha} \rangle$. Therefore, $s \diamond A_1 \diamond \cdots \diamond A_n$ is in the final state. □

Hence, if an agent has correct knowledge of the world, then the conclusions drawn from reliable observations must also be correct. Reliable action descriptions can be understood to describe infallible sensing actions.

4.1.4 Representing Existing Epistemic Action Languages

In Chapter 1, we briefly outlined two epistemic extensions of $\mathcal{A}$ that have been proposed in the literature[68, 86]. In this section, we demonstrate that $\mathcal{A}_\Box$ subsumes both of these extensions. In the next section, we will show that $\mathcal{A}_\Box$ is able to represent some problems that are not representable in these existing action languages, which indicates that the subsumption is strict.

Let O be a sensing action. Recall from Chapter 1 that $\mathcal{A}_B$ is the language obtained by extending $\mathcal{A}$ with propositions of the form

$$O \textbf{ determines } F.$$

The semantics is defined in terms of pointed belief states. In particular, for sensing actions, Φ^B_{AD} is defined such that

$$\Phi^B_{AD}(\langle s, \kappa\rangle, O) = \langle s', \kappa'\rangle$$

if and only if $s' = s$ and κ' is the subset of κ that agrees with s on the truth value assigned to F.

We define a translation σ from $\mathcal{A}_B$ to $\mathcal{A}_\Box$. Given AD, we construct $\sigma(AD)$ by replacing each proposition of the form

$$O \textbf{ determines } F$$

with two propositions

$$O \textbf{ causes } \Box F \textbf{ if } F$$
$$O \textbf{ causes } \Box\neg F \textbf{ if } \neg F$$

We have the following result.

Proposition 15 *Let AD be a set of $\mathcal{A}_B$ propositions, let O be a sensing action and let $\langle s, \kappa\rangle$ be a pointed knowledge state. Then $\Phi^B_{AD}(\langle s, \kappa\rangle, O) = \Phi_{\sigma(AD)}(\langle s, \kappa\rangle, O)$.*

Proof Let F be a fluent symbol that occurs in AD in a proposition of the form

$$O \textbf{ determines } F$$

If $s \models F$, then it follows that $O[s]$ is the set of interpretations I such that $I \models F$. Hence,

$$\begin{aligned}\Phi^B_{AD}(\langle s, \kappa\rangle, O) &= \langle s, \kappa \cap O[s]\rangle \\ &= \langle s, \kappa * O[s]\rangle \quad [\text{since } \kappa \cap O[s] \neq \emptyset] \\ &= \Phi_{\sigma(AD)}(\langle s, \kappa\rangle, O)\end{aligned}$$

By the symmetry of $\sigma(AD)$, we get the same result if $s \models \neg F$. □

We remark that Proposition 15 does not hold for pointed belief states in general. The underlying assumption in $\mathcal{A}_B$ is that the agent's knowledge is correct, but incomplete. This assumption is captured by restricting attention to pointed knowledge states.

The translation from $\mathcal{A}_L$ to $\mathcal{A}_\Box$ is similar. The main difference is that $\mathcal{A}_L$ allows sensing actions with conditional effects. Let O be a sensing action that occurs in a proposition of the form

$$O \textbf{ causes to know } F \textbf{ if } P.$$

The semantics of $\mathcal{A}_L$ specifies that $\Phi^L_{AD}(\kappa, O, \kappa')$ just in case κ' is non-empty and is defined by one of the following conditions:

1. $\{s \in \kappa \mid s \models F \wedge P\}$
2. $\{s \in \kappa \mid s \models \neg F \wedge P\}$
3. $\{s \in \kappa \mid s \models \neg P\}$.

Again, there is an underlying assumption that knowledge is correct in $\mathcal{A}_L$ representations. This is embodied in the semantics by restricting the outcome of a sensing action to be a non-empty subset of the initial belief state. Note that we have simplified the discussion by defining the semantics in terms of belief states, whereas the original semantics involves sets of belief states. This is not a significant simplification; the translation that we provide below could easily be reformulated to deal with sets of belief states. We remark also that we have only considered the deterministic portion of $\mathcal{A}_L$. Ontic action effects in $\mathcal{A}_\Box$ are given by standard $\mathcal{A}$ propositions, so we are not able to represent non-deterministic effects.

We now give the translation from $\mathcal{A}_L$ to $\mathcal{A}_\Box$. Let AD be an action description in $\mathcal{A}_L$. The $\mathcal{A}_\Box$ action description $\tau(AD)$ is obtained from AD by replacing every sensing proposition with potential sensing effect F and knowledge precondition P by the following propositions:

$$O \textbf{ causes } \Box(F \wedge P) \textbf{ if } F \wedge P$$
$$O \textbf{ causes } \Box(\neg F \wedge P) \textbf{ if } \neg F \wedge P$$
$$O \textbf{ causes } \Box\neg P \textbf{ if } \neg P.$$

The following proposition illustrates the correspondence with $\mathcal{A}_\Box$.

Proposition 16 *Let AD be an $\mathcal{A}_L$ action description, let O be a sensing action in AD and let κ be a belief state. Then $\Phi^L_{AD}(\kappa, O, \kappa')$ if and only if there is some $s \in \kappa$ such that $\Phi_{\tau(AD)}(\langle s, \kappa\rangle, O) = \langle s, \kappa'\rangle$.*

Proof Let ϕ be an arbitrary formula. With respect to $\tau(AD)$, we make the following observations.

1. If $s \models F \wedge P$ then $O[s] = |F \wedge P|$.
2. If $s \models \neg F \wedge P$ then $O[s] = |\neg F \wedge P|$.
3. If $s \models \neg P$ then $O[s] = |\neg P|$.

Suppose that $\Phi^L_{AD}(\kappa, O, \kappa')$, so there are three possible definitions for κ'. Consider the case where $\kappa' = \{s \in \kappa \mid s \models F \wedge P\}$. Since $\kappa' \neq \emptyset$, it follows that there is some $s \in \kappa$ such that $s \models F \wedge P$. But then

$$\begin{aligned} \kappa' &= \kappa \cap O[s] \\ &= \kappa * O[s] \\ &= \Phi_{\tau(AD)}(\langle s, \kappa \rangle, O) \end{aligned}$$

The same argument holds for the other two possibilities for κ'.

The converse holds because every $s \in \kappa$ must satisfy exactly one of the formulas $F \wedge P$, $\neg F \wedge P$, and $\neg P$. $\square$

Note that $\mathcal{A}_L$ differs from $\mathcal{A}_B$ and $\mathcal{A}_\Box$ in that there is no distinguished state representing the actual world. Proposition 16 illustrates that $\mathcal{A}_L$ action descriptions are interpreted disjunctively, by determining all possible outcomes κ' under the assumption that the actual world is in κ.

4.1.5 Increased Expressive Power

We have illustrated that $\mathcal{A}_B$ and the deterministic portion of $\mathcal{A}_L$ can be naturally embedded in the language $\mathcal{A}_\Box$. In this section, we revisit two previous examples to illustrate that $\mathcal{A}_\Box$ is actually more expressive.

Both $\mathcal{A}_B$ and $\mathcal{A}_L$ are intended to be used for action domains where the initial belief state is correct and every observation is reliable. By definition, for any sensing action O and any action description AD, it is the case that $\Phi^B_{AD}(\langle s, \kappa \rangle, O) = \langle s, \kappa' \rangle$ implies $\kappa' \subseteq \kappa$. This property states that the belief state following a sensing action is always a subset of the initial belief state. The same property holds for Φ^L_{AD}.

By contrast, consider the earlier example in which an agent may look out the window to determine if it is raining or not. Suppose that the initial belief state is the set of all non-raining worlds, but the actual state s is in the set W_r of raining worlds. Intuitively,

after looking out the window, the new belief state should be a non-empty subset of W_r. However, due to the property mentioned above, this can not be the case in $\mathcal{A}_B$ or $\mathcal{A}_L$. We have already seen that this problem can be represented in $\mathcal{A}_\Box$.

It would be relatively straightforward to modify $\mathcal{A}_B$ or $\mathcal{A}_L$ to allow incorrect beliefs. For example, $\mathcal{A}_B$ could be extended by allowing pointed belief states $\langle s, \kappa \rangle$ where $s \notin \kappa$, and propositions of the form

$$O \textbf{ determines } F$$

could be replaced by propositions of the form

$$O \textbf{ determines } F \textbf{ is } (true/false).$$

However, if the effects of sensing and non-sensing actions are simply performed successively, then the resulting language will still give an unsatisfactory treatment of litmus-type problems. The advantage of $\mathcal{A}_\Box$ is that the semantics is based on belief evolution, so the interaction between update and revision can be non-elementary. We illustrate by revisiting the litmus paper problem. We have already seen that the belief change that occurs in the litmus paper problem can not be captured by determining the effects of dipping and looking at the paper in succession. Hence, neither $\mathcal{A}_B$ nor $\mathcal{A}_L$ provides a suitable representation of this problem.

Example (revisited) Consider the extended litmus paper example. This domain can be represented by the action description AD containing the following propositions.

$$dip \textbf{ causes } Red \textbf{ if } Litmus \wedge Acid$$
$$dip \textbf{ causes } Blue \textbf{ if } Litmus \wedge \neg Acid$$
$$look \textbf{ causes } \Box Red \textbf{ if } Red$$
$$look \textbf{ causes } \Box Blue \textbf{ if } Blue$$

Suppose that the actual state is $s = \emptyset$, so the paper is not litmus paper and the beaker contains a base. As previously, the initial belief state is

$$\kappa = \{\{Litmus\}, \{Litmus, Acid\}\}.$$

If the agent dips the paper and then observes that it is still white, then the new belief state is given by

$$\Phi_{AD}(\langle s, \kappa \rangle, \langle dip, look \rangle) = \langle \emptyset, \kappa' \rangle$$

where κ' is the final belief state in $\kappa \circ \langle dip, look \rangle$. If we use the Hamming distance to define the underlying revision operator then we get the final belief state $\{\emptyset, \{Acid\}\}$, which we have previously suggested is the intuitively correct solution.

The litmus paper example is just one natural action domain in which the interaction between sensing effects and non-sensing effects must be considered. Of the three epistemic action languages we have considered, $\mathcal{A}_\Box$ is the only one that respects the interaction properties P1-P5.

4.2 Implementing a Belief Evolution Solver

In this section, we are interested in illustrating how we can implement a belief evolution solver by translating into answer set programming. Rather than assuming an underlying transition system T, we assume an underlying action description in $\mathcal{A}$. Starting with an action description is convenient, since it allows us to base our work on existing translations into logic programming.

We proceed as follows. First, we define a revision operator based on path length. Next, we introduce an informal procedure that can be used to solve belief evolution problems with respect to this operator. We then present a translation from $\mathcal{A}$ action descriptions to extended logic programs, where answer sets correspond to solutions to belief evolution problems. The results in this section are intended as a simple proof of concept, and we do not focus on the details of the proposed implementation.

4.2.1 Topological Revision Operators

We introduce a new class of AGM revision operators that is defined in terms of path length in a transition system. Let $T = \langle S, R \rangle$ be a transition system, and let $\kappa \subseteq S$. Assume that every state in S is accessible by a finite path from κ. Let S_i be the set of all states in S that are reachable from κ by a path of length at most i. Let $\mathcal{S}_T = \{S_i \mid 0 \leq i\}$. It is easy to demonstrate that $\mathcal{S}_T$ is a system of spheres centered on κ. Let $*_T$ denote the revision operator defined by $\mathcal{S}_T$; we refer to this as the *topological revision operator* defined by T.

Example Suppose that an agent has 10 marbles labeled with the digits 0–9 along with a bag that may hold any number of marbles. Marbles may be added or removed from the

bag, one at a time. Initially, the agent believes that the bag contains marbles 0–4. However, after weighing the bag, the agent comes to believe that the bag contains 7 marbles. We can represent the appropriate belief change in this context in terms of a transition system and a topological revision operator.

For each marble i, there are two actions: $Add(i)$ and $Remove(i)$. There are 10 fluent symbols $InBag(i)$ indicating if marble i is in the bag or not. The transition system T giving the action effects is the transition system described by the action description containing the following propositions for each $i \leq 9$:

$$Add(i) \textbf{ causes } InBag(i)$$
$$Remove(i) \textbf{ causes } \neg InBag(i).$$

The initial belief state κ is the singleton set $\{s_0\}$ where s_0 is the state satisfying the following condition:

$$s_0 \models InBag(i) \iff 0 \leq i \leq 4.$$

The observation that the bag contains 7 marbles is represented by the set of worlds α where $s \in \alpha$ if and only if $s \models InBag(i)$ for exactly 7 distinct values of i.

To determine $\kappa *_T \alpha$, we need to identify which worlds in α are reachable from κ by a minimal length path. It is clear that the shortest path from κ to α is obtained by performing 2 adding actions. Hence $\kappa *_T \alpha$ is the set of states s such that $s \models InBag(i)$ for every $i \leq 4$ and for exactly 2 values of $i > 4$.

Under the topological revision operator $*_T$, the degree of similarity between two states is given by the number of actions required to get from one to the other. Revision by α essentially involves postulating a minimal sequence of initial actions that explain the observation α. As such, topological revision operators are useful for action domains in which an agent cannot be certain about the exogenous actions that have occurred at an earlier point in time. One nice feature of this approach is that the revision operator is defined by the transition system; we do not require any independent notion of similarity. Also, there is a straightforward extension of topological revision in which we can represent the plausibility of actions by assigning weights to edges in the transition system. We do not require such an extension for our present purposes.

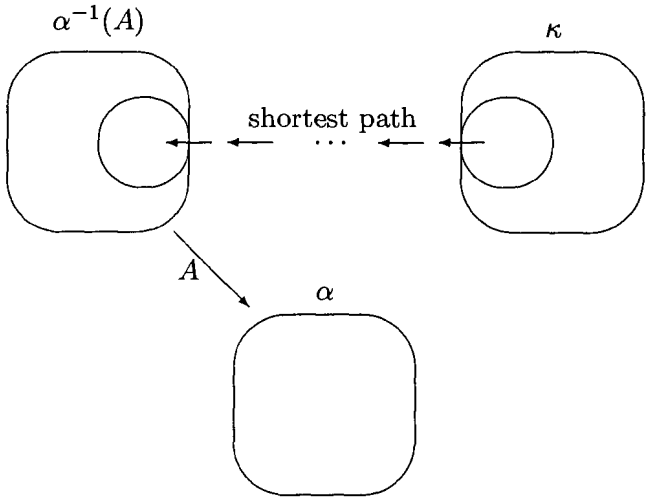

Figure 4.1: Visualizing Topological Evolution

4.2.2 Belief Evolution Under Topological Revision

In this section, we illustrate that belief evolution under topological revision can be reduced to finding shortest paths in the underlying transition system. We start by considering trajectories of length 1. Let κ denote a belief state, let A denote an action symbol, and let α denote an observation. We are interested in determining

$$\kappa \circ \langle\langle A \rangle, \langle \alpha \rangle\rangle.$$

Recall that, informally, this corresponds to the iterated belief change $\kappa \diamond A * \alpha$ and it is given by the following belief trajectory

$$\langle \kappa * \alpha^{-1}(A), \kappa * \alpha^{-1}(A) \diamond A \rangle.$$

Figure 4.1 illustrates how this is calculated with the topological revision function. The figure shows a large box representing $\alpha^{-1}(A)$; these are the states that can reach α by executing the action A. The circle inside $\alpha^{-1}(A)$ represents the subset that is minimally distant from κ, which in this context means the elements that can be reached from κ by a minimal length path. In other words, the circle inside $\alpha^{-1}(A)$ represents $\kappa * \alpha^{-1}(A)$. This gives a simple procedure for computing $\kappa \circ \langle\langle A \rangle, \langle \alpha \rangle\rangle$.

1. Determine $\alpha^{-1}(A)$.
2. Let $PATH$ denote the set of shortest paths from κ to $\alpha^{-1}(A)$.
3. Let κ_0 be the set of terminal nodes on paths in $PATH$.
4. Let $\kappa_1 = \kappa_0 \diamond A$.

Clearly $\kappa \circ \langle\langle A \rangle, \langle \alpha \rangle\rangle = \langle \kappa_0, \kappa_1 \rangle$. Hence, this procedure allows us to compute the outcome of belief evolution for trajectories of length 1.

If we consider action trajectories and observation trajectories of fixed finite length, the procedure can be generalized as follows. Let $\bar{A}$ be an action trajectory of length n and let $\bar{\alpha}$ be an observation trajectory of length n.

1. Determine $\alpha_{PRE} = \bigcap_i \alpha_i^{-1}(A_1, \ldots, A_i)$.
2. Let $PATH$ denote the set of shortest paths from κ to α_{PRE}.
3. Let κ_0 be the set of terminal nodes on paths in $PATH$.
4. For $i \geq 1$, $\kappa_i = \kappa_0 \diamond A_1 \diamond \ldots A_i$.

Steps 1, 3 and 4 are straightforward. In order to implement a solver for belief evolution under topological revision, we need some mechanism for determining the set of shortest paths from κ to α_{PRE}.

4.2.3 Translation to Answer Set Programming

We have seen that solving belief evolution problems under topological revision involves path-finding in the underlying transition system. We illustrate how this process can be automated by using the techniques of answer set planning. Answer set planning refers to the approach to planning in which a problem is translated into a logic program where the answer sets correspond to plans [66]. Many action languages have been translated into extended logic programs for answer set planning. We demonstrate how one existing translation can be modified for our purposes.

We need a translation from $\mathcal{A}$-descriptions to extended logic programs. Our translation from $\mathcal{A}$ is obtained by modifying a well known translation from $\mathcal{C}$ [66]. Let AD be an action description in the action language $\mathcal{A}$. For any natural number n, we define an associated

logic program $\tau_n(AD)$ with the property that answer sets for $\tau_n(AD)$ correspond to paths of length n in the transition system T described by AD. The language of $\tau_n(AD)$ consists of two disjoint classes of atoms, defined as follows. For each $i \leq n$ and each $F \in \mathbf{F}$, the language of $\tau_n(AD)$ contains an atom $F(i)$. For each $i < n$ and each $A \in \mathbf{A}$, the language of $\tau_n(AD)$ contains an atom $A(i)$. The logic program $\tau_n(AD)$ consists of the following rules:

1. for every proposition of the form

$$A \textbf{ causes } F \textbf{ if } F_1 \wedge \cdots \wedge F_p$$

in AD, the rules

$$F(i+1) \leftarrow A(i), F_1(i), \ldots, F_p(i)$$

for every $i < n$

2. the rules

$$\neg B \leftarrow not\ B$$
$$B \leftarrow not\ \neg B$$

where B is either an action atom or B is $F(0)$ for some fluent F

3. the rules

$$F(i+1) \leftarrow not\ \neg F(i+1), F(i)$$
$$\neg F(i+1) \leftarrow not\ F(i+1), \neg F(i)$$

for every fluent symbol F and $i < n$

4. the rules

$$\neg A_1(i) \leftarrow A_2(i)$$

for every $i < n$ and every pair of distinct action symbols A_1, A_2.

The first two sets of rules are taken directly from Lifschitz and Turner's translation of $\mathcal{C}$[66]. The set of rules given by (1) encodes the effects of actions and the set of rules given by (2) forces exactly one of each complementary pair B, $\neg B$ to be true. The rules given by (3) and (4) have been added to capture the distinct features of $\mathcal{A}$. In particular, (3) states that all fluents are inertial, and (4) states that at most one action occurs at each point in time. The following proposition restates Lifschitz and Turner's main result for our translation.

Proposition 17 *A complete set X is an answer set for $\tau_n(AD)$ if and only if it has the form*

$$\bigcup_{i=0}^{n}\{F(i) \mid F \in s_i\} \cup \bigcup_{i=0}^{n-1}\{A(i) \mid A = A_i\}$$

for some path $\langle s_0, A_0, s_1, \ldots, A_{n-1}, s_n\rangle$ in the transition system described by AD.

Proof It is sufficient to note that $\mathcal{A}$ is equivalent to the restriction of $\mathcal{C}$ in which there are no static laws, every fluent is inertial, and actions are non-concurrent. The result then follows from the main result of [66]. □

Hence every answer set for $\tau_n(AD)$ corresponds to a path in the transition system.

For the purpose of planning, it is useful to add a few rules to $\tau_n(AD)$ that restrict the admissible answer sets. In particular, we would like to introduce a formula K representing the initial beliefs of an agent. In order to simplify the discussion, we restrict K to be a conjunction of literals. Let K be the following conjunction of literals:

$$K = K_1 \wedge \cdots \wedge K_p$$

We can now extend the translation τ_n so that it takes two arguments: an action description AD and a formula K. Define $\tau_n(AD, K)$ to be the logic program obtained by adding the following rules to $\tau_n(AD)$:

$$K_1(0), \ldots, K_p(0)$$

It is easy to see that the answer sets for $\tau_n(AD, K)$ correspond to all paths of length n which start in a state where K is true.

We need to make two assumptions about the underlying action signature AD. First, we assume that the associated transition system T has the property that every state in T is reachable from every other state. Second, for any literal L, we assume that there is at most one proposition in AD of the form

$$A \textbf{ causes } L \textbf{ if } F_1 \wedge \cdots \wedge F_p.$$

This assumption rules out the so-called *similar* propositions defined in [33] and it facilitates the specification of effect preconditions.

Let $\diamond$ denote the update operator defined by AD, let $*$ denote the corresponding topological revision operator, and let $\circ$ denote the corresponding belief evolution operator. Let

K be a conjunction of literals representing the initial belief state, let A be an action symbol and let L be a literal representing an observation. We are interested in using answer sets to determine the final belief state in the following belief evolution:

$$|K| \circ \langle A, |L| \rangle.$$

Note that, as a special case, we also get a solver for topological belief revision.

We need to introduce some notation. For any literal L, if L occurs in a proposition of the form

$$A \textbf{ causes } L \textbf{ if } F_1 \wedge \cdots \wedge F_p,$$

then define $PRE(A, L) = F_1 \wedge \cdots \wedge F_p$, and define $PRE(A, L) = \bot$ otherwise. Hence $PRE(A, L)$ is a formula with the property that, if $PRE(A, L)$ is true then L will be true after executing A. Let $\bar{L}$ denote the complementary literal to L. It follows that $PRE(A, \bar{L})$ is a formula with the property that, if $PRE(A, \bar{L})$ is true then L will be false after executing A.

Proposition 18 *Let AD be an $\mathcal{A}$ action description with corresponding update operator $\diamond$. If s is a state in the transition system defined by AD, L is a literal, and A is an action symbol, then*

$$s \diamond A \models L \iff s \models PRE(A, L) \vee (L \wedge PRE(A, \bar{L})).$$

Proof Follows from the semantics of $\mathcal{A}$, since every fluent is inertial. □

Using the notation introduced in the definition of belief evolution, Proposition 18 states that

$$|L|^{-1}(A) = |PRE(A, L) \vee (L \wedge PRE(A, \bar{L}))|.$$

We are now in a position to give a basic procedure for the implementation of a belief evolution solver. We define the procedure informally.

evol(K,A,f)
Inputs: $K \in \bigwedge Lits$, $A \in \mathbf{A}$, $L \in Lits$
Output: $\langle \kappa_0, \kappa_1 \rangle \in 2^S \times 2^F$
Procedure:
1. Set $n = 1$.
2. Determine all answer sets for $\tau_n(AD, K)$.
3. Let $PATH$ be the corresponding set of paths.

4. Remove all paths where the final state fails to satisfy $PRE(A, L) \vee (L \wedge PRE(A, \bar{L}))$.
   (a) If $PATH = \emptyset$, set $n = n + 1$ and goto 2.
   (b) If $PATH \neq \emptyset$, then continue.
5. Let κ_0 denote the set of final states in $PATH$.
6. Let $\kappa_1 = \{s \diamond A \mid s \in \kappa_0\}$.
7. Return $\langle \kappa_0, \kappa_1 \rangle$.

We prove a correctness result.

Proposition 19 *If K is a conjunction of literals, $A \in \mathbf{A}$, and $L \in Lits$, then* $\mathbf{evol}(K, A, L) = |K| \circ \langle A, |L| \rangle$.

Proof It is sufficient to show that $\kappa_0 = |K| * |L|^{-1}(A)$.

The answer sets of $\tau_n(AD, K)$ correspond to paths of length n starting at K-states. Let n be the smallest natural number such that $PATH$ is non-empty. So n is the length of the shortest path from $|K|$ to $|PRE(A, L) \vee (L \wedge PRE(A, \bar{L})|$. Therefore, $PATH$ is the collection of minimal length paths from $|K|$ to $|L|^{-1}(A)$. This means that the elements of κ_0 are precisely the states in $|L|^{-1}(A)$ that can be reached from $|K|$ by a minimal length path. Equivalently, $\kappa_0 = |K| * |L|^{-1}(A)$. □

Hence, we have given a procedure that returns the result of belief evolution under some restrictive conditions. There are two computational problems that need to be addressed in the procedure. First, we need to find all answer sets for a given logic program at step 2; this can be accomplished by using an existing answer set solver such as smodels [77] or DLV [26]. The second computational task involves checking if each final state entails $|L|^{-1}(A)$. We remark that the second task could be avoided by moving to disjunctive logic programming, because $PRE(A, \bar{L})$ can be expressed by employing a disjunctive rule. The main limitation of our approach is that it fixes a specific revision operator based on path length.

We have only considered the case where the observation is given by a literal L. This restriction allows us to use the propositions in AD to give a relatively simple characterization of $|L|^{-1}(A)$. However, in principle, our procedure will work for observations given by any formula. In particular, given a state s and a formula ϕ, let $\phi(s)$ denote the conjunction of every fluent that is true in s together with the negation of every fluent that is false in s. It

is easy to see that $|\phi|^{-1}(A)$ can be defined as follows:

$$|O|^{-1}(A) = \bigvee\{\phi(s) \mid s \diamond A \models O\}.$$

However, unless the underlying transition system is very small, this approach will not be useful for practical examples.

A *projection problem* is a problem in which a series of actions *Acts* is given, and we are asked to determine which fluents will be true after executing *Acts*. An epistemic projection problem is similar, except that it asks which fluents will be believed after *Acts*. Using our basic approach to solving belief evolution problems, we can solve epistemic projection problems for $\mathcal{A}_\Box$. Let AD be an action description in $\mathcal{A}_\Box$ satisfying the following conditions:

1. the effect of every modal proposition has the form $\Box L$ for some literal L
2. the precondition of every modal proposition is empty.

In this case, if $\Box L$ is the modal effect of the sensing action O, then

$$\Phi_{AD}(\langle s, |K|\rangle, \langle A, O\rangle) = \langle s \diamond A, |K| \circ \langle A, |L|\rangle\rangle.$$

The actual state can be computed by the standard translation from $\mathcal{A}$ into logic programming, and the belief state can be computed as above.

4.3 Cryptographic Protocol Verification

4.3.1 Motivation

Cryptographic protocols are structured sequences of messages used with cryptographic algorithms to relay secure messages in a hostile environment. Even if we assume that our cryptosystem is perfect, communication may be compromised if the protocol is poorly designed. Checking a protocol for potential breeches is very difficult by hand, so formal methods are employed for verification [73]. Logical methods have proven to be very useful in the design and analysis of cryptographic protocols, starting with the pioneering work on BAN logic [13]. The basic idea behind BAN is to use a modal logic to represent the beliefs of each principle, and then formalize security goals as statements about the beliefs of protocol participants. In this framework, protocol goals can be established by proving that they are logical consequences of some underlying set of assumptions.

One serious problem with BAN logic is the fact that there is no agreed upon semantics[2]. Without a formal semantics, the logical stature of BAN is dubious. Several different semantics have been proposed [1, 9, 91], and new protocol logics have been introduced based on the standard semantics for epistemic logic [92]. We remark that such logics have typically focused on the representation of belief, with comparatively little emphasis on belief change. Belief change is normally addressed by introducing some ad hoc axioms or rules of inference describing specific instances of belief change. However, since changing beliefs are a fundamental problem in protocol logics, we suggest that a more principled approach to belief change would be beneficial. Belief evolution operators provide a natural framework for reasoning about protocol goals, because protocols involve belief change in the context of both ontic and epistemic actions.

4.3.2 Authentication Tests

An authentication protocol is used to ensure that all parties in a communication session know with whom they are communicating. Authentication protocols may have additional goals as well, such as establishing a shared key for communication [91]. Typically, an authentication protocol is composed of several *authentication tests*. Using the vocabulary of [40], an outgoing authentication test is an exchange where a message M is sent in encrypted form, and then a future message is received indicating comprehension of M. We will illustrate with an example, but first we need to introduce some notation.

If P and Q denote principals in a communication session, then

$$P \rightarrow Q : X$$

means that principal P sends the message X to principal Q. A symmetric key that is shared by principals P and Q will be denoted by K_{pq}. If the message X is encrypted with the key K, then we write $\{X\}_K$. Finally, we let N_p denote a nonce generated by the principal P. A nonce is simply a random number that is generated by a principal during a communication session.

The standard model for cryptographic protocol analysis assumes that an intruder can read every message that is sent, and the intruder may choose to prevent messages from reaching the desired recipient. Moreover, when a message is received, it is assumed that the sender is always unknown. The only way that P can be certain Q sent a particular message is if the message contains information that is only available to Q. It is assumed

that cryptography is strong in that encrypted messages can not be unencrypted during a protocol run without the proper key.

The following simple protocol is executed by agent P in order to determine if agent Q is alive in the communication session. The protocol involves a single outgoing authentication test.

The Challenge-Response Protocol

1. $P \rightarrow Q : \{N_p\}_{K_{pq}}$
2. $Q \rightarrow P : N_p$

In this protocol, P generates a random number and encrypts it with the shared key K_{pq} before sending it to Q. Informally, if this message is intercepted by an intruder, it will not be possible for the intruder to determine the value N_p. So if P receives the message N_p, then it is natural to conclude that Q must have decrypted the original message. This establishes that Q is alive on the network.

Establishing the correctness of this protocol requires a model of belief change. At the time that the first message is sent, P need not believe that Q is alive. In order to prove that the protocol is correct, we need to establish that P's beliefs change by the end of the protocol. In BAN logic, changing beliefs are modeled by introducing new rules of inference. For example, BAN logic handles authentication tests by introducing the following rule of inference:

$$\frac{P \textit{ believes } P \overset{K}{\longleftrightarrow} Q \qquad P \textit{ received } \{X\}_K}{P \textit{ believes } Q \textit{ said } X}.$$

The notation $P \overset{K}{\longleftrightarrow} Q$ is used to state that K is a key shared only by P and Q. So, if P believes that K is a key shared with Q and P receives the a message encrypted with K, then P concludes that Q sent the message.

The following attack illustrates that there is a problem with the Challenge-Response protocol.

An Attack on the Challenge-Response Protocol

1. $P \rightarrow I_Q : \{N_p\}_{K_{pq}}$
   1'. $I_Q \rightarrow P : \{N_p\}_{K_{pq}}$
   2'. $P \rightarrow I_Q : N_p$
2. $I_Q \rightarrow P : N_p$

In this attack, I_Q intercepts the original message and then initiates a new protocol run by sending it back to P. After P receives the message encrypted with K_{pq}, then P follows the protocol and returns the decrypted nonce. At the last step, I_Q sends the same decrypted nonce to P. Note that, at the conclusion of the protocol, Q has not sent any messages. Hence P has no assurance that Q is actually alive on the network, which was the stated goal of the protocol.

Problems of this sort are handled in an ad hoc manner in BAN logic. In particular, it is simply assumed that an agent can recognize the messages that they have sent. Under this assumption, the preceding attack can not occur. However, in real world applications, this assumption is often unjustified. We propose that a more flexible approach to protocol verification can be defined by introducing belief change operators rather than ad hoc assumptions and rules of inference.

4.3.3 Incorporating Belief Change

The intuition behind an outgoing authentication test is that the interpretation of a received message is dependent upon the messages that have been sent previously. In particular, if an agent P receives a message X, then P should believe that the actual history of the world is one in which it is possible to receive the message X. In the challenge response protocol, when P receives the response N_p, it is not reasonable to conclude that Q decrypted $\{N_p\}_{K_{pq}}$. The strongest conclusion that P should draw is that either Q decrypted $\{N_p\}_{K_{pq}}$ or else P decrypted it. In this section, we illustrate how this kind of reasoning can be made explicit using belief evolution.

In order to give a precise treatment of belief change in a cryptographic protocol, one would need to consider multiple agents with nested beliefs. Such a detailed treatment is beyond the scope of this dissertation, but we can still illustrate the basic idea.

We identify sent messages with actions and we identify received messages with observations. From the perspective of a single agent, cryptographic protocols generally have the following form, where each A_i is an action and each α_i is an observation.

Generic Protocol

1. A_1
2. α_1
3. $\vdots$

2n-1.A_n

2n. α_n

In protocol verification, we typically assume that the principle agent has some initial belief state κ, and we are interested in proving that some property holds after every protocol run. If the desired property can be given by a set of states $PROP$, then protocol verification consists in answering the following question. If W is a world view containing the subsequence

$$\langle A_1, \alpha_1, \ldots, A_n, \alpha_n \rangle$$

does it always follow that the final belief state in $\kappa \circ W$ is a subset of $PROP$?

Note that this approach to protocol verification does not require any ad hoc rules describing belief change, instead we have framed the problem as a simple application of belief evolution. We illustrate how this procedure can be applied in the case of the Challenge-Response protocol.

Example Let **F** be the set containing the fluent symbols

$$HasKey(X), HasEncryptedMessage(X), HasDecryptedMessage(X)$$

where X ranges over the agent names P, Q, and I_Q. Let **A** contain the action symbols

$$SendEncryptedMessage(X), SendDecryptedMessage(X)$$

where X again ranges over P, Q, and I_Q. Let T be the non-deterministic transition system where $SendDecryptedMessage(X)$ makes $HasDecryptedMessage(Y)$ become true for some $Y \neq X$. Similarly, $SendEncryptedMessage(X)$ makes $HasEncryptedMessage(Y)$ true for some $Y \neq X$, and also makes $HasDecryptedMessage(Y)$ true if $HasKey(Y)$. It is straightforward to define a transition system satisfying these conditions, with every other fluent being inertial. We remark that we interpret $HasDecryptedMessage(X)$ to mean that the agent X has received the decrypted nonce during the protocol run.

We consider the belief change that occurs in a run of the Challenge-Response protocol. Let the initial belief state κ be the set of states s such that

1. $s \models HasKey(X)$ iff $X = P$ or $X = Q$
2. $s \models HasEncryptedMessage(X)$ iff $X = P$

3. $s \models HasDecryptedMessage(X)$ iff $X = P$.

The final message received by P is identified with the observation α defined as follows:

$$\alpha = |HasDecryptedMessage(Q) \vee HasDecryptedMessage(P)|.$$

So α consists of all states where either P or Q has received the first message. The goal of the protocol is to establish that Q received the first message, so we define the goal to be the set $|HasDecryptedMessage(Q)|$.

According to our approach, proving that the protocol is correct amounts to proving that, for all world views $\langle \bar{A}, \bar{\alpha} \rangle$ containing the subsequence

$$SendEncryptedMessage(P), |HasDecryptedMessage(P)|,$$

if $\kappa \circ \langle \bar{A}, \bar{\alpha} \rangle = \langle \kappa_0, \ldots, \kappa_n \rangle$ then $\kappa_n \subseteq |HasDecryptedMessage(Q)|$.

The attack on the Challenge-Response protocol is given by the world view $\langle \bar{A}, \bar{\alpha} \rangle$ where

$$\begin{aligned} A_1 &= SendEncryptedMessage(P) \\ \alpha_1 &= |HasEncryptedMessage(P)| \\ A_2 &= SendDecryptedMessage(P) \\ \alpha_2 &= |HasDecryptedMessage(P)|. \end{aligned}$$

Regardless of the underlying revision operator, the final belief state in $\kappa \circ \langle \bar{A}, \bar{\alpha} \rangle$ contains the state

$$\{HasEncryptedMessage(P), HasDecryptedMessage(P)\}.$$

Clearly this is not an element of $|HasDecryptedMessage(Q)|$, so the protocol fails to establish the goal.

We have framed protocol verification as a problem in belief evolution. One major advantage of this approach is that it allows us to model non-monotonic belief change. By contrast, existing protocol logics tend to model belief change through monotonic rules of inference. Using a general purpose belief change framework for reasoning about protocol goals is also useful because it allows us to easily apply the same methods to more complex protocol goals, such as non-repudiation and anonymity. However, using belief evolution operators as the underlying approach is somewhat limited in that agents can not explicitly reason about failed or exogenous actions. For example, we might be interested in agents

that are able to make inferences of the form: if α is believed at time 2, then it is believed that a message was intercepted at time 1. In the next chapter, we present a generalization of belief evolution that is suitable for reasoning about problems of this kind.

Chapter 5

Extending the Framework

In Chapter 1, we stated that our primary motivation was to formalize the belief change that occurs in problems of the following form:

$$(\mathit{InitialBeliefs}) \cdot (\mathit{Action}) \cdot (\mathit{Observation}) \cdots (\mathit{Action}) \cdot (\mathit{Observation}).$$

However, thus far we have only considered problems of this form in which an agent has perfect knowledge about the history of ontic actions. This is essentially the content of property P2, which states that an observation is discarded if it is not a possible consequence of the preceding action. However, this is not always a reasonable approach. For instance, in action domains where exogenous actions may occur, it may be more reasonable to explain an inconsistent observation by assuming that an exogenous action occurred. More generally, if an agent is uncertain about the history of ontic actions, then it may not be reasonable to simply discard observations that conflict with the perceived history of actions. In this chapter, we introduce a new kind of world view that is suitable for reasoning about iterated belief change in action domains where the history of ontic actions may be incorrect. Although the formal methods employed in this chapter differ superficially from the previous chapters, we will prove that the new class of world views is in fact a generalization of the original definition.

We give a schematic example illustrating the kind of problem with which we will be concerned. Suppose that an agent believes that the action trajectory $\bar{A}$ gives the sequence of actions that have been executed. Now suppose that the agent observes α, where α consists entirely of states that cannot possibly occur following the action sequence $\bar{A}$. The agent has two options.

1. Reject α.

2. Accept α, and modify $\bar{A}$ accordingly.

The first option is the case considered in Chapter 3, where the action trajectory $\bar{A}$ is deemed to represent more reliable information than the observation α. The second option corresponds to the situation where the agent believes that the observation α is more likely to be correct than the action trajectory $\bar{A}$. In order to resolve conflicts of this nature, an agent needs some mechanism for comparing the plausibility of an action occurrence with the plausibility of an observation.

We use Spohn-style ranking functions to reason about belief change following an alternating sequence of actions and observations. At each instant, an agent assigns a plausibility value to every action and every state; the most plausible world histories are obtained by combining these values through a suitable aggregate function. Since plausibility is given a quantitative rank, an agent is able to compare the plausibility of actions and observations. This allows action occurrences to be postulated or refuted in response to new observations. By allowing action histories and observation histories both to be incorrect, we are able to reason about iterated epistemic action effects in the context of fallible beliefs, erroneous perception, exogenous actions and failed actions.

In the remainder of this chapter, we proceed as follows. In §5.1, we introduce an illustrative running example. In §5.2, we define graded world views, which are sequences of Spohn-style ranking functions. We illustrate how graded world views can be used to represent a wide range of epistemic action domains in §5.3. In §5.4, we illustrate the generality of graded world views by proving subsumption results for AGM revision, belief evolution, and Spohn's conditionalization. We conclude in §5.5 by introducing constraints on graded world views, and using the notion of a constrained world view to prove a non-subsumption result for belief extrapolation.

A preliminary version of the work presented in this chapter previously appeared in [47].

5.1 Motivating Example

We introduce a common-sense example in which an agent needs to compare the plausibility of certain actions with the plausibility of observations. We will return to this example periodically as we introduce our formal machinery.

We consider a simple action domain involving four agents: Bob, Alice, Eve, and Trent. Bob places a chocolate chip cookie on his desk and then leaves the room; he believes that no one is likely to eat his cookie while he is gone. At time 1, Bob knows that Alice is at his desk. At time 2, Bob knows that Eve is at his desk. After Eve leaves his desk, Trent comes and tells Bob that a bite has been taken from the cookie on his desk.

Given the preceding information, Bob can draw three reasonable conclusions: Alice bit the cookie, Eve bit the cookie, or Trent gave him poor information. If Bob has no additional information about the world, then each conclusion is equally plausible. However, we suppose that Bob does have some additional information. In particular, suppose that Alice is a close friend of Bob and they have shared cookies in the past. Moreover, suppose that Bob believes that Trent is always honest. Bob's additional information about Alice and Trent provides a sufficient basis for determining which of the three possible conclusions is the most plausible.

Informally, prior to Trent's report, Bob believes that his cookie was unbitten at all earlier points in time. After Trent tells him the cookie is bitten, he must determine the most plausible world history consistent with this information. In this case, the most plausible solution is to conclude that Alice bit the cookie. Note that this conclusion requires Bob to alter his subjective view of the action history. There is a non-monotonic character to belief change in this context, because Bob may be forced to postulate and retract actions over time in response to new observations. The ramifications of changing the action history are determined by the underlying transition system.

We remark that this example does not admit a reasonable representation in terms of belief evolution because Bob does not have certain knowledge about which actions have occurred. In order to represent this kind of reasoning, we need to be able to compare the plausibility of action occurrences at different points in time.

5.2 Ranking Functions over Actions and States

5.2.1 Plausibility Functions

We are interested in action domains where action histories may be incorrect. In this context, the action that is executed at any given point in time can be represented by a pre-order over all possible actions. The minimal elements of such a pre-order represent the actions that were most likely executed, and moving higher in the ordering gives increasingly implausible possibilities. Representing actions in this manner allows an agent to determine plausible

alternative actions in the face of conflicting evidence. Similarly, an agent needs a mechanism for ordering states in order to represent fallible observations and fallible beliefs. Moreover, we would like to be able to compare orderings over actions with orderings over states. One natural way to create mutually comparable orderings is by assigning quantitative plausibility values to every action and state at every point in time. Towards this end, we define plausibility functions.

Definition 30 *Let X be a non-empty set. A plausibility function over X is a function $r : X \rightarrow \mathbf{N}$.*

If r is a plausibility function and $r(x) \leq r(y)$, then we say that x is at least as plausible as y. We will only be interested in plausibility functions over finite sets, where there is always a non-empty set of minimally ranked elements.

Plausibility functions are inspired by Spohn's ordinal conditional functions [88], but there are some important differences. First, we allow plausibility functions over an arbitrary set X, rather than restricting attention to propositional interpretations. This allows us to treat actions in the same manner that we treat observations. Another important difference is that ordinal conditional functions must always assign rank 0 to a non-empty subset of elements of the domain. Plausibility functions are not restricted in this manner; the minimal rank for a given plausibility function may be greater than 0. We have defined plausibility functions in this manner because we will be interested in taking sums over plausibility functions, and we need to ensure that such sums also define plausibility functions.

We remark that Darwiche and Pearl also consider ranking functions that do not necessarily assign rank 0 to any states [19]. However, Darwiche and Pearl define the belief state associated with r to be the set of states that are assigned rank 0. Under this convention, ranking functions that never assign rank 0 are associated with the empty belief state. By contrast, we associate a non-empty belief state with every plausibility function.

We introduce some useful terminology and notation. Let r be a plausibility function over X. The minimum and maximum values obtained by r are denoted by $\min_r$ and $\max_r$, respectively. We define $Bel(r)$ to be the set $\{w \mid r(w) = \min_r\}$. This notation is intended to suggest that $Bel(r)$ is the set of actions or states that are believed. For $\alpha \subseteq X$, we define $r(\alpha)$ to be the minimum value obtained by r for some $w \in \alpha$. The *degree of strength* of a plausibility function r is the least n such that $\min_r + n = r(v)$ for some $v \notin Bel(r)$. Hence, the degree of strength of r is the span between the plausibility of the minimally ranked

elements and the non-minimally ranked elements. The degree of strength in r indicates how much the plausibility rank increases if we choose some $v \notin Bel(r)$. There are two natural interpretations of the degree of strength of a plausibility function r over a set of states. If we think of r as an initial epistemic state, then the degree of strength is an indication of how strongly it is believed that the actual state is in $Bel(r)$. If we think of r as an observation, then the degree of strength is a measure of reliability. In the case where X is a set of states, we use the terms degree of strength and degree of belief interchangeably.

Note that Spohn defines the degree of strength of a subset of X, rather than the degree of strength of a ranking function. Our definition coincides with Spohn's definition if we identify the degree of strength of r with Spohn's degree of strength of the set $Bel(r)$. Hence, we use the same conception of degree of strength, but we are only interested in the strength of belief in the minimally ranked elements.

In order to illustrate the application of plausibility functions over different domains, we continue our simple example.

Example (cont'd) We describe how the cookie problem can be represented with plausibility functions.

Let $\mathbf{F} = \{BiteTaken\}$ and let $\mathbf{A} = \{BiteAlice, BiteEve\}$. Both actions have the same effect, namely they both make the fluent $BiteTaken$ become true. We represent the problem with 3 plausibility functions: a_1, a_2, and o_2.

1. a_1 is a plausibility function over actions at Time 1
2. a_2 is a plausibility function over actions at Time 2
3. o_2 is a plausibility function over states at Time 2

Informally, each function should obtain a minimum value at the event that Bob considers the most plausible at the given point in time. Since Bob initially believes that no one will eat his cookie, both a_1 and a_2 should obtain a minimum value at the null action λ. Trent's report that the cookie has been bitten at Time 2 is represented as a plausibility function over states, by defining o_2 with a minimum at the set of worlds where the cookie has a bite out of it. Note that we will generally treat reported information in this manner; the degree of strength of a report is an indication of trust in the agent providing the report. The additional soft constraints about Bob's relationships are used to determine the magnitude of the values for each event. Define a_1 and a_2 by the values in the following table.

	λ	*BiteAlice*	*BiteEve*
a_1	0	1	10
a_2	0	10	3

The fact that Alice is more likely to bite the cookie is represented by assigning a low plausibility value to *BiteAlice* at time 1. Define o_2 as follows.

	$\emptyset$	$\{BiteTaken\}$
o_2	9	0

Hence, the observation $\{BiteTaken\}$ is assigned the minimum plausibility value, and the only alternative observation is assigned a very high plausibility value. This reflects the fact that Trent's report is understood to supersede the assumption that Alice and Eve do not bite the cookie.

Note that the degree of strength of a_1 is less than the degree of strength of a_2 and o_2. This gives an indication that Bob has comparatively less confidence in his beliefs about the action at Time 1.

5.2.2 Graded World Views

In Chapter 3, we essentially defined a world view to be an alternating sequence of actions and observations. We have now suggested that, in the context of imperfect action histories, actions and observations can both be represented by plausibility functions. This leads to a natural extension of the definition of a world view. In particular, we can define a *graded world view* to be an alternating sequence of plausibility functions over $\mathbf{A}$ and plausibility functions over $2^{\mathbf{F}}$. We have the following formal definition.

Definition 31 *A graded world view of length n is a $(2n+1)$-tuple*

$$\langle OBS_0, ACT_1, OBS_1, \ldots, ACT_n, OBS_n \rangle$$

where each OBS_i is a plausibility function over $2^{\mathbf{F}}$ and each ACT_i is a plausibility function over $\mathbf{A}$.

At time i, the most plausible actions are the minimally ranked actions of ACT_i and the most plausible states are the minimally ranked states of OBS_i. We take OBS_0 to represent the initial belief state, and each subsequent OBS_i to represent a new observation. If

$ACT = \langle ACT_1, \ldots, ACT_n \rangle$ and $OBS = \langle OBS_0, \ldots, OBS_n \rangle$, then we write $\langle ACT, OBS \rangle$ as a shorthand for the graded world view $\langle OBS_0, ACT_1, OBS_1, \ldots, ACT_n, OBS_n \rangle$. We use the notation $\langle ACT, OBS \rangle$ to emphasize the similarity with a world view $\langle \bar{A}, \bar{\alpha} \rangle$. In both cases, we are representing an agent's view of the history of actions and observations.

We remark briefly on the intuition behind graded world views. We are interested in action domains involving actions that are both partially observable and fallible. However, for the moment we do not consider failed actions. The plausibility of an action A represents the likelihood that A was successfully executed at a given instant. Hence, the lowest plausibility values will be assigned to actions that an agent has executed, or actions that an agent has observed directly. Higher plausibility values will be assigned to exogenous actions that are assumed to be unlikely, or action occurrences that are only believed based on external reports. In §5.2.4, we provide some additional motivation for plausibility values by illustrating a correspondence with subjective probability functions.

Note that graded world views differ from the world views of Chapter 3 in that a graded world view includes the initial belief state. Our original notion of a world view was framed in the context of AGM revision functions, where initial beliefs are always abandoned in favour of a new observation. By contrast, the underlying assumption in a graded world view is that the initial belief state is no different than any subsequent observation; there is no reason to automatically prefer the initial beliefs over new information, nor is there any reason to automatically disregard the initial beliefs given new information. In order to make this assumption salient, we represent the initial belief state as an observation at time 0.

5.2.3 Aggregate Plausibility Functions

Given a graded world view $\langle ACT, OBS \rangle$, we would like to be able to determine the most plausible history of the world. We formally define the notion of a *history* over a transition system.

Definition 32 *Let $T = \langle S, R \rangle$ be a transition system. A history of length n is a tuple $\langle w_0, A_1, \ldots, A_n, w_n \rangle$ where for each i:*

1. $w_i \in S$,
2. $A_i \in \mathbf{A}$, *and*
3. $\langle w_i, A_{i+1}, w_{i+1} \rangle \in R$.

Let $HIST_n$ denote the set of histories of length n.

Ideally, we would like to use graded world views to assign plausibility values to histories. However, a graded world view does not provide sufficient information to define a unique plausibility function over histories. For example, a graded world view does not indicate the relative weight of recent information versus initial information. In order to determine the most plausible history, we need some mechanism for combining a sequence of plausibility functions.

Although a graded world view does not define a unique plausibility function over histories, we can define a general notion of consistency between graded world views and plausibility functions on histories. Let $r_0, \ldots, r_n$ be plausibility functions over $X_0, \ldots, X_n$, respectively. Let r be a plausibility function over $X_0 \times \cdots \times X_n$. We say that r is *consistent with* $\langle r_0, \ldots, r_n \rangle$ if, for every i and every $x_i, x_i' \in X_i$

$$r_i(x_i) < r_i(x_i')$$
$$\Longleftrightarrow$$
$$r(\langle x_0, \ldots, x_i, \ldots, x_n \rangle) < r(\langle x_0, \ldots, x_i', \ldots, x_n \rangle)$$

So r is consistent with $\langle ACT, OBS \rangle$ just in case r increases monotonically with respect to each component of $\langle ACT, OBS \rangle$. Any plausibility function r that is consistent with $\langle ACT, OBS \rangle$ provides a potential candidate ranking over histories.

Define an *aggregate plausibility function* to be a function that maps every graded world view of length n to a plausibility function over $HIST_n$. We are interested in aggregate plausibility functions in which the output is always consistent with the input. Hence, we say that an aggregate plausibility function *agg* is *admissible* if, for every $\langle ACT, OBS \rangle$, the function $agg(\langle ACT, OBS \rangle)$ is consistent $\langle ACT, OBS \rangle$.

We provide some examples. Note that aggregate plausibility functions return a function as a value; we can specify the behaviour of an aggregate by specifying a plausibility value for each pair consisting of a graded world view and a history. Let $h = \langle w_0, A_1, \ldots, A_n, w_n \rangle$. One admissible aggregate is obtained by taking the sum of plausibility values.

$$sum(\langle ACT, OBS \rangle)(h) = \sum_{i=1}^{n} ACT_i(A_i) + \sum_{i=0}^{n} OBS_i(w_i)$$

A weighted sum can be used to reflect the relative importance of different time points. For each i, let b_i be a positive integer.

$$sum_w(\langle ACT, OBS\rangle)(h) = \sum_{i=1}^{n} ACT_i(A_i) + \sum_{i=0}^{n} b_i \cdot OBS_i(w_i).$$

By setting $b_i = 2^i$, the aggregate function sum_w can be used to represent a strict preference for recent information. The functions sum and sum_w are just two simple examples; many more examples can be defined by specifying aggregate functions that increase monotonically with each component.

We return to the cookie example to illustrate how the reasoning involved can be captured with graded world views and aggregate plausibility functions.

Example (cont'd) We have already defined plausibility functions a_1, a_2 and o_2. In order to give a complete graded world view, we need to define two more plausibility functions over states. In particular, we need to give a plausibility function o_0 representing Bob's initial beliefs and we need to give a plausibility function o_1 representing the null observation that Bob makes at Time 1.

First, we reiterate the description of a_1 and a_2 in the following table.

	λ	*BiteAlice*	*BiteEve*
a_1	0	1	10
a_2	0	10	3

The fact that Alice is more likely to bite the cookie is represented by assigning a lower plausibility value to *BiteAlice* at Time 1.

The plausibility function o_0 should assign a minimum value to the state where the cookie is unbitten. The plausibility function o_1 should assign the same value to every state. The plausibility function o_2 (given previously) represents Trent's report that the cookie has been bitten. As noted previously, we treat reported information as an observation, and we use the degree of strength of the reported information as an indication of the reliability of the source. In this case, the degree of strength of o_2 is an indication of trust in Trent. We define o_0, o_1, o_2 in the next table.

Note that the degree of strength of o_2 is higher than the degree of strength of a_1 or a_2. This reflects the fact that Trent's report is understood to supersede the assumption that Alice and Eve do not bite the cookie. Graded world views have been defined precisely for

	$\emptyset$	$\{BiteTaken\}$
o_0	0	9
o_1	0	0
o_2	9	0

this kind of comparison between action plausibilities and state plausibilities.

If we use the aggregate function *sum*, then we are interested in finding the minimal sum of plausibilities over $\langle o_0, a_1, o_1, a_2, o_2 \rangle$. By inspection, we find that the minimum plausibility is obtained by the following history:

$$h = \langle \emptyset, BiteAlice, BiteTaken, \lambda, BiteTaken \rangle.$$

This history represents the sequence of events in which Alice bites the cookie at time 1. Intuitively, this is the correct solution: given the choice between Alice and Eve, Bob believes that Alice is the more plausible culprit.

We remark that graded world views bear a resemblance to the generalized belief change framework proposed by Liberatore and Schaerf [62]. However, the Liberatore-Schaerf approach associates a "penalty" with state change, which is minimized when determining plausible models. As such, it is difficult to represent problems where non-null actions are strictly more plausible than null actions. By contrast, graded world views have no implicit preference for null actions. Moreover, our approach differs in that we allow actions with conditional effects given by a transition system.

5.2.4 Subjective Probabilities

One issue that arises from our definition of a graded world view is the fact that it is not clear how plausibility values should be assigned in practical problems. We address this problem by illustrating a correspondence between plausibility functions and *probability functions*. We simplify the discussion by restricting attention to rational-valued probability functions as follows.

Definition 33 *Let X be a non-empty set. A probability function over X is a function $Pr : X \to \mathbf{Q}$ such that*

- *for all $x \in X$, $0 \leq Pr(x) \leq 1$*

- $\sum_{x \in X} Pr(x) = 1$.

We do not need any other axioms of probability theory for our present purposes. At a common-sense level, it is clear what it means to say that "action A occurred at time t with probability p." By contrast, the problem with plausibility values is that there is no obvious sense of scale; it is difficult to assign numerical plausibility values, because the numbers have no clear meaning. We illustrate how probability functions can be translated uniformly into plausibility functions, thereby giving a sense of scale and meaning to plausibility values.

Let Pr be a probability function over a finite set X. Let Q denote the least common denominator of all rational numbers $\frac{p}{q}$ such that $Pr(x) = \frac{p}{q}$ for some $x \in X$. Define the plausibility function r as follows.

1. If $Pr(x)$ is minimal, set $r(x) = Q$.
2. Otherwise, if $Pr(x) = \frac{p}{Q}$, then set $r(x) = Q - p$.

Hence, every probability function can be translated into a plausibility function.

Example (cont'd) Consider the following probability functions for the cookie example.

	λ	*BiteAlice*	*BiteEve*
Pr_{a_1}	.5	.45	.05
Pr_{a_2}	.5	.15	.35

	$\emptyset$	$\{BiteTaken\}$
Pr_{o_0}	.9	.1
Pr_{o_1}	.5	.5
Pr_{o_2}	.1	.9

The corresponding plausibility functions are given in the following tables.

	λ	*BiteAlice*	*BiteEve*
a'_1	10	11	20
a'_2	10	20	13

	$\emptyset$	$\{BiteTaken\}$
o'_0	1	10
o'_1	2	2
o'_2	10	1

It is easy to see that these plausibility functions are obtained from the plausibility functions given earlier by adding a constant to each value. In §5.3.2, we illustrate that adding a constant in this manner does not affect the class of minimally ranked histories.

Connecting plausibility functions with subjective probabilities provides some justification for the use of the aggregate function *sum*. In particular, if we assume that the subjective probability functions are *independent*, then the probability of a given sequence of events is

determined by taking a product. In the cookie example, we can compare the probability of Alice biting the cookie versus Eve biting the cookie:

1. $Pr(\langle \emptyset, BiteAlice, BiteTaken, \lambda, BiteTaken \rangle)$
   $= .9 \times .45 \times .5 \times .5 \times .9 = .091125$

2. $Pr(\langle \emptyset, \lambda, \emptyset, BiteEve, BiteTaken \rangle)$
   $= .9 \times .5 \times .5 \times .35 \times .9 = .070875$

It is easy to check that the history where Alice bites the cookie is actually the most probable history. So, in this example, the minimally ranked history according to the aggregate function *sum* is also the most probable history according to the sequence of probability functions. This is a general property of our translation: maximizing probability over independent probability functions corresponds to minimizing the sum over plausibility values.

5.2.5 The Summation Convention

In order to ground the discussion, it is useful to choose a fixed aggregate function for assigning plausibility values to histories. As such, unless otherwise indicated, we will assume that plausibility values are assigned to histories by the aggregate function *sum*. Although this is not the only approach to combining plausibility functions, it provides a simple admissible aggregate function that is appropriate in many cases. In particular, we saw in the previous section that *sum* is appropriate for domains where the plausibility functions have been obtained from subjective probabilities.

We introduce some notation that will simplify the results in the next few sections. Recall that $sum(\langle ACT, OBS \rangle)$ is a plausibility function on histories. When the underlying graded world view is clear from the context, we will write $plaus(h)$ as a shorthand for $sum(\langle ACT, OBS \rangle)(h)$.

It is useful to introduce an operator that maps a graded world view to the most plausible histories.

Definition 34 *Let WV denote the set of graded world views of length n for a fixed action signature. Define $\Phi : WV \rightarrow 2^{HIST_n}$ as follows:*

$$\Phi(\langle ACT, OBS \rangle) = \{h \mid plaus(h) \leq plaus(g) \text{ for all } g \in HIST_n\}.$$

We have the following obvious equivalence

$$\Phi(\langle ACT, OBS\rangle) = Bel(sum(\langle ACT, OBS\rangle)).$$

It is also useful to use *plaus* to define a plausibility function over states.

Definition 35 *Let $\langle ACT, OBS\rangle$ be a graded world view. For any state w, define*

$$plaus\_state(\langle ACT, OBS\rangle)(w)$$

to be the least n such that $plaus(\langle ACT, OBS\rangle)(h) = n$ for some history h with final state w.

So the plausibility of the state w is the rank of the most plausible history ending with w. When the underlying graded world view is clear from context, we simply write $plaus\_state(w)$ for the plausibility of the state w. We extend the operator $Bel(\cdot)$ to graded world views by defining $Bel(\langle ACT, OBS\rangle)$ to be $Bel(plaus\_state(\langle ACT, OBS\rangle))$. Hence, *Bel* takes a graded world view as an argument and returns the most plausible set of terminal states.

5.3 Using Graded World Views

5.3.1 Pointwise Minima

Suppose that the underlying set $\mathbf{F}$ of fluent symbols and the underlying set $\mathbf{A}$ of action symbols are both finite. Let $W = \langle ACT, OBS\rangle$ be a graded world view where

$$ACT = \langle ACT_1, \ldots, ACT_n\rangle$$

and

$$OBS = \langle OBS_0, \ldots, OBS_n\rangle.$$

The easiest way to determine a minimally ranked history is to simply take the most plausible actions and the most plausible worlds at each point in time. The following definition makes this notion more precise.

Definition 36 *Given a history $h = \langle w_0, A_1, \ldots, A_n, w_n\rangle$, we say h is a pointwise minimum for $\langle ACT, OBS\rangle$ if, for all i,*

1. *for all $A \in \mathbf{A}$, $ACT_i(A_i) \leq ACT_i(A)$, and*

2. *for all* $w \in 2^{\mathbf{F}}$, $OBS_i(w_i) \leq OBS_i(w)$.

The following proposition states that, if a graded world view has any pointwise minima, then those will be the most plausible histories.

Proposition 20 *Let* $W = \langle ACT, OBS \rangle$ *be a graded world view and let* M *be the set of pointwise minima for* W. *If* $M \neq \emptyset$, *then* $\Phi(W) = M$.

Proof It is sufficient to note that, for $h \in M$, $plaus(h) \leq plaus(g)$ for all histories g. □
Note, however, that histories are restricted in that each world must be the outcome of the preceding action. As such, it is possible that a graded world view will have no pointwise minimum.

5.3.2 Equivalence

Clearly it is possible for two distinct graded world views to have the same set of minimally ranked world histories. In fact, it is possible for two distinct graded world views to induce the same preference ordering over histories. In this section, we define a natural equivalence relation over graded world views with an eye towards categorical representations. We start by defining a relation on plausibility functions.

Definition 37 *Let* r_1 *and* r_2 *be plausibility functions over a set* X. *We say that* $r_1 \cong r_2$ *if, for every* $x, y \in X$,

$$r_1(x) - r_1(y) = r_2(x) - r_2(y).$$

It is clear that $\cong$ is an equivalence relation.

Let r be a plausibility function. For any integer z, the *translation* of r by z is the plausibility function $x \mapsto r(x) + z$. It is easy to prove that $r \cong r'$ if and only if r' is a translation of r. We define the *normalization* of r to be the translation by $-\min_r$. The normalization of r is the unique plausibility function equivalent to r that obtains a minimum of 0.

We can extend the notion of equivalence to graded world views.

Definition 38 *Let* WV_1 *and* WV_2 *be graded world views over histories for a fixed action signature. We say that* $WV_1 \cong WV_2$ *if, for every pair of histories* g *and* h,

$$sum(WV_1)(g) - sum(WV_1)(h) = sum(WV_2)(g) - sum(WV_2)(h).$$

Unlike plausibility functions, it is possible to construct equivalent pairs of graded world views that are not obtained by translations.

The following proposition illustrates that every graded world view is equivalent to a graded world view consisting of normalized plausibility functions.

Proposition 21 *Let $\langle ACT, OBS\rangle$ be a graded world view. If $\langle ACT', OBS'\rangle$ is obtained by normalizing each component of ACT and OBS, then*

$$\langle ACT, OBS\rangle \cong \langle ACT', OBS'\rangle.$$

Proof Let g, h be histories. For ease of readability, let $plaus_1$ and $plaus_2$ denote $sum(\langle ACT, OBS\rangle)$ and $sum(\langle ACT', OBS'\rangle)$, respectively. Then the following equalities are immediate:

$$\begin{aligned} plaus_2(g) - plaus_2(h) &= plaus_1(g) - \min_{plaus_1} -plaus_1(h) + \min_{plaus_1} \\ &= plaus_1(g) - plaus_1(h). \end{aligned}$$

□

Hence, although we allow plausibility functions with minimum values larger than 0 in a graded world view, we can always pass to an equivalent graded world view consisting of normalized plausibility functions. We remark, however, that a graded world view defined by a sequence of normalized plausibility functions need not obtain a minimum of 0. In this case, the minimum will be 0 if and only if the graded world view has a pointwise minimum. It is also important to note that Proposition 21 only holds under the aggregate plausibility function *sum*.

5.3.3 Representing Belief States

Graded world views can be defined that simply pick out a distinguished set of elements of the domain. If $\alpha \subseteq X$ and c is an integer, let $\alpha\!\uparrow\! c$ denote the function defined as follows:

$$\alpha\!\uparrow\! c\ (w) = \begin{cases} 0 & \text{if } w \in \alpha \\ c & \text{otherwise} \end{cases}$$

If c is a positive integer, then $\alpha\!\uparrow\! c$ denotes a plausibility function in which the elements of α are the most plausible, and everything else is equally implausible. Plausibility functions of the form $\alpha\!\uparrow\! c$ will be called *simple*. If X is a set of states, then simple plausibility functions

correspond to belief states; if X is a set of actions, then simple plausibility functions pick out the actions that are believed to have occurred. Using the terminology introduced earlier, we say that α is held with degree of belief c.

If $c > 0$, then $\alpha\uparrow -c$ does not actually define a plausibility function. However, allowing negative values leads to a simple symmetry in our notation. In the following proposition, $\tilde{\alpha}$ denotes the complement of α.

Proposition 22 *For any set α and positive integer c*

$$\alpha\uparrow c \cong \tilde{\alpha}\uparrow -c.$$

Proof Let w, v be states. By definition, we have

$$\alpha\uparrow c\ (w) - \alpha\uparrow c\ (v) = \begin{cases} c & \text{if } w \in \alpha, v \notin \alpha \\ -c & \text{if } w \notin \alpha, v \in \alpha \\ 0 & \text{otherwise.} \end{cases}$$

and

$$\tilde{\alpha}\uparrow -c\ (w) - \tilde{\alpha}\uparrow -c\ (v) = \begin{cases} -(-c) & \text{if } w \notin \tilde{\alpha}, v \in \tilde{\alpha} \\ -c & \text{if } w \in \tilde{\alpha}, v \notin \tilde{\alpha} \\ 0 & \text{otherwise.} \end{cases}$$

Clearly, the right hand sides of each equality are the same. □

Suppose that

$$ACT = \langle ACT_1, \ldots, ACT_n \rangle$$

and

$$OBS = \langle OBS_0, \ldots, OBS_n \rangle$$

where each ACT_i and OBS_i is simple, with maximum plausibility c. Hence, we essentially have belief states with no plausibility ordering. In this case, it is easy to show that

$$\langle w_0, A_1, \ldots, A_n, w_n \rangle \in \Phi(\langle ACT, OBS \rangle)$$

if and only if the cardinality of

$$\{A_i \mid A_i \in ACT_i\} \cup \{w_i \mid w_i \in OBS_i\}$$

is maximal among all histories. In other words, the most plausible histories are those that agree with $\langle ACT, OBS \rangle$ at the highest number of components. This is a reasonable approach to take in the trivial case where we have no prior ranking over states or actions.

5.3.4 Graded World Views as Epistemic States

Recall that an *epistemic state* is a representation of an agent's beliefs that defines a total pre-order $\preceq$ over all states [19]. If $w \preceq v$, then the underlying agent believes that it is more likely that the actual state of the world is w than v. The current belief state is given by the set of $\preceq$-minimal states. Recall also that a graded world view defines a plausibility function *plaus_state* over states. So a graded world view clearly defines an ordering over states, and we can think of a graded world view as defining an epistemic state. The worlds that receive minimal rank in a graded world view are the worlds that are supported by the most reliable observations and actions. Using this ranking to define a plausibility ordering is tantamount to assuming that the plausibility of w is completely determined by the reliability of the source reporting that w occurs.

By viewing graded world views as epistemic states, we can define belief change operations in a more familiar manner. In particular, we can define belief change through a simple concatenation operator $\bullet$ on graded world views. Given a sequence of plausibility functions $\bar{r} = \langle r_1, \ldots, r_n \rangle$ and a plausibility function r, we let $\bar{r} \cdot r$ denote the sequence $\langle r_1, \ldots, r_n, r \rangle$. Let $\langle ACT, OBS \rangle$ be a graded world view, let r_A be a plausibility function over actions and let r_S be a plausibility function over states. Define $\bullet$ as follows:

$$\langle ACT, OBS \rangle \bullet \langle r_A, r_S \rangle = \langle ACT \cdot r_A, OBS \cdot r_S \rangle.$$

In this context the initial epistemic state is given by $\langle ACT, OBS \rangle$, which represents an agent's a priori beliefs about the history of observed actions and states. New actions and observations are incorporated by simply concatenating the new plausibility functions on to the initial graded world view. The new graded world view defines a new ordering over histories, but it also includes all historical information required for future belief change. As a special case of this simple concatenation operation, we get a new approach to update. For any set X, let 0 denote the plausibility function that uniformly assigns 0 to every element of X. We can identify the update $\langle ACT, OBS \rangle \diamond r_A$ with the following operation:

$$\langle ACT, OBS \rangle \bullet \langle r_A, 0 \rangle.$$

We can also define a natural approach to revision in this manner. Let *null* denote a plausibility function that assigns plausibility 0 to the null action λ, and assigns everything else a plausibility larger than the maximum value obtained by $sum(\langle ACT, OBS \rangle)$. We identify

the revision $\langle ACT, OBS\rangle * r_S$ with the following operation:

$$\langle ACT, OBS\rangle \bullet \langle null, r_S\rangle.$$

Using plausibility functions to represent observations allows us to represent some natural problem domains that can not be easily represented if we restrict observations to sets of possible worlds. In particular, consider an action domain in which observations have varying degrees of reliability. In such domains, when an agent makes an observation that is inconsistent with the current belief state, there are two factors that should be considered: the strength of belief in the current belief state and the reliability of the observation. There is an obvious conflict that arises if we attempt to address both factors simultaneously. For example, suppose that the underlying agent strongly believes that w is a possible state of the world. Now suppose that the agent makes two observations.

1. One observation suggests that w is possible, but comes from an unreliable source.
2. Another observation suggests that w is not possible, and it comes from a very reliable source.

It can be difficult to determine the appropriate belief change in this scenario, particularly if strength of belief and observational reliability are treated independently. By quantifying the reliability of every observation, graded world views make it easy to resolve this kind of issue. We remark that problems of this form have also been addressed recently through the use of prioritized merging operators [21].

There is an interesting asymmetry in the definition of revision and update through the $\bullet$ operator. In the case of update, we assume that the final observation assigns the same plausibility to every state. The symmetric definition for a single observation would be defined as follows:

$$\langle ACT, OBS\rangle \bullet \langle 0, r_S\rangle.$$

However, this definition allows an arbitrary action to occur immediately before the observation. If we want to assume that the graded world view $\langle ACT, OBS\rangle$ gives a complete picture of the world at the time of the observation, then we need to assume that any intermediary action is null. Hence, the asymmetry is not due to any significant difference between actions and observations; the asymmetry is simply due to the fact that graded world views involve alternating sequences of actions and observations, with actions occurring first by default.

In this section we have illustrated that a graded world view defines an epistemic state. If we take an epistemic state to be a pre-order on states, then the converse is clearly false: an ordering on states does not provide enough information to define numerical ranking functions over states. The move from epistemic states to graded world views is motivated by the same kind of concern that motivates the move from belief states to epistemic states. In particular, belief states in AGM revision can be understood to represent the minimal elements in some ordering of states. Hence, a belief state can provide a partial description of an ordering, and an ordering can in turn provide a partial description of a graded world view. A belief state is sufficient for single-shot revision, provided that an ordering is implicit in the revision operator. However, a belief state is not sufficient if we need to explicitly reason about the way plausibility orderings are modified. Similarly, orderings on states are sufficient for reasoning about preferences over states, but they are not sufficient if we need to explicitly reason about action histories.

5.3.5 Representing Natural Action Domains

In this section, we illustrate how some interesting phenomena can be represented by graded world views. The simplest examples involve graded world views of length 1. In particular, we initially focus on graded world views of the form

$$\langle INIT \rangle \bullet \langle r_A, r_S \rangle.$$

In this context, $INIT$ represents the initial beliefs of an agent, r_A represents an agent's beliefs about the action that has been executed, and r_S represents the observed state of the world. To be clear, $INIT$, r_A, and r_S are all *plausibility functions.* As such, we can define the degree of strength of each. To facilitate the exposition, we denote the degrees of strength by $deg(INIT)$, $deg(r_A)$, and $deg(r_S)$ respectively. Varying the magnitudes of these values allows us to capture several different underlying assumptions.

1. Fallible initial beliefs: $deg(INIT) < deg(r_A)$ and $deg(INIT) < deg(r_S)$.
2. Erroneous perception: $deg(r_S) < deg(INIT)$ and $deg(r_S) < deg(r_A)$.
3. Fallible action history: $deg(r_A) < deg(INIT)$ and $deg(r_A) < deg(r_S)$.

As a simple example, suppose that an agent believes a certain lamp is initially on, then the power switch is toggled, and then the agent observes that the lamp is actually still

on. Clearly this sequence of events can not consistently be believed by a rational agent. Manipulating the degrees of strength of $INIT$, r_A and r_O gives an agent some mechanism for resolving such conflicts. In case (1), the agent is not completely certain that the lamp was initially on. As such, the easiest way to incorporate the new information is to change the initial belief state. By contrast, in case (2), the agent is not completely certain that the lamp is still on after toggling the switch. In this case, since the agent is confident the lamp was initially on and the switch was toggled, it is natural to reject the observation and believe that the lamp is now off. The distinction between these two cases cannot be captured without some notion of reliability.

The special case in which the degree of strength is 0 also captures some important phenomena. Note that a plausibility function r has degree of strength 0 just in case there is some constant c such that $r(x) = c$ for all x. As such, a degree of 0 indicates that every element of the domain receives minimal rank. We consider the informal interpretation of a degree 0 for each plausibility function in our schematic example.

1. If $deg(INIT) = 0$, then every initial state is equally plausible. The agent has no a priori beliefs about the state of the world.

2. If $deg(r_O) = 0$, then r_O represents a null observation. The observation OBS does not provide evidence for any particular state.

3. If $deg(r_A) = 0$, then every action is equally likely. So the agent is completely ignorant about the action that has occurred, and we can think of r_A as an exogenous action beyond the agent's control.

These are relatively crude distinctions, but they still capture important classes of problems. Roughly speaking, the problems that we have addressed thus far can be captured by a plausibility ordering over sequences of the form

$$\kappa \diamond A_1 * \alpha_1 \diamond \cdots \diamond A_n * \alpha_n$$

where κ is a belief state, each A_i is an action symbol, and each α_i is an observation. Recall from Chapter 3 that belief evolution operators are only useful for problems in which the underlying plausibility ordering is given as follows, for some permutation $p_1, \ldots, p_n$ of

$1, \ldots, n$.

$$\left.\begin{array}{c} A_1 \\ \vdots \\ A_n \end{array}\right\} \prec \alpha_{p_1} \prec \alpha_{p_2} \prec \cdots \prec \alpha_{p_n}$$

By contrast, graded world views are suitable for any total pre-order over $A_1, \alpha_1, \ldots, A_n, \alpha_n$. But this is not the entire class of problems representable by graded world views. By using a ranking function for each event, we are able to draw two additional distinctions that can not be represented by a simple ordering. First, we are able to represent changes in plausibility that do not affect the ordering of states. This is useful for representing action domains where an agent must observe a single piece of evidence multiple times before believing it is correct. Second, we are able to represent graded evidence that supports several conclusions with different degrees of confidence. We conclude this section with two examples illustrating action domains that are hard to represent if we only have an ordering over the plausibility of events.

Example (Additive Evidence) Bob believes that he turned the lamp off in his office, but he is not completely certain. As he is leaving the building, he talks first to Alice and then to Eve. If only Alice tells him his lamp is still on, then he will believe that she is mistaken. Similarly, if only Eve tells him his lamp is still on, then he will believe that she is mistaken. However, if both Alice and Eve tell Bob that his lamp is still on, then he will believe that it is in fact still on.

This example can easily be represented by a graded world view as follows. We assume that the underlying action signature contains, among others, a fluent symbol *LampOn* and an action symbol $TurnLampOff$. The underlying transition system defines the effects of turning the lamp off in the obvious manner. Let ON denote the set of states in which *LampOn* is true. The following plausibility functions define a graded world view that represents this action domain.

1. $OBS_0 = ON \uparrow 10$
2. $ACT_1 = \{TurnLampOff\} \uparrow 3$
3. $OBS_1 = ON \uparrow 2$
4. $ACT_2 = \lambda \uparrow 10$

5. $OBS_2 = ON \uparrow 2$

Note that $\Phi(\langle OBS_0, ACT_1, OBS_1 \rangle)$ consists of all histories where the lamp is turned off at time 1. However, $\Phi(\langle OBS_0, ACT_1, OBS_1, ACT_2, OBS_2 \rangle)$ consists of all histories where the lamp is not turned off at time 1. Two observations of ON are required to make Bob believe that he did not turn the lamp off.

Example (Graded Evidence) Bob receives a gift that he estimates to be worth approximately \$7. He is curious about the price, so he tries to glance quickly at the receipt without anyone noticing. He believes that the receipt says the price is \$3. This is far too low to be believable, so Bob concludes that he must have mis-read the receipt. Since a "3" looks very similar to an "8", he concludes that the price on the receipt must actually have been \$8.

To represent this example, we first define $ACT_1 = \lambda \uparrow 10$ because Bob believes that no ontic actions have occurred. We assume that there are fluent symbols $Cost1, Cost2, \ldots, Cost9$ interpreted to represent the cost of the gift. We define a plausibility function OBS_0 representing Bob's initial beliefs.

$$OBS_0(w) = \begin{cases} 0 \text{ if } w = \{Cost7\} \\ 1 \text{ if } w = \{Cost6\} \text{ or } w = \{Cost8\} \\ 3 \text{ otherwise} \end{cases}$$

Note that Bob initially believes that the cost is \$7, but it is comparatively plausible that this cost is one dollar more or less. Finally, we define a plausibility function OBS_1 representing the observation of the receipt.

$$OBS_1(w) = \begin{cases} 0 \text{ if } w = \{Cost3\} \\ 1 \text{ if } w = \{Cost8\} \\ 3 \text{ otherwise} \end{cases}$$

Bob believes that the observed digit was most likely a "3", with the most plausible alternative being the visually similar digit "8".

Given these plausibility functions, the most plausible state of the world is the state in which the price is \$8. In order to draw this conclusion, Bob needs observations that provide graded evidence about states of the world and he needs to be able to weight this information

against his initial beliefs.

The preceding examples illustrate that there are natural common-sense reasoning problems in which an agent needs to consider aggregate plausibilities over a sequence of actions and observations. Graded world views are well-suited for reasoning about such problems.

5.3.6 Non-Deterministic and Failed Actions

In this section, we consider actions with non-deterministic effects. Note that actions that may fail can be represented as actions with non-deterministic effects, so we address fallible actions in this section as well. Our basic approach is the following. We introduce some new machinery for the representation of non-deterministic actions, and then we demonstrate that the new machinery is unnecessary when we use summation to determine the plausibility of histories. As such, we can reasonably restrict attention to deterministic actions when proving formal expressibility results for graded world views.

Given a non-deterministic transition system $T = \langle S, V, R \rangle$ and a graded world view W, it is not clear how we should choose the effects of each action in the most plausible world histories. This problem can be solved by following [11], and attaching a plausibility value to the possible effects of each action. For each action A and state s, let $EFF(A, s)$ denote the set of states s' such that $(s, A, s') \in R$. Hence $EFF(A, s)$ is the set of states that may result, given that action A is executed in state s.

Definition 39 *An effect ranking function is a function δ that maps every action-state pair (A, s) to a plausibility function over $EFF(A, s)$.*

Informally, an effect ranking function gives the likelihood of each possible effect for each action.

A *non-deterministic graded world view* is a pair $\langle W, \delta \rangle$ where W is a graded world view and δ is an effect ranking function. We illustrate with an example.

Example Consider an action domain involving a single fluent symbol *LampOn* indicating whether or not a certain lamp is turned on. There are two action symbols *Press* and *ThrowPaper* respectively representing the acts of pressing on the light switch, or throwing a ball of paper at the light switch. Informally, throwing a ball of paper at the light switch is not likely to turn on the lamp. But suppose that an agent has reason to believe that a piece

of paper was thrown at the lamp and, moreover, the lamp has been turned on. We illustrate how non-deterministic graded world views can provide a representation of this problem.

Both actions have non-deterministic effects in that both may cause *LampOn* to become true, but both may also fail to do so. We define a graded world view $\langle ACT, OBS \rangle$ of length 1. First, we define ACT so that $ThrowPaper$ is the most likely action at time 1.

	λ	$Press$	$ThrowPaper$
ACT_1	10	2	1

Next we define OBS so that initially the light is off, and then the light is on.

	$\emptyset$	$LightOn$
OBS_0	10	0
OBS_1	0	10

Finally, we define an effect ranking function δ that represents the fact that pressing is more likely to turn the light on.

	$\emptyset$	$\{LightOn\}$
$\delta(Press, LightOn)$	0	10
$\delta(Press, \emptyset)$	1	2
$\delta(ThrowPaper, LightOn)$	0	10
$\delta(ThrowPaper, \emptyset)$	2	1

In the preceding example, there are two possible solutions: either a plausible event occurs with an unlikely outcome, or a less plausible event occurs with an expected outcome. There is no a priori preference given to occurrence plausibilities or to effect plausibilities; the framework is flexible enough to represent either possibility.

Introducing effect ranking functions makes the distinction between action occurrences and action effects explicit, which in turn gives a straightforward treatment of failed actions. However, we need to introduce some extra machinery in order to determine the most plausible action history. The most general approach is to extend the definition of an aggregate plausibility function: a *non-deterministic aggregate plausibility function* takes a non-deterministic graded world view as an argument, and it returns a plausibility function over histories. An *admissible* non-deterministic aggregate plausibility function is one that increases monotonically with respect to the given graded world view, as well as the given effect ranking function.

We have been using the function *sum* as our standard aggregate plausibility function. The natural extension of *sum* to non-deterministic graded world views is the following. For any history $h = w_0, A_1, \ldots, A_n, w_n$, define

$$sum(\langle ACT, OBS\rangle, \delta)(h) = \sum_i OBS_i(w_i) + ACT_i(A_i) + \delta(A_i, w_{i-1})(w_i).$$

It is easy to see that this is an admissible non-deterministic aggregate function. Returning to the lamp example, there are two minimally ranked histories under this function: one in which the lamp was turned on by pressing on the switch and one in which the lamp was turned on by throwing a piece of paper at the switch.

In the remainder of this section, we will assume that *sum* is the default aggregate function for non-deterministic world views. Under this assumption, we demonstrate that non-deterministic graded world views can be translated into graded world views in an extended action signature.

Let $T = \langle S, V, R\rangle$ be a non-deterministic transition system over the action signature $\langle \mathbf{A}, \mathbf{F}\rangle$. Let $\langle\langle ACT, OBS\rangle, \delta\rangle$ be a non-deterministic graded world view. We extend the action signature to a new action signature $\mathbf{A}'$ where every edge in T corresponds to an action symbol. In particular, let $\mathbf{A}' = \{A_{(s,A,t)} \mid (s, A, t) \in R\}$. Let $T' = \langle S, V, R'\rangle$ where R' is the closure of the set $\{s, A_{(s,A,t)}, t \mid s, t \in S\}$. Suppose that $ACT = ACT_1, \ldots, ACT_n$. Define $ACT' = ACT'_1, \ldots, ACT'_n$ where, for each i, $ACT'_i(A_{(s,A,t)}) = ACT(A) + \delta(A, s)(t)$.

Proposition 23 *For any non-deterministic transition system T, a history*

$$h = w_0, A_1, \ldots, A_n, w_n$$

obtains the minimum rank in $\langle\langle ACT, OBS\rangle, \delta\rangle$ if and only if

$$h' = w_0, A_{(w_0,A_1,w_2)}, \ldots, A_{(w_{n-1},A_n,w_n)}, w_n$$

obtains the minimum rank in $\langle ACT', OBS\rangle$.

Proof The plausibility of h is obtained by taking the sum

$$\sum_i OBS_i(w_i) + ACT_i(A_i) + \delta(A_i, w_{i-1})(w_i),$$

which is clearly the same sum taken to to determine the plausibility of h'. □

Hence non-deterministic actions and failed actions can be represented in a graded world view, simply by setting up the plausibility functions carefully.

We remark that there is a conceptually interesting distinction that is lost in this translation. Informally, there is a distinction between an action that fails to occur and an action that occurs, but fails to produce an expected effect. This distinction is clear if we consider the difference between failing to drop a glass on the ground, and dropping a glass that fails to break when it hits the ground. In the first case, the agent executes the drop action but it fails to occur; perhaps the glass sticks to the agent's hand. In the second case, the glass is successfully dropped without breaking. In our framework, both of these events are represented by a dropping action with the null effect. We suggest that this is an acceptable treatment, because in both cases the sequence of actions and states is identical. As such, we can not distinguish between these scenarios based on our definition of a history. However, we may be able to distinguish indirectly based on the values of other fluents. For instance, the location of the glass is only going to change in the case where it is successfully dropped.

5.4 Comparison with Related Formalisms

5.4.1 Representing Single-Shot Belief Change

In this section, we consider graded world views from the perspective of single-shot belief change; that is, belief change that occurs following a single ontic or epistemic action. Recall that we defined $*$ and $\diamond$ on graded world views as shorthand notation for the associated concatenation operations. Based on the results in this section, it will be clear that this shorthand is natural and appropriate.

We first consider the case of a single ontic action.

Proposition 24 *Let $\langle ACT, OBS\rangle$ be a graded world view. For any plausibility function r over* **A**,

$$Bel(\langle ACT, OBS\rangle \bullet \langle r, 0\rangle) = Bel(\langle ACT, OBS\rangle) \diamond Bel(r).$$

Proof Follows immediately from the assumption that every action is always executable. □

Proposition 24 is important if we are primarily interested in belief states and ontic actions. Basically, in this case, graded world views are unnecessary. The most plausible final belief state can be determined by simply looking at the belief state associated with the initial graded world view.

We now consider the case of a single observation. In the present section, we are primarily interested in comparing the expressive power of graded world views with AGM revision operators. There is one sense in which graded world views are clearly more expressive than AGM operators. In particular, a new observation need not be incorporated into an agent's beliefs if the observation does not come from a reliable source. We will demonstrate that, in the context of a single observation, this is essentially the only difference between a graded world view and an AGM revision operator. More specifically, we will see that the belief change defined by concatenating a single observation onto a graded world view can be captured by an AGM operator, provided that the observation has degree of strength higher than some fixed threshold.

First, we prove that every plausibility function defines a system of spheres. Let r be a plausibility function over X with minimum value $\min_r$. For any n, let $r[n]$ denote the set of complete, consistent theories that are satisfied by some I with $r(I) \leq n$.

Proposition 25 *Let r be a plausible function over a finite action signature. The collection $\mathcal{R} = \{r[n] \mid n \geq \min_r\}$ is a system of spheres centered on $r[\min_r]$.*

Proof Clearly, for each n, $r(n) \subseteq r(n+1)$. Hence $\mathcal{R}$ is totally ordered by $\subseteq$.

If $T \in r[\min_r]$, then T is satisfied by some I with $r(I) \leq \min_r$. But then, for any n, T is satisfied by some I with $r(I) \leq n$. Hence $r[\min_r] \subseteq r[n]$ for all $r[n]$.

Since the action signature is finite, there are only finitely many states. Hence there is a state that is assigned a maximum plausibility, say $\max_r$. Therefore, $r[\max_r]$ is the set of complete, consistent theories.

Let ϕ be a consistent formula. Since there are only finitely many states, there must be a state $w \in \kappa$ such that $r(w) \leq r(v)$ for all $v \in \kappa$. Let $n = r(w)$. Clearly $r(n) \cap \kappa \neq \emptyset$. Now suppose that $U \in \mathcal{S}$ and $U \cap \kappa \neq \emptyset$. Suppose that $U = r(m)$, so U is the set of complete, consistent theories satisfied by some I with $r(I) \leq m$. Since some elements of U are also in κ, it follows $m \geq n$. Therefore $r[n] \subseteq U$, and $r[n]$ is the least sphere intersecting κ. □

Using this result, we can show that single-shot revision under graded world views can be captured by AGM revision operators. We make this claim precise in the next proposition.

Proposition 26 *Let $\langle ACT, OBS \rangle$ be a graded world view. There is an AGM revision function $*$ and a natural number n such that, for any plausibility function r over states with*

degree of strength larger than n,

$$Bel(\langle ACT, OBS\rangle \bullet \langle \lambda \uparrow n, r\rangle) = Bel(\langle ACT, OBS\rangle) * Bel(\alpha).$$

Proof Recall that *plaus* is a plausibility function over histories that is defined by minimizing sums over $\langle ACT, OBS\rangle$, and *plaus_state* is the corresponding plausibility function over final states.

Let n be a natural number such that $n > plaus(h)$ for every history h. Let r be a plausibility function with rank n. It follows that $w \in Bel(\langle ACT, OBS\rangle\langle \lambda \uparrow n, r\rangle)$ if and only if the following conditions hold:

1. $w \in Bel(r)$
2. *plaus_state*(w) is minimal among all states satisfying 1.

By Proposition 25, *plaus_state* defines a system of spheres centered on $Bel($*plaus_state*$)$. It follows from Grove's representation result [39] that there is an AGM revision function $*$ such that, for any observation α, $w \in Bel($*plaus_state*$) * \alpha$ if and only if the following conditions hold:

1. $w \in \alpha$
2. *plaus_state*(w) is minimal among all states satisfying 1.

Setting $\alpha = Bel(r)$ gives the desired result. □

Proposition 26 illustrates that, for a single observation, the most plausible worlds can be determined without considering the history of actions and observations. We can determine the most plausible worlds following an observation by simply abstracting a belief state from a graded world view, then performing AGM revision. It is easy to show that the converse is also true: every AGM revision operator can be represented by a graded world view. More precisely, we have the following result.

Proposition 27 *Let $*$ be an AGM revision operator and let κ be a belief state. There is a graded world view $\langle ACT, OBS\rangle$ with $Bel(\langle ACT, OBS\rangle) = \kappa$ and a natural number n such that, for every non-empty observation α,*

$$\kappa * \alpha = Bel(\langle ACT, OBS\rangle \bullet \langle \lambda \uparrow n, r\rangle)$$

where r is any plausibility function over states where the minimal ranked elements α have degree larger than n.

Proof By Grove's representation result, $*$ can be captured by a system of spheres $\mathcal{S}$. It is straightforward to define $\langle ACT, OBS \rangle$ such that $\mathcal{S}$ is the system of spheres given by Proposition 25. Set n such that $n > plaus(h)$ for every history h. The result is immediate. □

Taken together, Propositions 26 and 27 illustrate that graded world views are equivalent to AGM revision if we restrict attention to a single observation with a sufficiently high degree of reliability. Hence, for single-shot belief change, the full expressive power of graded world views is unnecessary. For both ontic actions and observations, we can define the same belief change operations if we start with just a belief state. Again, there is a correspondence here with Nayak's work on iterated revision [75]; if an observation is sufficiently plausible, then every state in that observation ends up being strictly more plausible then every other state.

5.4.2 Representing Belief Evolution Operators

There are two underlying assumptions in the definition of belief evolution.

1. The plausibility of an observation is determined by some ordering, recency by default.
2. The action history is assumed to be correct.

Both of these assumptions can be represented in a graded world view by setting up the plausibility functions appropriately. Assume that we have a fixed initial belief state κ_I, along with a metric transition system defining a revision operator $*$ and an update operator $\diamond$. Let $\circ$ be the belief evolution operator obtained from $*$ and $\diamond$. Let

$$\bar{A} = \langle A_1, \ldots, A_n \rangle$$

be an action trajectory, and let

$$\bar{\alpha} = \langle \alpha_1, \ldots, \alpha_n \rangle$$

be an observation trajectory. We want to construct a graded world view W_{ev} that assigns minimal plausibility value to all histories corresponding to $\kappa_I \circ \langle \bar{A}, \bar{\alpha} \rangle$.

We define $W_{ev} = \langle ACT, OBS \rangle$ presently. By combining κ_I with the underlying metric d, we can define a plausibility function $BASE$ that represents the initial ordering of states

implicit in $*$. In particular, for any w, set

$$BASE(w) = \min(\{d(w,k) \mid k \in \kappa_I\}).$$

Using this plausibility function, we can define the observation trajectory OBS. Let max denote the maximum value obtained by $BASE$.

$$OBS_i = \begin{cases} BASE & \text{if } i = 0 \\ \alpha_i \uparrow (2^i + \max) & \text{otherwise} \end{cases}$$

By incrementing the plausibility of false observations exponentially, we can assure that recent observations will be given greater credence.

Informally, each action symbol A_i is translated into a plausibility function that obtains the minimum value on the set $\{A_i\}$. Formally, we have the following, for $1 \leq i \leq n$:

$$ACT_i = A_i \uparrow (2^{n+1} + \max).$$

Proposition 28 *If* $\kappa_I \circ \langle \bar{A}, \bar{\alpha} \rangle = \langle \kappa_0, \ldots, \kappa_n \rangle$, *then*

$$h \in \Phi(W_{ev})$$

$$\Longleftrightarrow$$

$$h = \langle w_0, A_1, \ldots, A_n, w_n \rangle \text{ where } w_i \in \kappa_i \text{ for each } i.$$

Proof Assume for the moment that $\langle \bar{A}, \bar{\alpha} \rangle$ is consistent. Let $h = \langle v_0, B_1, \ldots, B_n, v_n \rangle$. By definition $h \in \Phi(W_{ev})$ if and only if the sum

$$\sum_{i=1}^{n} ACT_i(B_i) + \sum_{i=0}^{n} OBS_i(v_i) \tag{5.1}$$

is minimal. Since $\langle \bar{A}, \bar{\alpha} \rangle$ is consistent, there exist histories $\langle w_0, A_1, \ldots, A_n, w_n \rangle$ where each $w_i \in \alpha_i$. For such histories, the sum (5.1) becomes

$$\sum_{i=1}^{n} ACT_i(A_i) + \sum_{i=0}^{n} OBS_i(w_i) = 0 + OBS_0(w_0)$$

We remark that this sum is less than any sum that can be obtained by a history where there is some i such that either $B_i \in A_i$ or $w_i \notin \alpha_i$. Therefore $h \in \Phi(W_{ev})$ if and only if the following three conditions hold:

1. $B_i = A_i$ for each $i > 0$

2. $v_i \in \alpha_i$ for each $i > 0$

3. $OBS_0(v_0)$ is minimal among states satisfying 1 and 2.

In order to satisfy condition 2, it must be the case that v_0 is in the set

$$V = \bigcap_i \alpha_i^{-1}(\bar{A}_i).$$

In order to simultaneously satisfy condition 3, it must also be the case that v_0 is minimally distant from κ_I according to the metric d. In other words, $v_0 \in \kappa_I * V$. Therefore, $h \in \Phi(W_{ev})$ if and only if each $B_i = A_i$ and the following conditions hold:

1. $v_0 \in \kappa_I * \bigcap_i \alpha_i^{-1}(\bar{A}_i)$, and

2. $v_i = v_0 \diamond \bar{A}_i$.

This is the definition of $\kappa_I \circ \langle \bar{A}, \bar{\alpha} \rangle$, so this completes the proof.

The case where $\langle \bar{A}, \bar{\alpha} \rangle$ is inconsistent is similar. The only difference is that we need to notice that the degree of strength of each observation increases by a power of 2. We use the fact that, for any natural number p, 2^p is larger than every sum of terms 2^i with $i < p$. As such, in order to minimize the sum (5.1), we need to work backwards through the observations, keeping each observation if it is consistent with the observations that followed. This is just an equivalent specification of $\tau(W_{ev})$, as given in Definition 19 – increasing powers exponentially forces a strict preference for recent observations. The details of the proof in this case are tedious, but not difficult. □

Proposition 28 demonstrates that graded world views can represent any belief evolution operator defined with respect to a distance function. From the perspective of graded world views, the assumption that action histories are infallible is essentially just a restriction on the admissible plausibility functions.

We conclude this section with some brief remarks about the use of orderings to resolve inconsistency in iterated belief change. The Darwiche-Pearl postulates are only satisfied when we assume that the most recent observation takes precedence over previous observations. By contrast, Papini illustrates an alternative approach to iterated revision in which earlier observations take precedence over later observations[80]. More generally, we defined belief evolution operators with respect to an arbitrary total ordering over the observations.

The most natural extension of belief evolution would extend the ordering to include all observations and actions. Using the techniques in this section, it is easy to see that this extended conception of belief evolution corresponds to the class of graded world views with an arbitrary initial observation followed by plausibility functions of the form $\alpha \uparrow 2^i$, where each i is distinct. Hence, even the most general extension of belief evolution can be represented by a relatively restricted class of graded world views.

5.4.3 Representing Conditionalization

Spohn uses ranking functions to define a form of belief change called *conditionalization* [88]. The idea is that new evidence is presented as a pair (α, m), where α is a set of states and $m \geq 0$; the value of m is an indication of the strength of the observation α. Informally, the conditionalization of r is a new function where the minimally ranked α-worlds receive rank 0 and the non-α worlds are all "shifted up" by m. In this section, we illustrate how conditionalization can be defined in terms of graded world views.

First, we define conditionalization formally. Let r be a plausibility function with $\min_r = 0$ and let α be a subset of the domain of r. Let $\min(\alpha)$ denote the minimum value $r(w)$ for $w \in \alpha$. Spohn defines the the plausibility function $r(\cdot|\alpha)$ over α as follows:

$$r(w|\alpha) = r(w) - \min(\alpha).$$

We call $r(w|\alpha)$ the *α-part of r*. The conditionalization of r, written $r_{(\alpha,m)}$, is the following plausibility function.

$$r_{(\alpha,m)}\ (w) = \begin{cases} r(w|\alpha) & \text{if } w \in \alpha \\ m + r(w|\tilde{\alpha}) & \text{if } w \notin \alpha \end{cases}$$

So the conditionalization of r is the α-part of R together with the $\tilde{\alpha}$-part shifted appropriately.

We illustrate that conditionalization can easily be represented by taking minimal sums over plausibility functions.

Definition 40 *Let r be a plausibility function over $2^{\mathbf{F}}$, let α be a non-empty subset of $2^{\mathbf{F}}$, and let m be a natural number. Define $r_C(\alpha, m)$ as follows:*

$$r_C(\alpha, m)\ (w) = \begin{cases} 0 & \textit{if } w \in \alpha \\ m + \min(\alpha) & \textit{if } w \notin \alpha \end{cases}$$

We refer to $r_C(\alpha, m)$ as the *conditionalizer* of r with respect to α and m. The following proposition illustrates how we can define the conditionalization of a plausibility function by taking an appropriate sum.

Proposition 29 *Let r be a plausibility function with $\min_r = 0$. For any α, m, the normalization of $r + r_C(\alpha, m)$ is the conditionalization $r_{(\alpha,m)}$.*

Proof If $w \in \alpha$, then

$$r(w) + r_C(\alpha, m)(w) = r(w) + 0 = r(w).$$

If $w \notin \alpha$, then

$$r(w) + r_C(\alpha, m)(w) = r(w) + m + \min(\alpha).$$

Since $r(w) \geq 0$ and $m \geq 0$, it follows that the minimum value obtained by $r + r_C(\alpha, m)$ is $\min(\alpha)$. Hence, the normalization of $r + r_C(\alpha, m)$ is the plausibility function r' defined as follows.

$$r'(w) = r(w) + r_C(\alpha, m)(w) - \min(\alpha)$$

It is easy to verify that this is equal to $r_{(\alpha,m)}$. □

Proposition 29 illustrates that the conditionalization of r by (α, m) can be defined by taking a minimal sum over two plausibility functions. We have restricted attention to plausibility functions with minimum 0 because this class coincides more closely with Spohn's ranking functions. However, we can define the conditionalizer in the same manner for plausibility functions with non-zero minimums. We can also define the conditionalization of a graded world view. Informally, we simply conditionalize the associated plausibility function on states. Hence, we identify the conditionalization with respect to $\langle \alpha, m \rangle$ with the following operation:

$$\langle ACT, OBS \rangle \bullet \langle null, plaus\_state_C(\alpha, m) \rangle.$$

It is straightforward to show that this gives the desired result.

5.5 Limitations and Advantages

5.5.1 Constrained World Views

Our focus in previous sections has been on establishing the expressive power of graded world views, as compared with existing frameworks for reasoning about belief change. As a result,

we have focused on problems involving an a priori graded world view, along with some "new" information. However, the restriction to new information is artificial. In the general case, there is no reason to restrict attention to problems in which an agent only receives information about actions and observations occurring at the most recent point in time. An agent could certainly receive new information about earlier events and actions. Hence, a more general problem involves an agent with an underlying graded world view, together with a set of constraints on the most plausible histories. In this section, we consider the representation of problems that have this more general form.

Suppose that $\langle ACT, OBS\rangle$ is a graded world view of length n. An *action constraint* is a pair (A, i) where A is an action symbol and $i \leq n$. Define $\Phi(\langle ACT, OBS\rangle) \restriction (A, i)$ to be the set of histories with minimal plausibility, subject to the restriction that the i^{th} action executed is A. We define *observation constraints* in the analogous manner, and we let $\Phi(\langle ACT, OBS\rangle) \restriction (\alpha, i)$ be the set of minimally ranked histories where the i^{th} state is in α. If Ω is a set of constraints, then we define $\Phi(\langle ACT, OBS\rangle) \restriction \Omega$ to be the set of minimally ranked histories satisfying every constraint in Ω. We will refer to such histories as *constrained histories* and we will refer to a graded world view together with a set of constraints as a *constrained world view.*

We have presented constrained world views to illustrate that graded world views are useful for many problems beyond those that are normally considered to be in the realm of a standard "belief change operator." For example, suppose that Bob sends an encrypted email message to Alice, inviting her to a party at his house. Bob is aware that Eve is the system administrator, and that she could potentially manipulate the message before delivering it. When Alice does not show up, Bob concludes that Eve did not deliver the message. Bob is concerned that Eve read the message and had hurt feelings that she was not invited. However, looking at every possible action Eve could take, Bob concludes that Eve could not have decrypted the message.

In the preceding example, Bob needs to consider all possible actions that Eve could have executed. The conclusion that Bob draws is that Eve's knowledge of the party is invariant with respect to her actions. We can formally define invariance as follows.

Definition 41 *Let $\langle ACT, OBS\rangle$ be a graded world view. We say that a set of worlds α is an i-invariant of $\langle ACT, OBS\rangle$ if and only if, for every $A \in \mathbf{A}$, $Bel(\langle ACT, OBS\rangle \restriction (A, i)) \subseteq \alpha$.*

The intuition behind i-invariance is that, regardless of the action at time i, the underlying

agent will always believe that the actual world is in α. Reasoning about invariant properties is essential if an agent is trying to ensure some property must hold in an action domain involving exogenous actions. This is required, for example, in reasoning about cryptographic protocols.

Reasoning about invariance is just one new kind of problem that can be addressed by constrained world views. We suggest that constraints can also be used to provide natural representations of hypothetical reasoning and abductive reasoning. In the next section, we use constraints to compare graded world views with belief extrapolation operators.

5.5.2 Comparison with Belief Extrapolation

Constrained world views are similar to belief extrapolation operators. We briefly introduced belief extrapolation in Chapter 1, and we refer the reader to [23] for a complete introduction. For the present purposes, it is sufficient to recall that a belief extrapolation operator is defined with respect to an ordering over histories. Given an ordering over histories together with a sequence of formulas, a belief extrapolation operator returns the most plausible sequences of states. In the case of constrained world views, we essentially do the same thing. The given graded world view defines an ordering over states, and the constraints give a sequence of conditions that need to be satisfied. The difference is that the mapping from graded world views to orderings on histories is not surjective; there are orderings on histories that can not be described by a graded world view. For example, a graded world view can not capture plausibilities of the form "if A_1 occurs at time i, then A_2 is likely to occur at time $i+1$." Informally, a graded world view can only represent domains where the ordering on histories is built up in a pointwise manner by the plausibilities at each point in time. In this section, we use this limitation to establish a difference in expressive power between constrained world views and belief extrapolation operators.

First, we need to formalize the problem that we would like to address more precisely. Given a belief extrapolation operator, we would like to be able to find a graded world view that captures the same information.

Definition 42 *Let $\downarrow$ be a belief extrapolation operator. We say that $\downarrow$ is representable if there is a graded world view $\langle ACT, OBS\rangle$ such that, for every scenario Σ of length n,*

$$Traj(\Sigma\downarrow) = \Phi(\langle ACT, OBS\rangle) \restriction \Sigma.$$

If $\updownarrow$ is representable, then the behaviour of $\updownarrow$ can be simulated with a graded world view. We remark that we have abused notation in the definition in that $Traj(\Sigma\ \updownarrow)$ is a collection of sequences of states, whereas $\Phi(\langle ACT, OBS\rangle) \upharpoonright \Sigma$ is a collection of histories. We interpret the equality to mean that the two collections are equal if we ignore the action symbols in the latter.

The following proposition indicates that belief extrapolation operators have an expressive advantage.

Proposition 30 *There is a belief extrapolation operator $\updownarrow$ that is not representable.*

Proof Let $\preceq$ be an ordering in which the following trajectories are minimal.

1. $\langle\{a,b\},\{a,b\},\{\neg a,b\}\rangle$
2. $\langle\{a,b\},\{a,b\},\{a,b\}\rangle$
3. $\langle\{a,\neg b\},\{a,b\},\{\neg a,b\}\rangle$

Let $\updownarrow$ be the associated belief extrapolation operator. We will show that $\updownarrow$ is not representable.

Let $\Sigma = \langle a, a \wedge b, b\rangle$. Note that Σ is satisfied by all three minimal trajectories. Therefore $Traj(\Sigma\ \updownarrow)$ is precisely the set of minimal trajectories.

Now suppose that $\langle ACT, OBS\rangle$ is a graded world view such that

$$\langle ACT, OBS\rangle \upharpoonright \Sigma$$

assigns minimal plausibility to 1, 2, and 3. Hence, there exist actions $A_1, A_2, A_3, B_1, B_2, B_3$ such that the following sums all obtain the minimum possible rank:

1. $OBS_0(\{a,b\}) + ACT_1(A_1) + OBS_1(\{a,b\}) + ACT_2(B_1) + OBS_2(\{\neg a,b\})$
2. $OBS_0(\{a,b\}) + ACT_1(A_2) + OBS_1(\{a,b\}) + ACT_2(B_2) + OBS_2(\{a,b\})$
3. $OBS_0(\{a,\neg b\}) + ACT_1(A_3) + OBS_1(\{a,b\}) + ACT_2(B_3) + OBS_2(\{\neg a,b\})$

It must be the case that $ACT_1(A_1) = ACT_1(A_2)$, because otherwise either 1 or 2 could be reduced by changing the first action. Similarly, it must be the case that $ACT_2(B_1) = ACT_2(B_3)$, because otherwise either 1 or 3 would not be minimal. So, we can rewrite the sums as follows:

1. $OBS_0(\{a,b\}) + ACT_1(A_1) + OBS_1(\{a,b\}) + ACT_2(B_1) + OBS_2(\{\neg a,b\})$
2. $OBS_0(\{a,b\}) + ACT_1(A_1) + OBS_1(\{a,b\}) + ACT_2(B_2) + OBS_2(\{a,b\})$
3. $OBS_0(\{a,\neg b\}) + ACT_1(A_3) + OBS_1(\{a,b\}) + ACT_2(B_1) + OBS_2(\{\neg a,b\})$

From 1 and 2, it follows from basic algebra that

$$ACT_2(B_1) + OBS_2(\{\neg a,b\}) = ACT_2(B_2) + OBS_2(\{a,b\}).$$

Substituting this in 3 gives another minimal sum:

$$OBS_0(\{a,\neg b\}) + ACT_1(A_3) + OBS_1(\{a,b\}) + ACT_2(B_2) + OBS_2(\{a,b\}).$$

This corresponds to the trajectory

$$\langle\{a,\neg b\},\{a,b\},\{a,b\}\rangle.$$

Hence, any graded world view assigning minimum plausibility to 1-3, must also assign minimum plausibility to this fourth trajectory. Informally, if 1-3 are preferred trajectories according to a graded world view, then we are forced to accept another preferred trajectory. But we already saw that $Traj(\Sigma\,\updownarrow)$ consists only of 1-3. Therefore $\updownarrow$ is not representable. □

Note that the proof of Proposition 30 is constructive and it demonstrates that there is a simple, concrete, extrapolation operator that is not representable.

Informally, Proposition 30 follows from the fact that some orderings on histories can not be defined by a graded world view. This is particularly important in applications where an agent has preferences over the order in which events occur. In such applications, it can be useful to assign plausibilities to certain sequences of actions. We suggest, however, that the class of orderings definable by graded world views is a natural class of orderings. In particular, there are many action domains where an agent has no preconceived assumptions about the order that exogenous actions will occur. Graded world views provide a reasonable tool for the representation of such action domains. However, if an agent has some information about the order in which actions tend to occur, then we need arbitrary orderings over histories.

5.5.3 Expressive Advantages of Graded World Views

In the previous section, we saw that constrained world views can not capture every belief extrapolation operator. However, it would be a mistake to conclude that belief extrapolation provides a more expressive framework for reasoning about belief change. In this section, we discuss some of the advantages of graded world views.

First of all, note that belief extrapolation operators are defined for a fixed history length. Given an ordering over histories of length n, it is not clear how to incorporate a new action followed by a new observation; there is no fixed method for extending orderings over n-tuples to orderings over $n+1$ tuples. In the case of a graded world view, however, it is clear how the new ordering is defined when more actions are performed. As such, graded world views are more appropriate for the representation of epistemic action domains where we expect new observations and actions to occur.

The main advantage of graded world views over all of the related formalisms that we have discussed is that graded world views provide a mechanism for dealing with imperfect information. For example, one of the main assumptions underlying belief extrapolation is that every observation should be incorporated in the new scenario. Graded world views allow observations that need not be incorporated. There are two kinds of problems where an observation should not be incorporated immediately. First, there are problems where the observation comes from an unreliable source that may not be trusted. For example, the cookie example can easily be modified to represent the situation where Trent is known to be dishonest, and his report will tend to be ignored. Second, there are problems where an observation comes from a reliable source, but does not provide enough evidence to overthrow the current beliefs. Recall the example where Bob waits for 2 reports before concluding that he left his lamp on. The first report is not ignored; it simply doesn't provide sufficient information to immediately change Bob's beliefs. Graded world views provide a tool for the representation of both of these classes of problems.

We conclude with a brief remark about the overall approach taken in our framework. In Chapter 1, we saw that belief change caused by actions is often represented by starting with an action formalism and then adding revision operators. One problem with this approach is that it does not allow beliefs about action occurrences. In a sense, graded world views take the opposite approach. We start with ranking functions, which were originally defined for reasoning about belief change, and then we plug in actions. An agent's beliefs about

the actions that occur are independent of the formal representation of action effects. As such, although we have presented graded world views in terms of transition systems, it would certainly be possible to use a different action formalism. The key point is that, by using ranking functions to represent uncertainty about states and actions, we can define a framework for reasoning about epistemic action effects in which primary importance is placed on the evolution of an agent's beliefs.

Chapter 6

Conclusion

We have considered iterated belief change in the presence of ontic actions and epistemic actions. Our approach has been based on a simple action formalism in which action effects are given by a transition system. In this final chapter, we offer some concluding remarks.

6.1 Summary

We briefly summarize the results presented in the body of this dissertation. Our goal has been to formalize the belief change that occurs in problems of the following form:

$$(\mathit{InitialBeliefs}) \cdot (\mathit{Action}) \cdot (\mathit{Observation}) \cdots (\mathit{Action}) \cdot (\mathit{Observation}).$$

We assume that the state of the world can be represented by an interpretation over a fixed propositional signature $\mathbf{F}$, and we assume that the effects of actions are given by an underlying transition system T. Under these assumptions, we consider the manner in which an agent's beliefs should change due to a sequence of actions and observations.

We define a belief state to be a set of states, informally the states that are considered possible. An observation is also a set of states, informally the states that are observed to be possible. We define update and revision operators such that our prototypical problem can be restated as follows

$$\kappa \diamond A_1 * \alpha_1 \diamond \cdots \diamond A_n * \alpha_n.$$

Our notion of belief update is actually a form of action progression in which an agent predicts the outcome of an action with conditional effects. We illustrate that successively

applying update and revision operations leads to unintuitive results. We present a set of formal properties that we expect to be satisfied in problems involving alternating updates and revisions, and we introduce a new belief change operator that satisfies our properties. The new operator is called a *belief evolution* operator, and it is defined with respect to a given pair $(\diamond, *)$ consisting of an update operator $\diamond$ and a revision operator $*$. We compare belief evolution operators with some existing approaches for reasoning about belief change caused by action, and we prove a representation result in terms of systems of spheres.

We consider several applications of belief evolution. In particular, we use belief evolution operators to define the semantics of a new epistemic extension of $\mathcal{A}$. We illustrate how to implement a solver for projection problems in the new language though answer set planning. Also, as a somewhat speculative application, we illustrate that belief evolution operators can be used to model the reasoning involved in the verification of authentication protocols.

One limitation of belief evolution operators is that they can not represent uncertainty about the action history. In order to address this problem, we use Spohn-like ranking functions to represent actions and observations. In this case, our prototypical problem consists of a sequence of ranking functions of the form

$$\langle r_I, r_{A_1}, r_{O_1}, \ldots, r_{A_n}, r_{O_n} \rangle$$

where r_I represents the initial belief state, each r_{A_i} is a ranking function representing an action, and each r_{O_i} is a ranking function representing an observation. This is a natural extension of the class of problems addressed by belief evolution operators. Framing the problem in terms of ranking functions allows us to represent uncertainty about the actions that have occurred, as well as observational reliability. We illustrate how to use sequences of ranking functions of this form to represent a wide range of phenomena, and we prove that this approach subsumes several related approaches to belief change.

6.2 Contributions to Existing Research

6.2.1 The Fundamental Contribution

Many different formal approaches have been proposed for reasoning about belief change caused by action [4, 11, 24, 35, 44, 48, 60, 68, 84, 85, 86, 95]. Often, the problem has been treated in a modular fashion by adding revision operators to existing formalisms for reasoning about action effects. As a result, existing work either ignores the interaction

between actions and observations, or else the interaction is treated implicitly. We have illustrated that the interpretation of an observation may depend on the preceding sequence of actions, and it may also depend on the reliability of the observation. Hence, our work makes it clear that a complete treatment of belief change caused by action involves more than simply adding a revision operator to an existing action formalism.

The fact that sequences of actions and observations can not be treated iteratively is the key insight motivating this dissertation. The problems that we have addressed involve an agent that has a formal mechanism for determining the belief change that follows an action, along with a formal mechanism for determining the belief change that follows an observation. As noted above, this is the situation in many existing epistemic action formalisms. Our goal has been to specify and formalize, at a high-level, the manner in which the effects of individual actions and observations should be combined to compositionally determine the belief change following a sequence of actions and observations. Hence, the fundamental contribution of this dissertation is an explicit treatment of *iterated* belief change caused by actions and observations. To the best of our knowledge, there has been no other formal work that explicitly treats this problem.

In the next few sections, we review some of the specific contributions of this work.

6.2.2 Interaction Between Update and Revision

In terms of belief change, every transition system defines a belief update operator. We define an epistemic transition system to be a transition system together with an AGM revision operator. We introduce several natural approaches to revision in a transition system framework, based on notions of distance and path length. Under these revision operators, epistemic transition systems provide a simple, graphically-motivated formalism for reasoning about belief change due to actions and observations.

Reasoning about iterated belief change in an epistemic transition system involves reasoning about alternating sequences of updates and revisions. We present a set of so-called interaction properties P1-P5 that should intuitively be satisfied whenever an update is followed by a revision. There has been no previous work on explicitly specifying the manner in which alternating updates and revisions should be treated. The properties P1-P5 are based on the AGM postulates, suitably translated by the effects of an action. As such, if one is inclined to support the AGM postulates for a single revision, then P1-P5 are easily justified as natural properties for iterated belief change.

Although we have used transition systems to represent the effects of actions, our work makes it clear that the only role played by the transition system is to allow the underlying agent to project the current state to an outcome state that will result due to an action. As such, we could easily frame our results in any action formalism defining action progression and belief revision.

6.2.3 Evaluating Existing Formalisms

Interpreted descriptively, the interaction properties allow us to discern between reasonable approaches to iterated belief change, and naive approaches that do not satisfy the intuitions of the AGM framework. As noted previously, formalisms that determine the effects of actions and observations successively do not satisfy the properties. The existing epistemic extensions of $\mathcal{A}$ are examples of such formalisms. As such, the existing extensions of $\mathcal{A}$ are not appropriate for reasoning about the iterated belief change that occurs in problems like the litmus paper problem. Hence, our work in this dissertation can be used to provide limits on the range of application of certain epistemic action formalisms.

We have illustrated that the epistemic extension of the SitCalc is well suited for the representation of iterated belief change. This result is important for two reasons. First, it provides an interesting and useful epistemic action formalism that satisfies our interaction properties. Second, our work provides some formal justification for the treatment of observations in the SitCalc. From a naive point of view, the fact that "revision actions" do not satisfy the AGM postulates may be seen as a negative. However, our evaluation illustrates that the SitCalc approach is appropriate for iterated belief change; the AGM postulates need not be satisfied when an observation follows a sequence of actions.

6.2.4 Belief Evolution

Belief evolution operators give a prescriptive approach to iterated belief change that respects the interaction between update and revision. The underlying assumption in belief evolution is that action histories are infallible. Under this assumption, belief evolution operators give a reasonable solution to the litmus paper problem. We have demonstrated that the naive combination of update and revision is inappropriate for this problem. As such, one of the contributions of our work is a general solution to litmus-type problems in terms of given update and revision operators. More generally, belief evolution operators intuitively capture

"experimental reasoning" where an agent explores the world by performing state-changing actions, then observing the effects. This kind of reasoning requires an agent to explicitly consider the sequence of actions preceding an observation.

We prove that belief evolution operators can be characterized in terms of modified systems of spheres. This result illustrates that ontic actions can be understood to apply a shifting operation on a system of spheres. This is a simple operation that corresponds closely with our intuitions about the effects of actions, given an a priori plausibility ordering on states of the world. By providing a representation result in terms of systems of spheres, we illustrate that belief evolution is a natural operation in the traditional AGM framework.

Another distinguishing feature of our approach is that, given a sequence of actions and observations, the result of belief evolution is a complete belief trajectory. This makes it salient that agents can not simply focus on determining the new belief state when dealing with both actions and observations. The fact that a complete belief trajectory is returned ensures that an agent's beliefs will be consistent over time. This extends the notion of consistency required for AGM revision. In particular, in the AGM framework, an agent's beliefs must be consistent following an observation. However, if an agent is aware that actions have been performed, then this notion of consistency is too weak. An agent should also believe that the history of states of the world is consistent with the actions that have been executed.

6.2.5 Applications Involving Actions and Observations

This dissertation has made several contributions in terms of applications involving actions and observations. First of all, we have introduced an epistemic extension of $\mathcal{A}$ that provably subsumes two prominent existing extensions of $\mathcal{A}$. Moreover, our extension differs in that we allow erroneous initial beliefs, and it is the only extension that allows non-Markovian belief change respecting the properties P1-P5.

There are relatively few implemented solvers for standard belief change operators. As such, the proposed solver for belief evolution represents a useful contribution because it suggests a novel approach to automating the solution of belief change problems. To the best of our knowledge, answer set planning has not previously been used to implement a solver for belief revision or related belief change operators. We remark that a prototype of our solver has actually been implemented, and it is presently undergoing testing.

The treatment of cryptographic protocols in terms of belief evolution operators is admittedly superficial, but it still makes a contribution to existing research. Protocol logics use monotonic rules of inference to represent changing beliefs, and this is clearly not appropriate in many applications. However, to date there has been no work on the use of non-monotonic belief change operators in reasoning about cryptographic protocols. Since cryptographic protocols involve alternating sequences of actions and observations, belief evolution operators provide a useful model of the belief change that is involved. Hence, using belief evolution operators for protocol verification is useful for two communities. For the security community, belief evolution provides a more accurate model of belief change than that embodied by existing protocol logics. For the belief change community, cryptographic protocols provide an interesting and important class of examples.

6.2.6 Reasoning with Fallible Action Histories

Belief evolution operators provide a complete picture of iterated epistemic action effects under the relatively strict assumption that action histories are infallible. If this restriction is lifted, then it becomes much more difficult to say anything definite based solely on the effects of individual actions. Our work on graded world views makes it clear that the plausibility of actions and observations plays a central role in iterated belief change. Note that the plausibility of an action occurrence is completely independent of the effects of the action. As such, our work illustrates that reasoning about iterated epistemic action effects requires more than an action formalism and an ordering over states. In the general case, we need to have some definite assumptions about the likelihood that an action actually occurs at a given point in time.

Graded world views provide a single formalism that is suitable for reasoning about epistemic action effects involving fallible beliefs, fallible perception, exogenous actions, and failed actions. We are not aware of any other existing action formalism that is able to represent all of these phenomena.

We have precisely delineated a large class of problems that can be captured by graded world views. In particular, graded world views are well-suited for reasoning about problems where each action and observation comes with an attached plausibility, and these plausibilities are mutually independent. We have proved that this class of problems includes the problems addressed by subjective probability functions, AGM revision operators, and Spohn's conditionalization. As such, graded world views are a natural extension of existing

work in belief change. Further, we have illustrated by example that graded world views can also capture interesting new classes of problems, such as those involving graded evidence. Such problems are not typically captured in formal approaches to representing belief change due to actions.

We see two main applications for graded world views. First, as suggested above, graded world views provide a flexible framework that can be used to capture a variety of phenomena that may not be captured in existing formalisms. Second, graded world views can serve as a unifying tool that can be used to compare and contrast the expressive power of related formalisms.

6.3 Future Work

There are several interesting directions for future work. One obvious direction would be to work towards a syntactic treatment of iterated belief change in the presence of multiple observations. We have only provided postulates for a single action followed by a single observation. Belief evolution operators and graded world views both provide semantic tools for resolving conflicting observations at different points in time, but we have not attempted to syntactically capture this phenomenon. Existing work in this area includes [21], in which properties of iterated sequences of observations are described through a set of postulates for prioritized merging. The postulates do not characterize a specific approach to iterated belief change, and they do not incorporate ontic actions. However, axiomatizing prioritized merging may be a reasonable first step towards characterizing belief evolution operators for multiple observations.

Another direction for future research would address iterated belief change caused by actions in a multi-agent environment. We remark that even defining update and revision operators in this context can be difficult, because a single action can affect the beliefs of each agent in a different way. However, the multi-agent belief structures of [44] provide a reasonable tool for the representation of a single update or a single revision. By employing the methodology of belief evolution, it should be possible to extend multi-agent belief structures to represent iterated belief change in a manner that respects our interaction properties. Alternatively, one could also extend the formalism with plausibility functions for each agent to represent imperfect information.

On the practical side, we would like to work towards a more complete treatment of

cryptographic protocol verification. Since cryptographic protocols typically involve several agents, this project would necessarily have to follow the aforementioned multi-agent extension of our work. However, there are real advantages to be gained by framing protocol verification in terms of a general approach to belief change. First of all, we have suggested that there are obvious problems with the use of monotonic logics to reason about belief change in a protocol run. Moreover, existing protocol logics have often been defined specifically for authentication protocols. Related protocol goals, such as anonymity and non-repudiation, often require the introduction of new forms of reasoning. Using a general approach to belief change may facilitate the representation of a wide range of protocol goals in a single framework.

6.4 Final Remarks

Iterated belief change due to actions and observations occurs in many natural problem domains. Despite this fact, there has been very little work on the formalization of this kind of reasoning. We suggest that, in this respect, there is a close parallel between iterated revision and iterated epistemic action effects. In particular, AGM revision operators were originally defined for a single revision. Over time, it became clear that iterated revision introduces new complications that must be explicitly addressed. Similarly, belief change caused by actions has typically been addressed for a single ontic action or a single epistemic action. Just as iterated revisions introduce new complications, we have illustrated that iterated actions and observations also introduce new complications. We suggest that the epistemic effects of iterated actions and observations should be discussed and addressed as a distinct problem, just as iterated revision has been discussed and addressed as a distinct problem. We view this dissertation as the initial formal treatment of a natural phenomenon in belief change, and we hope that the formal tools introduced will provide a useful foundation for future work.

Bibliography

[1] M. Adadi and M. Tuttle. A semantics for a logic of authentication. In *Proceedings of the 10th ACM Symposium on Principles of Distributed Computing*, pages 201–216. ACM Press, August 1991.

[2] N. Agray, W. van der Hoek, and E. de Vink. On BAN logics for industrial security protocols. In B. Dunin-Keplicz and E. Nawarecki, editors, *Proceedings of CEEMAS 2001*, volume 2296 of *Lecture Notes in Artificial Intelligence*, pages 29–36. Springer-Verlag, 2002.

[3] C.E. Alchourrón, P. G ardenfors, and D. Makinson. On the logic of theory change: Partial meet functions for contraction and revision. *Journal of Symbolic Logic*, 50(2):510–530, 1985.

[4] A. Balgtag, L. Moss, and S. Solecki. 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, 1998.

[5] C. Baral and M. Gelfond. Reasoning about effects of concurrent actions. *Journal of Logic Programming*, 31(1-3):85–117, 1997.

[6] C. Baral, M. Gelfond, and A. Provetti. Representing actions: Laws, observations and hypothesis. *Journal of Logic Programming*, 31(1-3):201–243, 1997.

[7] C. Baral and N. Tran. Representation and reasoning about evolution of the world in the context of reasoning about actions. In *Proceedings of The Workshop on Nonmonotonic Reasoning, Action, and Change (NRAC'03)*, pages 31–36, August 2003.

[8] B. Bart, J.P. Delgrande, and O. Schulte. Knowledge and planning in an action-based multi-agent framework: A case study. In E. Stroulia and S. Matwin, editors, *Advances in Artificial Intelligence, 14th Biennial Conference of the Canadian Society for Computational Studies of Intelligence*, volume 2056 of *Lecture Notes in Computer Science*, pages 63–136. Springer-Verlag, 2001.

[9] A. Bleeker and L. Meertens. A semantics for BAN logic. In *Proceedings of DIMACS Workshop on Design and Formal Verification of Security Protocols*, February 1997.

[10] R. Booth, S. Chopra, and T. Meyer. Restrained revision. In *Proceedings of The Workshop onNonmonotonic Reasoning, Action, and Change (NRAC'05)*, pages 15–21, August 2005.

[11] C. Boutilier. Generalized update: Belief change in dynamic settings. In *Proceedings of the Fourteenth International Joint Conference on Artificial Intelligence (IJCAI 1995)*, pages 1550–1556, 1995.

[12] M. Burrows, M. Abadi, and R. Needham. Authentication: A practical study of belief in action. In M. Vardi, editor, *Proceedings of the 2nd Conference on Theoretical Aspects of Reasoning about Knowledge*, pages 325–342. Morgan Kaufmann, March 1988.

[13] M. Burrows, M. Abadi, and R. Needham. A logic of authentication. Technical Report 39, Digital Systems Research Center, February 1989.

[14] M. Burrows, M. Abadi, and R. Needham. A logic of authentication. *ACM Transactions on Computer Systems*, 8(1):18–36, 1990.

[15] L. Carlucci Aiello and F. Massacci. An executable specification language for planning attacks to security protocols. In P. Syverson, editor, *Proceedings of the IEEE Computer Security Foundation Workshop*, pages 88–103. IEEE Computer Security Press, 2000.

[16] L. Carlucci Aiello and F. Massacci. Verifying security protocols as planning in logic programming. *ACM Transactions on Computational Logic*, 2(4):542–580, 2001.

[17] B. Chellas. *Modal Logic: An Introduction.* Cambridge University Press, 1980.

[18] M. Dalal. Investigations into a theory of knowledge base revision. In *Proceedings of the National Conference on Artificial Intelligence (AAAI88)*, pages 475–479, 1988.

[19] A. Darwiche and J. Pearl. On the logic of iterated belief revision. *Artificial Intelligence*, 89(1-2):1–29, 1997.

[20] J.P. Delgrande. Preliminary considerations on the modelling of belief change operators by metric spaces. In *Proceedings of the 10th International Workshop on Non-Monotonic Reasoning (NMR 2004)*, pages 118–125, June 2004.

[21] J.P. Delgrande, D. Dubois, and J. Lang. Iterated revision as prioritized merging. In *Proceedings of the 10th International Conference on Principles of Knowledge Representation and Reasoning (KR2006)*, June 2006.

[22] D. Dolev and A.C. Yao. On the security of public key protocols. *IEEE Transactions on Information Theory*, 2(29):198–208, 1983.

[23] F. Dupin de Saint-Cyr and J. Lang. Belief extrapolation (or how to reason about observations and unpredicted change). In *Proceedings of the 8th International Conference on Principles of Knowledge Representation and Reasoning (KR2002)*, pages 497–508, 2002.

[24] T. Eiter, W. Faber, N. Leone, G. Pfeifer, and A. Polleres. Planning under incomplete knowledge. In J. Lloyd, V. Dahl, U. Furbach, M. Kerber, K. Lau, C. Palamidessi, L.M. Pereira, Y. Sagiv, and P.J. Stuckey, editors, *Proceedings of the First International Conference on Computational Logic (CL 2000)*, volume 1861 of *Lecture Notes in Artificial Intelligence*, pages 807–821. Springer-Verlag, July 2000.

[25] T. Eiter, W. Faber, N. Leone, G. Pfeifer, and A. Polleres. Answer set planning under action costs. In G. Ianni and S. Flesca, editors, *Proceedings of the 8th European Conference on Artificial Intelligence (JELIA)*, volume 2424 of *Lecture Notes in Artificial Intelligence*, pages 541–544. Springer-Verlag, September 2002.

[26] T. Eiter, W. Faber, N. Leone, G. Pfeifer, and A. Polleres. The DLV $\mathcal{K}$ planning system: Progress report. In G. Ianni and S. Flesca, editors, *Proceedings of the 8th European Conference on Artificial Intelligence (JELIA)*, volume 2424 of *Lecture Notes in Artificial Intelligence*, pages 541–544. Springer-Verlag, September 2002.

[27] T. Eiter and G. Gottlob. On the computational cost of disjunctive logic programming: Propositional case. *Annals of Mathematics and Artificial Intelligence*, 15(3/4):289–323, 1995.

[28] R. Fagin, J. Halpern, Y. Moses, and M. Vardi. *Reasoning About Knowledge.* MIT Press, 1995.

[29] K. Forbus. Introducing actions into qualitative simulation. In *Proceedings of the International Joint Conference on Artificial Intelligence (IJCAI89)*, pages 1273–1278, 1989.

[30] M. Gelfond, C. Baral, and G. Gonzalez. Alan: An action language for non-Markovian domains. In *IJCAI-03 Workshop on Nonmonotonic Reasoning, Action and Change (NRAC03)*, 2003.

[31] M. Gelfond and V. Lifschitz. The stable model semantics for logic programming. In R. Kowalski and K. Bowen, editors, *Proceedings of the 5th International Conference on Logic Programming*, pages 1070–1080. MIT-Press, 1988.

[32] M. Gelfond and V. Lifschitz. Classical negation in logic programs and disjunctive databases. *New Generation Computing*, 9(3/4):365–385, 1991.

[33] M. Gelfond and V. Lifschitz. Representing action and change by logic programs. *Journal of Logic Programming*, 17:301–321, 1993.

[34] M. Gelfond and V. Lifschitz. Action languages. *Linköping Electronic Articles in Computer and Information Science*, 3(16):1–16, 1998.

[35] J. Gerbrandy and W. Groenveld. Reasoning about information change. *Journal of Logic Language and Information*, 6(2), 1997.

[36] E. Giunchiglia, J. Lee, V. Lifschitz, N. McCain, and H. Turner. Nonmonotonic causal theories. *Artificial Intelligence*, 153(1-2):49–104, 2003.

[37] E. Giunchiglia and V. Lifschitz. An action language based on causal explanation: preliminary report. In *Proceedings of The National Conference on Artificial Intelligence(AAAI98)*, pages 623–630, 1998.

[38] E. Giunchiglia and V. Lifschitz. Action languages, temporal action logics and the situation calculus. *Linköping Electronic Articles in Computer and Information Science*, 4(40):1–19, 1999.

[39] A. Grove. Two modellings for theory change. *Journal of Philosophical Logic*, 17:157–170, 1988.

[40] J. Guttman and J. Thayer. Authentication tests. In *Proceedings 2000 IEEE Symposium on Security and Privacy*, May 2000.

[41] J. Halpern and R. Pucella. On the relationship between strand spaces and multi-agent systems. *ACM Transactions on Information and System Security (TISSEC)*, 6(1), February 2003. (to appear).

[42] S. Hanks and D. McDermott. Nonmonotonic logic and temporal projection. *Artificial Intelligence*, 33(3):379–412, 1987.

[43] J. Hernández-Orallo and J. Pinto. Especificación formal de protocolos criptográficos en cálculo de situaciones. *Novatica*, 143:57–63, 2000. English version at http://www.dsic.upv.es/ jorallo/escrits/escritsa.htm.

[44] A. Herzig, J. Lang, and P. Marquis. Revision and update in multi-agent belief structures. In *Proceedings of LOFT 6*, 2004.

[45] A. Hunter. Adding modal operators to the action language $\mathcal{A}$. In *Proceedings of the 10th International Workshop on Non-Monotonic Reasoning (NMR 2004)*, pages 219–226, June 2004.

[46] A. Hunter and J.P. Delgrande. Iterated belief change: A transition system approach. In *Proceedings of the International Joint Conference on Artificial Intelligence (IJCAI05)*, pages 460–465, August 2005.

[47] A. Hunter and J.P. Delgrande. Belief change in the context of fallible actions and observations. In *Proceedings of the National Conference on Artificial Intelligence(AAAI06)*, July 2006.

[48] Y. Jin and M. Thielscher. Representing beliefs in the fluent calculus. In *Proceedings of the European Conference on Artificial Intelligence(ECAI04)*, 2004.

[49] O. Kahramanoğulları and M. Thielscher. A formal assessment result for fluent calculus using the action description language $\mathcal{A}_k$. In R. Kruse, editor, *Proceedings of the German Annual Conference on Artificial Intelligence (KI)*, volume 2821 of *Lecture Notes in Artificial Intelligence*, pages 209–223. Springer-Verlag, 2003.

[50] L. Karlsson and J. Gustafsson. Reasoning about actions in a multi-agent environment. *Linköping Electronic Articles in Computer and Information Science*, 2(14), 1997.

[51] H. Katsuno and A.O. Mendelzon. On the difference between updating a knowledge base and revising it. In *Proceedings of the Second International Conference on Principles of Knowledge Representation and Reasoning (KR 1991)*, pages 387–394, 1991.

[52] H. Katsuno and A.O. Mendelzon. On the difference between updating a knowledge base and revising it. In Peter G ardenfors, editor, *Belief Revision*, pages 183–203. Cambridge University Press, 1992.

[53] H. Kautz and B. Selman. Planning as satisfiability. In *Proceedings of the European Conference on Artificial Intelligence (ECAI92)*, pages 359–363, 1992.

[54] P. Koksal, N. Cicekli, and I. Toroslu. Specification of workflow processes using the action description language $\mathcal{C}$. *http://citeseer.nj.nec.com/488894.html*, 2001.

[55] S. Konieczny, J. Lang, and P. Marquis. Distance-based merging: a general framework and some complexity results. In *Proceedings of the 8th International Conference on Principles of Knowledge Representation and Reasoning (KR2002)*, pages 97–108, 2002.

[56] S. Konieczny and R. Pino-Perez. On the logic of merging. In *Proceedings of the Sixth International Conference on Principles of Knowledge Representation and Reasoning (KR98)*, pages 488–498, 1998.

[57] S. Kremer and J. Raskin. A game-based verification of non-repudiation and fair exchange protocols. In *CONCUR 2001-Concurrency Theory*, volume 2154 of *Lecture Notes in Computer Science*. Springer-Verlag, 2001.

[58] J. Lang. About time, revision, and update. In *Proceedings of the 11th International Workshop on Non-Monotonic Reasoning (NMR 2006)*, June 2006.

[59] D. Lehmann. Belief revision, revised. In *Proceedings of the Fourteenth International Joint Conference on Artificial Intelligence (IJCAI95)*, pages 1534–1541, August 1995.

[60] H. Levesque. What is planning in the presence of sensing? In *Proceedings of the Thirteenth National Conference on Artificial Intelligence (AAAI96)*, pages 1139–1146, 1996.

[61] H. Levesque, F. Pirri, and R. Reiter. Foundations for the situation calculus. *Linköping Electronic Articles in Computer and Information Science*, 3(18):1–18, 1998.

[62] P. Liberatore and M. Schaerf. Brels: A system for the integration of knowledge bases. In *Proceedings of KR2000*, pages 145–152. Morgan Kaufmann Publishers, 2000.

[63] O. Lichtenstein and A. Pnueli. Propositional temporal logics: Decidability and completeness. *Logic Journal of the IGPL*, 8(1):55–85, 2000.

[64] V. Lifschitz. Two components of an action language. *Annals of Mathematics and Artificial Intelligence*, 21:305–320, 1997.

[65] V. Lifschitz. Action languages, answer sets and planning. In K.R. Apt, V.W. Marek, M. Truszczynski, and D.S. Warren, editors, *The Logic Programming Paradigm: A 25-Year Perspective*. Springer-Verlag, 1999.

[66] V. Lifschitz and H. Turner. Representing transition systems by logic programs. In *Proceedings of the Fifth Int'l Conf. on Logic Programming and Nonmonotonic Reasoning (LPNMR 99)*, pages 92–106, 1999.

[67] F. Lin. Embracing causality in specifying the indirect effects of action. In *Proceedings of IJCAI-95*, pages 1985–1991, 1995.

[68] J Lobo, G. Mendez, and S.R. Taylor. Knowledge and the action description language $\mathcal{A}$. *Theory and Practice of Logic Programming*, 1(2):129–184, 2001.

[69] N. McCain and H. Turner. Causal theories of action and change. In *Proceedings of AAAI-97*, pages 460–465, 1997.

[70] N. McCain and H. Turner. Satisfiability planning with causal theories. In A. Cohn, L. Schubert, and S. Shapiro, editors, *Proceedings of the Sixth International Conference on Principles of Knowledge Representation and Reasoning*, pages 212–223, 1998.

[71] J. McCarthy. Programs with common sense. In *Proceedings Teddington Conference on the Mechanization of Thought Processes*, pages 75–91. Her Majesty's Stationary Office, 1959.

[72] J. McCarthy. Elaboration tolerance. In *Proceedings of Common Sense 98*, 1998.

[73] C. Meadows. Open issues in formal methods for cryptogrphic protocol analysis. In *Proceedings of DISCEX 2000*, pages 237–250. IEEE Computer Society Press, January 2000.

[74] R.C. Moore. A formal theory of knowledge and action. In J.R. Hobbs and R.C. Moore, editors, *Formal Theories of the Commonsense World*, pages 319–358. Ablex Publishing, 1985.

[75] A.C. Nayak. Iterated belief change based on epistemic entrenchment. *Erkenntnis*, 41:353–390, 1994.

[76] A.C. Nayak, M. Pagnucco, and P. Peppas. Dynamic belief change operators. *Artificial Intelligence*, 146:193–228, 2003.

[77] I. Niemelä and P. Simons. Smodels - an implementation of stable model and well-founded semantics for normal logic programs. In *Proceedings of the 4th International Conference on Logic Programming and Non-Monotonic Reasoning*, volume 1265 of *Lecture Notes in Artificial Intelligence*, pages 141–151. IEEE Computer Society Press, 1992.

[78] I. Niemelä and P. Simons. Extending the smodels system with cardinality and weight constraints. In Jack Minker, editor, *Logic-Based Artificial Intelligence*, pages 491–521. Kluwer Academic Publishers, 2000.

[79] M. Pagnucco and P. Peppas. Causality and minimal change demystified. In *Proceedings of the Seventeenth International Joint Conference on Artificial Intelligence (IJCAI'01)*, pages 125 – 130, 2001.

[80] O. Papini. Iterated revision operations stemming from the history of an agent's observations. In H. Rott and M. Williams, editors, *Frontiers in Belief Revision*, pages 279–301. Kluwer Academic Publishers, 2001.

[81] P. Peppas, A. Nayak, M. Pagnucco, N. Foo, and M. Prokopenko. Revision vs. update: Taking a closer look. In *Proceedings of the Twelfth European Conference on Artificial Intelligence (ECAI96)*, pages 95–99, 1996.

[82] R.B. Scherl and H.J. Levesque. Knowledge, action, and the frame problem. *Artificial Intelligence*, 144(1):1–39, 2003.

[83] O. Schulte and J.P. Delgrande. Representing von neumann-morgenstern games in the situation calculus. In *Proceedings of the AAAI Workshop on Game Theoretic and Decision Theoretic Agents*, number WS-02-06 in AAAI Technical Report. The AAAI Press, 2002.

[84] S. Shapiro and M. Pagnucco. Iterated belief change and exogenous actions in the situation calculus. In *Proceedings of the Sixteenth European Conference on Artificial Intelligence (ECAI'04)*, pages 878–882, 2004.

[85] S. Shapiro, M. Pagnucco, Y. Lesperance, and H.J. Levesque. Iterated belief change in the situation calculus. In *Proceedings of the Seventh International Conference on Principles of Knowledge Representation and Reasoning (KR 2000)*, pages 527–538. Morgan Kaufmann Publishers, 2000.

[86] T. Son and C. Baral. Formalizing sensing actions: A transition function based approach. *Artificial Intelligence*, 125(1-2):19–91, 2001.

[87] T. Son, P. Huy, and C. Baral. Planning with sensing actions and incomplete information using logic programming. In *Proceedings of LPNMR7*, 2004.

[88] W. Spohn. Ordinal conditional functions. A dynamic theory of epistemic states. In W.L. Harper and B. Skyrms, editors, *Causation in Decision, Belief Change, and Statistics, vol. II*, pages 105–134. Kluwer Academic Publishers, 1988.

[89] V. S. Subrahmanian and Carlo Zaniolo. Relating stable models and AI planning domains. In *International Conference on Logic Programming*, pages 233–247, 1995.

[90] P. Syverson. Knowledge, belief, and semantics in the analysis of cryptographic protocols. *Journal of Computer Security*, 1:317–334, 1992.

[91] P. Syverson and I. Cervesato. The logic of authentication protocols. In R. Focardi and R. Gorrieri, editors, *Foundations of Security Analysis and Design*, volume 2171 of *Lecture Notes in Computer Science*, pages 63–136. Springer-Verlag, 2001.

[92] P. Syverson and P. van Oorschot. A unified cryptographic protocol logic. Technical Report 5540-227, Naval Research Lab, 1996.

[93] H. Turner. A logic of universal causation. *Artificial Intelligence*, 113:87–123, 1999.

[94] W. van der Hoek and M. Wooldridge. Tractable multiagent planning for epistemic goals. In *Proceedings of the First International Conference on Autonomous Agents and Multiagent Systems (AAMAS-02),*. ACM Press, 2002.

[95] H.P. van Ditmarsch. Descriptions of game actions. *Journal of Logic, Language and Information (JoLLI)*, 11:349–365, 2002.

Index