Conditional Logics for Default Reasoning and Belief Revision

by
Craig Boutilier

Technical Report 92-1
January 1992

Department of Computer Science
University of British Columbia
Rm 333 - 6356 Agricultural Road
Vancouver, B.C.
CANADA V6T 1Z2

Conditional Logics for Default Reasoning and Belief Revision

Craig Boutilier

Department of Computer Science
University of British Columbia
Vancouver, British Columbia, Canada

A Dissertation submitted in conformity with the requirements
for the Degree of Doctor of Philosophy in the University of Toronto

Printed January 10, 1992

Abstract

Much of what passes for knowledge about the world is defeasible, or can be mistaken. Our perceptions and premises can never be certain, we are forced to jump to conclusions in the presence of incomplete information, and we have to cut our deliberations short when our environment closes in. For this reason, any theory of artificial intelligence requires at its heart a theory of *default reasoning*, the process of reaching plausible, but uncertain, conclusions; and a theory of *belief revision*, the process of retracting and adding certain beliefs as information becomes available.

In this thesis, we will address both of these problems from a logical point of view. We will provide a semantic account of these processes and develop conditional logics to represent and reason with default or *normative* statements, about normal or typical states of affairs, and statements of belief revision. The conditional logics will be based on standard modal systems, and the possible worlds approach will provide a uniform framework for the development of a number of such systems.

Within this framework, we will compare the two types of reasoning, determining that they are remarkably similar processes at a formal level of analysis. We will also show how a number of disparate types of reasoning may be analyzed within these modal systems, and to a large extent unified. These include normative default reasoning, probabilistic default reasoning, autoepistemic reasoning, belief revision, subjunctive, hypothetical or counterfactual reasoning, and abduction.

Acknowledgements

I am indebted to a number of people without whose advice and encouragement this work would not have been possible. Foremost on this list is Ray Reiter, who has supervised this thesis, as well as my Master's thesis, over the past five years. His guidance and encouragement throughout the course of this work have been invaluable, as has his confidence to allow me to wander about only somewhat aimlessly at times. His view of AI has influenced my own beyond measure.

The members of my committee have had tremendous influence on my research, perhaps more than is usual. Hector Levesque's work on the logical foundations of knowledge representation lies at the heart of much of this work, and his perspective should be in evidence here. It was Alberto Mendelzon who piqued my interest in belief revision and at least half of this thesis is directly attributable to this. Discussions with both Hector and Alberto have been extremely productive. Graeme Hirst's valuable comments on the thesis have forced me to examine the philosophical foundations of this work more than I might have otherwise.

My research has, of course, been influenced by the work of many people, but the work of Jim Delgrande started me down the road of conditional logic and I haven't looked back yet (well, not too often). As is evident in the body of this thesis, the work of Peter Gärdenfors, Daniel Lehmann and Judea Pearl has had a tremendous impact on my own. I would also like to thank each of them for generously taking the time to discuss or comment on my research, and especially Professor Lehmann, who served as my external examiner. Thanks also to Alasdair Urquhart, who served as internal examiner, and David Spring, who sat on my committee as well.

Many other people have played a role in the development of this thesis, either through brief comments (of whose influence they may be unaware), extended discussion or periodic correspondence. I am grateful to the following people (hoping to keep omissions to a minimum): Fahiem Bacchus, Luis Fariñas del Cerro, Moisés Goldszmidt, Gösta Grahne, Russ Greiner, Jeff Horty, Hirofumi Katsuno, Gerhard Lakemeyer, Philippe Lamarre, David Makinson, Wiktor Marek, Michael Morreau, David Poole, Karl Schlechta, Bart Selman and Mirek Truszczyński.

I'd like to thank all my friends in Toronto for stimulating conversations on all sorts of (academic and not so academic) topics, lots of good times, and, of course, lots of soccer, hockey, softball, volleyball and one broken ankle. My family has been a constant source of inspiration and support, and I'd especially like to acknowledge my parents in this regard. But none of this would have been possible without my wife. Nicki, I don't know how you put up with it all for the past year, but ...*Phew!*

Contents

List of Figures

List of Symbols

SYMBOL	DEFINED ON PAGE (FIRST APPEARS ON)	MEANING
$\neg$	44 (11)	classical negation
$\wedge$	44 (11)	classical conjunction
$\vee$	44 (11)	classical disjunction
$\supset$	44 (12)	material implication
$\equiv$	44	material equivalence
$\Box$	44 (23)	necessity (accessible)
$\Diamond$	44 (23)	possibility (accessible)
$\overline{\Box}$	84 (76)	necessity (inaccessible)
$\overline{\overline{\Box}}$	84 (84)	necessity (all worlds)
$\overline{\Diamond}$	84 (84)	possibility (inaccessible)
$\overline{\overline{\Diamond}}$	84 (84)	possibility (all worlds)
$O(KB)$	106	only know KB (CO, etc.)
B	140 (140)	belief (OL)
O	140 (102)	only know (OL)
N	140	believe at most (OL)
$\boxtimes$	139 (137)	belief (CO, etc.)
$\overleftarrow{\boxtimes}$	143	believe at most (CO, etc.)
$\Rightarrow$	88 (7)	normal implication (CO, etc.)
$\Rightarrow$	47 (41)	normal implication (CT4, etc.)
$\xrightarrow{\text{KB}}$	108 (7)	revision conditional
$>$	92 (76)	conditional believe at most (CO, etc.)
$>$	124 (27)	abstract subjunctive (VC, C2)
$\rightarrow$	67	default rule (ε-semantics)
$\Rightarrow$	67	strict default rule (ε-semantics)
$\mid\!\sim$	61	nonmonotonic consequence relation
$\models$	11	satisfaction/entailment
$\vdash$	11	provability/derivability
$\|\|A\|\|$	24	worlds satisfying A
Cn	11	logical consequence
$\vdash_1$	80	1-entailment
$Cl(T)$	96	closure (default theory)
$R_i^\wedge$	96	closure axioms
$\models_\leq$	94	preferential entailment
Cn_P	80	preferential consequence
Cn_R	81	rational closure
$\mathcal{R}_e$	99	relevance relation

List of Definitions

Chapter 1

Introduction

They arrive at their favorite pub after Craig has just soundly trounced Pete in their weekly tennis match. Craig is thirstily awaiting the loser-bought first round when Pete casually mentions something about being disappointed that hockey season has ended. Impatiently, Craig chastises Pete for the stalling tactics and small talk, urging him to get up and buy some beer. Smugly, Pete announces that Craig still owes him a night out because of a wager on the Stanley Cup (hockey) championships, and Craig, conceding this round of gamesmanship, heads to the bar for the first round.

This scene illustrates a number of important aspects of commonsense reasoning. First, we notice Craig is reasoning on the basis of expectations about the typical or normal state of affairs. Normally the loser buys the first round and Craig has no reason to expect otherwise tonight, so he sits down fully anticipating Pete's going to get a pitcher. Craig also discounts Pete's initial comment about hockey as being irrelevant to the current situation. It doesn't affect Craig's expectation (or at least he hasn't yet admitted it!). Finally, Craig is forced by overwhelming evidence to the contrary to admit he will end up paying for the night. Not only is he forced to retract the belief in his initial conclusion, but also a number of other beliefs, such as the fact that he will get away cheaply tonight, and that he will get home early.

Reasoning about normal or prototypical situations, jumping to conclusions, or making assumptions is commonly referred to as *default reasoning*, and the ubiquity of this process in commonsense reasoning has been widely acknowledged. In just about any situation one can imagine, certain assumptions must be made, and certain inferences must be drawn in the face of incomplete or inconclusive information. We must expect that conclusions we reach are, in general, merely plausible (to greater or lesser extent) and fallible, rather than certain or infallible. If we were to ever suspend belief in facts of whose truth we were not entirely assured, we would be in a position to know very little (if anything), much to the delight of the philosophical skeptics. However, contemporary epistemologists have increasingly come to accept that knowledge may be based on *defeasible* reasons (Pollock 1986) and, hence, that beliefs (even those that count as knowledge) cannot be arrived at with logical certainty. That Pete buys the first beer might be a reasonable inference, but it is not certain. Believing that Aunt Martha's pet Tweety can fly because it is a bird is also not guaranteed — it might turn out to be a penguin, or dead.

Since most of our beliefs are arrived at through default reasoning and are based on defeasible reasons, it should not be surprising that our beliefs often turn out to be mistaken, justified though they might be. The conclusion that Pete buys the first round given that he has lost the match seems reasonable, but turned out to be wrong. If Tweety has a broken wing, inferring that Tweety can fly will probably be wrong as well. When we learn of the wing, we must revise our beliefs.

Not only do we add to our stock of beliefs that Tweety has a broken wing, but we also remove the fact that Tweety can fly (at the present time), perhaps that Aunt Martha keeps Tweety at a safe distance from the cat, or maybe even that Aunt Martha is a nice person (who doesn't mistreat pets).

Because default reasoning pervades commonsense (and not so common) reasoning, any general theory of artificial intelligence (AI) must include some account of the process of reasoning by default, as well as of *belief revision*, the process of revising states of belief in the presence of new, often contradictory, beliefs or evidence. It is usually admitted that any intelligent agent, artificial or otherwise, must have the capability to possess knowledge of its environment and itself, and to reason with that knowledge to infer new beliefs and expectations from which it can make decisions regarding future actions and goals. In order to characterize the types of knowledge required by such an agent, and the reasoning it should perform, an account of default reasoning and belief revision is crucial, as most knowledge and reasoning will be of this form.

A number of logical and procedural accounts[1] of default reasoning and belief revision have been proposed over the last ten to fifteen years, in both the AI and philosophical communities. The logical accounts have become predominant, and in this thesis we will provide a logical account of both default reasoning and belief revision, and examine the relationship between the two. Our treatment will be in terms of *conditional logic*, whereby we represent default rules and statements of revision in conditional form. For instance, a default about the social aspects of tennis playing may be phrased as "If Pete loses at tennis, he buys the beer," expressing a statement about typical or normal states of affairs. A statement of revision might be "If I am forced to believe I'm buying, then I'll cease believing I'll get home early." We will propose a logic within which one can represent and reason with statements of both types. While default reasoning and revision appear to be quite distinct (though complementary) forms of reasoning, the conditionals and logic for representing them will turn out to be the same in each case, allowing us to differentiate the two processes using a technically uniform framework for comparison. This framework will also be general enough to allow us to show a number of correspondences with existing characterizations of defaults and revision.

Before getting on with the business of developing such a characterization, a few words regarding the role of logic in AI, and especially in knowledge representation and reasoning (KR), are in order. In particular, a number of people have taken issue over the years with the dominant part played by logic in KR, so we will attempt to address some of these criticisms.

1.1 The Role of Logic in Knowledge Representation and Reasoning

There are two (extreme) ways one can attempt to build a program or agent that embodies knowledge of the world and acts with some degree of intelligence. At one end of the spectrum, one can write a program, using any means necessary, that exhibits the desired behavior. At the other, one can make explicit every piece of knowledge required by the program, and characterize the behavior formally in terms of this knowledge. Of course, a number of gradations lie between the two poles, but somewhat surprisingly, some AI researchers view the two approaches as mutually exclusive (of course, we have suggestively phrased the positions so they do not appear so). While the first view has been dubbed the *procedural* approach to AI, the second is known as the *declarative* or *logical*

[1] See the next section regarding this distinction.

approach, and has often been summed up by B.C. Smith's *Knowledge Representation Hypothesis*:

> Any mechanically embodied intelligent process will be comprised of structural ingredients that a) we as external observers naturally take to represent a propositional account of the knowledge that the overall process exhibits, and b) independent of such external semantical attribution, play a formal but causal and essential role in engendering the behavior that manifests that knowledge. (Smith 1982, p.82)

There are several reasons the logical view has emerged as predominant. If some program exhibits the desired (intelligent) behavior, then clearly the knowledge of the world and ability to reason with such required to exhibit this behavior is present within the program, though such knowledge might be only implicit and not readily apparent from the code. If we are to understand, reason about, and predict the behavior of this system, or combine this component with other (intelligent) programs, it makes sense to state explicitly the behavior of the program as a function of its knowledge and external environment.

But why should this knowledge be represented as sentences in some logic? The most compelling answer is that logics (usually) come equipped with formal semantic accounts that specify the meaning of sentences in the associated language and justify the notion of consequence determined by the logic. Because of this, knowledge represented as sentences in some logical language can be readily associated with facts in the domain of interest via the model theory. Hayes (1974), for instance, has argued that this is the primary reason for logical representations: we require no commitment by a system to represent knowledge explicitly in terms of logical sentences, and to reason with a (general-purpose) theorem prover, only that such a system be able to be understood in such terms. Notice that this view of the role of logic is much weaker than that espoused in the Knowledge Representation Hypothesis.[2]

Once the (explicit) Knowledge Representation Hypothesis has been abandoned, of course, one might claim that any formal system will do when it comes to characterizing precisely and in a principled manner the reasoning performed by a program or agent, and that logic should be accorded no special status in this regard. If a set of differential equations will accurately model the behavior of a program, why bother with logical accounts? While prediction of behavior might be accurate within any formal system, it is the model-theoretic semantics of logics that give logical representations their advantage in understanding behavior (Hayes 1974; Moore 1982). Now, one might argue that formal semantic analysis provides no real *meaning* to sentences, it is merely the mapping of one mathematical structure (the logical language) into another (an interpretation of the language). These so-called models may be any structure whatsoever. For instance, term models (or Herbrand models) of a first-order language cannot be said to provide meaning to sentences in any natural sense. Clearly, intuitive semantics is a matter of degree. Certain models might seem completely arbitrary and unnatural, though formally adequate. However, others might correspond more readily to our intuitions about the structure of the "real world" or relevant application domain. The notion of *natural semantics* is evidently subjective, but to the extent such semantic accounts exist for our formal characterizations, they can aid in the evaluation of the adequacy, propriety and utility of the corresponding system.[3] In any event, "formal semantics reduce an impossible problem, that of understanding the meaning of arbitrarily complex sentences of a recursive language, to a more tractable problem, that of understanding the meaning of an atomic set of primitive semantic features and functions" (Bacchus 1990, p.3).

---

[2]This idea is related to that developed in Newell's (1982) account of the *knowledge level*.

[3]See (Pollock 1984, Chapter 6) for a discussion of the philosophical importance *and* unimportance of formal semantics.

Given this view of the role of logic in KR, a number of other criticisms of logical representations remain unaddressed. Minsky (1974) has argued that translating "real world" knowledge into a formal language is a difficult task at best, and that determining the relevant facts in such a knowledge base for the job at hand is problematic. This point is at least partially valid: to the extent that the first-order predicate calculus has been the standard logic in knowledge representation, writing facts as logical sentences can be difficult. Expressing probabilistic, temporal, epistemic, and deontic notions, not to mention our subject, notions of typicality, in first-order logic (FOL) is a virtual impossibility, at least doing so in a natural or appealing manner. However, logics and other formalisms extending the expressive power of FOL with intuitive accounts of these notions have been proposed (for instance, Shoham (1988), Bacchus (1990), Levesque (1984a), and Horty (1991) present examples of such systems). This thesis is concerned with representing statements of normality, a problem addressed by a number of other logical systems (Reiter 1980; McCarthy 1980; McDermott and Doyle 1980; Moore 1985).[4] The epistemological adequacy and naturalness of expression afforded by such systems is a matter of debate, and we will present a logical representation that is indeed quite intuitively appealing. It is hard to see how a completely non-logical system can allow one to represent knowledge more naturally (and unambiguously) without disregarding all principles whatsoever.

Regarding the objection of selecting appropriate knowledge from a large set of facts for various reasoning tasks, suffice to say that if a program performs a reasoning task properly, it must "select" the relevant facts from the irrelevant facts, implying the designer of the program (or the program itself) is "aware" of such relevance. In this case, the specification of the program's behavior in terms of the knowledge it possesses and uses should not be problematic, in principle. At the very least, it should be a useful exercise to enhance the understanding of the program. The Knowledge Representation Hypothesis (at least on our weakened conception) does not require the program be a theorem prover running over some set of sentences; but, in principle, it could be (with appropriate heuristics determining some notion of relevance).

Minsky (1974) and McDermott (1987) have also suggested that logic is inappropriate for modeling human reasoning because of its "perfect" nature. For example, logical reasoning is both *sound* (all conclusions reached are valid or "true") and *complete* (all true facts can be deduced), while human reasoning possesses neither quality. Certainly these are not achievable properties in practical reasoning nor are they necessarily desirable. For instance, the problem addressed in this thesis regarding the ability to make assumptions or jump to conclusions illustrates the need for "unsoundness": if we know birds normally fly and that Tweety is a bird then the reasonable conclusion that Tweety flies is not guaranteed true, and the inference is not deductively valid. But this criticism neglects what we might call *inductive logic*, or logical accounts of plausible inference. The flying bird inference, while not (deductively) valid, is what Rankin (1988, following Fetzer) calls *inductively proper*. A suitable inductive logic would sanction precisely the inferences Minsky calls "unsound." Indeed, some properties of inductive inference are identified by Rankin that distinguish it from deductive inference. Inductive inference is:

(a) *nondemonstrative* — its conclusion could be false though its premises are true;

(b) *ampliative* — its conclusion contains more information than the premises;

(c) *nonadditive* — further premises could change the strength or conclusion of the argument.

[4] These and many other such systems will be surveyed in Chapter 2.

It is precisely the nondemonstrative nature of inductive inference that accounts for "unsoundness," but unsoundness only in a deductive sense. If by unsoundness one is referring to the mistakes and poor decisions for which human reasoners are notorious, then inductive inference is not unsound, for it sanctions only reasonable conclusions, and we should expect no less from our formal systems and (intelligent) programs. However, unsoundness in the deductive sense is not outside the province of inductive reasoning and is perfectly amenable to logical analysis.

The nonadditive nature of default inference is the essential cause of this nondemonstrativity, or unsoundness. If we come to believe that Tweety has a broken wing, the inference that Tweety flies is no longer forthcoming. Additional premises might change conclusions in a way that is impossible to represent deductively (see the *qualification problem* below). If the initial inference had been deductively sound, it would remain so when accompanied by new information. This leads to another criticism of logical formalisms: logic is *monotonic*. That is, if X and Y are sets of sentences (premises) and $X \subseteq Y$, then $Cn(X) \subseteq Cn(Y)$, where Cn is the logical consequence operator under discussion. New premises cannot invalidate old conclusions. Yet we have seen that inductive logics are not subject to this criticism as they have built into their structure a certain *nonmonotonicity*. A hallmark of current approaches to default reasoning is their nonmonotonic quality. Typically, such systems start with a classical (monotonic) propositional, first-order, or modal logic and add some extra-logical mechanism for deriving inductive conclusions. Because of this, while the (apparent) meaning of sentences of the logic is derived from the underlying classical semantics, any systematic account of the inference process is often unclear or uncompelling.

An alternative view, and one developed to a certain extent in this thesis, is that the logic characterizing reasoning need not be nonmonotonic (nor, in the strictest sense, inductive), even though it captures inductive patterns of reasoning. This view is implicit in the work of Moore (1985) and is made explicit by Levesque (1990). It asserts that while human reasoning is inductive in nature, we can account for the, say, nonadditive character of inference within a monotonic logic. For instance, consider again the flying bird inference. While learning new premises changes our conclusion, the change can be viewed monotonically: the conclusion that Tweety flies is derived when "all we know" about Tweety is that she is a bird, in particular, when we believe nothing about Tweety's wing being broken. When "broken wing" is added to our stock of beliefs, the conclusion is no longer (inductively) proper. But, on the view that "nothing about a broken wing is believed" is not a premise of the new inference, the consequence relation is no longer nonmonotonic. To model the nonadditive nature in our logical analysis, we take into account among our new premises that certain facts are no longer disbelieved; hence, our set of premises has not strictly increased. The logical model of reasoning developed in this thesis will adopt this stance: the nonmonotonic nature of inference can be captured logically (and monotonically).

Perhaps the hardest criticism to address is the completeness of logical reasoning. Obviously, human reasoning is not complete, and for computational reasons (see discussion below) such a goal is unattainable by any intelligent artifact. How can we propose logical accounts of reasoning (and the accompanying semantics) as characterizations of intelligent programs when we know full well that such specifications cannot come close to being met? Israel (1980) makes a related point, that logical rules of inference (the cause of completeness and computational difficulties) are not *real* rules of inference and AI should be more interested in such rules. He recommends turning to epistemological theories, which are (arguably) interested in just such rules. Unfortunately, theories of knowledge all seem to require notions of belief, truth, justification, and so on, and it appears most such theories (e.g., Levi (1980), Swain (1981), Pollock (1986)) rely on, to some degree, logical relations between sentences (or other "items of knowledge"), and, in fact, seem to be *converging toward* the same problems as AI, not solving them.

Certainly theories of knowledge can be either descriptive or normative; but each type influences the other, in particular, normative theories guiding the form of descriptive theories (MacLennan 1988). In this respect, logical accounts of KR can also be viewed as normative, or prescriptive, theories. They describe how an ideal, rational, computationally-unbounded agent should reason. It is to this lofty ideal that an intelligent artifact (or human being, even) should aspire, unattainable though it might be. However, as Glymour (1988, p.204) puts it: "Ought implies can, or so I think, and any normative epistemological theory, like any ethical theory, bears the burden of showing how its imperatives can be fulfilled, or if not fulfilled, how they can be better or worse approximated." It is on this score that most theories of KR can be criticized, but only to a certain extent.

Comparatively little effort has been put into developing effective and meaningful approximations to existing theories.[5] Unfortunately, there is, as yet, little consensus as to what constitutes an adequate normative theory of reasoning, especially one that encompasses default reasoning. Most would agree that no adequate theory yet exists, and implicit in Glymour's statement is a methodology for AI that requires an adequate (idealized) normative theory that can be approximated. While work should be progressing toward these notions of approximate and incomplete reasoning, research should also continue into the development of this prescriptive component.[6]

Theories of incomplete inference still require *some* notion of inference. For instance, Cherniak's (1986) theory of minimal rationality gives an account of why agents cannot be ideally rational and proposes certain weaker conditions that any successful agent should (attempt to) satisfy. But even minimal rationality is specified in terms of feasible and sound inference. A normative theory is a theory of sound inference on which such approximations may be based. Again, this is not to deny the importance of heuristics. "Formally incorrect heuristics need not in fact be irrational at all. They are not just unintelligible or inadvisable sloppiness, because they are a means of avoiding computational paralysis while still doing better than guessing" (Cherniak 1986, p.82). How do we know if a heuristic is formally incorrect? Perhaps we needn't be concerned with that question. How do we know when a heuristic is appropriate, when it "does better than guessing"? This is a question we must be able to answer; it is the case just when it sanctions sound inference more often than not.[7] Again, to judge the efficacy of heuristics an appeal to a normative theory is required.

As mentioned above, computational difficulties prevent us from realizing effective implementations of complete logical reasoning. Even classical propositional reasoning is intractable, yet it is generally agreed that at least the expressive power of FOL will be required of any reasonable KR service. Determining the consequences of a first-order theory is undecidable in general. One motivation for early investigation into default reasoning was to circumvent such difficulties in reasoning, by allowing one to ignore (almost) meaningless preconditions that would have to be proven true before being certain of a conclusion. An instance of what McCarthy (1977) calls the *qualification problem* is not having to prove, to *consider* starting a car, that the battery is connected, the tank is not empty, there is no potato in the tail pipe, and an endless list of such trivial, silly details.

---

[5]Some work using 4-valued semantics and weakened rules of inference for FOL (Patel-Schneider 1987; Lakemeyer 1987), and weakening the expressive power of default rules (Selman and Kautz 1988), for instance, has led to decidable or tractable fragments of KR schemes. This work is best viewed as first steps in this direction, and demonstrates the thesis of Levesque and Brachman (1985) that there exists a fundamental tradeoff in the expressive power of representations and the computational effort required to use them. However, these cannot be said to directly address the issue of "approximating completeness," since these logics give up entire classes of inference within the semantics, hence remaining complete with respect to the weakened model theory.

[6]Kelly (1988) expounds a somewhat different view of normative theories in which the effective component of a theory is of prime importance.

[7]We take the term "more often than not" to be sufficiently vague to allow a number of interpretations. Presumably, some decision theoretic criteria will be involved.

Even specifying such a list is a hopeless task. Standard default formalisms allow one to conclude that the car will start if *all one knows* is that the key has been turned, ignoring the possibility of these strange conditions; but, to discount these, default systems require at least they be shown not considered true, for if they are "known" the conclusion should be invalid. Thus it is required that the negation of these conditions (i.e. the qualifications) be consistent with existing knowledge, and one saves effort, it is claimed, by not having to *prove* these negations. Herein lies the problem, for the set of satisfiable formulae in FOL is not even recursively enumerable.

This fact has been the subject of much criticism (McDermott (1987), for example), since default reasoning is supposed to make reasoning easier, yet has made a semi-decidable problem completely undecidable. There are two ways to address this point. The first is to assert that, from a practical standpoint, both non-r.e. problems and r.e., non-recursive problems are unwieldy. If we are to keep the expressive power of FOL, approximation techniques for reasoning will be required that are decidable (or tractable), in which case (for this notion of "theoremhood") consistency checking will be decidable (or tractable, assuming consistency defined in terms of negation). The second answer is to notice that the gains afforded by default reasoning are not necessarily computational (in particular, with respect to the qualification problem). Having to prove the consistency of the qualification might not be an easier *reasoning* task than proving the qualification, but consistency-checking can certainly save much *investigative* effort if nothing is known about the condition; and it is precisely this (likely) physical effort that is certainly more costly than any reasoning that takes place.[8] Default reasoning liberates one from the task of exploring the environment to prove that unlikely conditions are false by permitting one to assume such.

1.2 Overview

In this thesis, we will propose conditional logics for representing default rules that have an intuitive semantic characterization and allow us to reason about and combine default rules in a manner that lies outside the scope of most existing default reasoning systems. Roughly speaking, a default rule will take the form $A \Rightarrow B$ and be read as "In the most normal situations in which A holds, B holds as well," or "A normally implies B." Hence, default statements can be interpreted as statements of normality or prototypicality, or as exception-allowing generalizations, such as "Birds normally fly." We will also develop a conditional account of belief revision in which we read $A \xrightarrow{KB} B$ as "If the initial state of belief KB were revised to incorporate A, then B would be believed." This account will be related to our logic of default rules, and we will demonstrate that revision and default reasoning are remarkably similar processes. Both accounts will be related to existing approaches and will extend them in meaningful ways.

Very little attention will be given to implementation issues. In this regard, the work here should be viewed as (a step toward) a normative theory of default reasoning and belief revision, against which methods of approximation and actual implementations can be measured.

Throughout, we assume a familiarity with classical propositional logic (CPL), and to some extent first-order logic. A brief survey of the motivations and techniques underlying modal and conditional logic will be given in Chapter 2, but only to the degree such matters are relevant to subsequent developments. An acquaintance with these topics is helpful, but not necessary.

[8] We draw attention to an analogy to file processing. The first lesson learned by any student in a basic file structures course is that the (computational) efficiency of a program, in terms of the number of CPU cycles for example, can be sacrificed to a great extent if it minimizes the number of disk accesses, which are orders of magnitude slower due to their physical component. Similarly, the lesson of default reasoning should be that it is desirable to expend quite a bit of "mental effort" in order to save the physical requirements of exploring the environment.

In Chapter 2 we survey a number of developments in default reasoning and its logical representations, as well as some applications of default reasoning. We then discuss briefly some details of modal and conditional logics. We show that conditional logics possess many properties that are desirable, in general, of default representations. This, in turn, motivates the perspective adopted in this thesis, that default rules can be captured naturally through the use of conditional logics (in particular, those defined in terms of modal logics).

In Chapter 3 we briefly survey the area of belief revision, concentrating on the work of Alchourrón, Gärdenfors and Makinson, and that related to it. We also discuss truth maintenance systems, the primary area of research about belief revision in AI.

The natural representation of defaults is addressed in Chapter 4, where we present a family of conditional logics for default reasoning that are equivalent to the modal logics extending S4. In fact, we adopt the view that these modal logics are themselves logics for default reasoning within which the conditional connective can be defined. We examine properties of these logics, in particular discussing the role background and evidence play in our representations, a distinction made in probabilistic reasoning systems but seldom in symbolic, qualitative systems. We then show how our modal framework subsumes several standard conditional approaches to default reasoning, namely preferential and rational entailment, and ε-semantics.

Some difficulties exist in standard conditional approaches to defaults, however, the key problem being that of *irrelevance*. In Chapter 5, we discuss this problem at some length and detail current solutions to the problem, all of which rely on extra-logical notions. We extend our modal logics with an additional unary modal connective, one corresponding to truth at *inaccessible* worlds. With the increased expressive power, the new logics CO and CO* are able to capture the assumptions made to deal with irrelevance purely axiomatically. This is demonstrated by proving the equivalence of a class of CO*-theories to the extra-logical systems determining rational closure and 1-entailment. We continue by comparing two notions of relevance, *statistical relevance*, based on intuitions of conditional independence, and *practical relevance*, a more coarse-grained notion to which our conditional notion of relevance corresponds.

In Chapter 6 we turn our attention to the problem of belief revision. We define a conditional connective for revision based on the extended bimodal logics presented in Chapter 5, showing it to be equivalent to the standard AGM model of revision. Unlike the AGM model, the approach based on CO* is again defined purely within the logic, with the "postulates" for revision taking the form of derived rules and theorems. Our approach allows the obvious generalization of AGM revision as revision can be defined for any of a number of modal logics. Furthermore, our approach makes no commitment to the *Limit Assumption* made by most current approaches to revision. Within this model, we can define naturally (using the modal language) concepts of entrenchment, plausibility, and various forms of integrity constraints, along with an account of subjunctive queries that improves on existing approaches in several ways.

Key to our model of revision are several epistemic notions, in particular the concept of *only knowing*. We discuss how our logics of revision are related to standard autoepistemic reasoning systems, and show that we can view CO* as a generalization of the autoepistemic logic OL.

In Chapter 7 we discuss the relationship of the normative conditional defined in Chapters 4 and 5 to the subjunctive conditional presented in Chapter 6. We conclude that the subjunctive and normative conditionals are, in fact, the same conditional, and that default reasoning and belief revision differ only in the way the logics are used, or the perspective that one adopts when applying the logic. Given this relationship, we demonstrate some and suggest many other possible connections between a number of diverse reasoning systems for defaults and revision. Among these are conditional approaches to default reasoning, including the probabilistically motivated system

of ε-semantics, autoepistemic logic (and hence logic programs, default logic, etc.), AGM belief revision and some generalizations of it, abduction, possibility theory, subjunctive (or hypothetical, or counterfactual) reasoning, and standard probabilistic reasoning.

In Chapter 8, we summarize the contributions of this thesis and examine interesting avenues of further research, reviewing some mentioned in earlier chapters. In discussing potential extensions to this, we will make explicit some of the assumptions underlying our approach and justify the possible worlds framework we adopt.

Chapter 2

Logical Foundations for Default Reasoning

In this chapter we will survey a number of developments in default reasoning. We distinguish default reasoning from the more general class of nonmonotonic reasoning. Roughly, a default reasoning system is one that can represent and reason with arbitrary facts of the form "If A then normally B," or the like. This excludes from consideration a number of nonmonotonic systems. For example, certain temporal reasoning schemes (Shoham 1988; Kautz 1986), though nonmonotonic, can make default assumptions only about the persistence of events. Similarly, systems based on the *closed-world assumption* (Reiter 1978a) can make assumptions only regarding negative information. Hence, these will be examined only cursorily.

We will follow by providing a very brief discussion of modal and conditional logics, introducing some of the notation and terminology we use in later chapters. We also point out some similarities of theorems of conditional logics to principles of default reasoning. In this way we motivate the viewpoint adopted in this thesis, that conditional logics may be used profitably to model nonmonotonic aspects of reasoning.

Probabilistic approaches to reasoning have played a crucial role in the development of nonmonotonic theories. While we do not survey the development of such systems in this chapter, several recent approaches to default reasoning based on probabilistic principles will be examined in subsequent chapters, and compared to the models of reasoning developed here.

2.1 Logical Preliminaries

We assume a familiarity with classical propositional logic (CPL) and some formulation of first-order logic (FOL), as well as an acquaintance with modal logics. We take $\mathbf{P}$ to denote a denumerable set of atomic variables and $\mathbf{L}_{CPL}$ to denote the propositional language over this set. Letters A, B, C, etc. sometimes denote propositional variables or atoms, but often are used for arbitrary formulae. Lower-case letters (p, q, r, e, etc.) are used similarly. Upper-case letters near the end of the alphabet (S, T, X, Y, etc.) typically denote sets of formulae. The primary exception is the use of a letter (say P) to abbreviate an atom with an intended reading (say penguin). Typeface words such as bird always indicate relation symbols (and usually atoms), and when capitalized refer to constants (e.g., DCS). Greek letters α, β, γ, etc., are variables ranging over arbitrary formulae. We will often lapse and present derived theorems or rules of inference in a form without α, β, etc., but with atoms A, B, and so on. Due to principles of substitution in our logics, these should be viewed

as, for instance, theorem schemata. The symbol $\top$ denotes the identically-true proposition (e.g., any truth-functional tautology) and $\bot$ denotes falsity (e.g., any contradictory sentence).

We use $\vdash$ to denote derivability generally, and in CPL in particular. When the symbol is subscripted (as in $\vdash_{CT4}$) we intend derivability in the system thus indicated. We write $\alpha, \beta \vdash \gamma$ for $\{\alpha, \beta\} \vdash \gamma$, and $\vdash \alpha$ for $\emptyset \vdash \alpha$ (thus, $\vdash \alpha$ indicates the theoremhood of α). We say α is *contingent* if $\nvdash \alpha$ and $\nvdash \neg\alpha$; and α is *contingent with respect to* β (or *given* β; or *on* β) just when $\beta \nvdash \alpha$ and $\beta \nvdash \neg\alpha$. We use Cn to denote logical consequence operations in the Tarskian sense (though generalized in the sense of Gabbay (1985) for nonmonotonic systems), and typically use it for consequence in CPL; so $Cn(X) = \{\alpha : X \vdash \alpha\}$. If S is a set of sentences and Cn denotes consequence for some logic L, then we say S is *L-consistent* iff $Cn(S) \neq Cn(\bot)$ (this is sufficient for our purposes). We say α is L-consistent to mean $\{\alpha\}$ is (which will be true in all systems of interest here just when $\nvdash_L \neg\alpha$.)

All logics will be defined on propositional languages. However, some examples will be clearer if first-order notation is used. If we write, say, $\mathtt{p(x)}$, we take this to mean that there exists some (finite) set of propositions $\mathtt{p1}, \mathtt{p2}, \ldots \mathtt{p}n$ that assert that predicate $\mathtt{p}$ is true of domain objects 1 through n. So $\forall \mathtt{xp(x)}$ stands for $\mathtt{p1} \wedge \mathtt{p2} \wedge \ldots \mathtt{p}n$, while $\exists \mathtt{xp(x)}$ means $\mathtt{p1} \vee \mathtt{p2} \vee \ldots \mathtt{p}n$.

We introduce some scoping conventions for logical connectives that will simplify parenthesized formulae. Unary connectives such $\neg$, $\Box$, and other modal operators will have the narrowest scope (or will bind most tightly). Among binary connectives, conditional connectives ($\supset$, $\nRightarrow$, $\Rightarrow$, $>$, etc.) will have the wider scope (will bind less tightly), than other binary connectives ($\wedge, \vee$). Thus

$$\neg A \wedge \Box B \supset C \vee D$$

will be interpreted as

$$((\neg A) \wedge (\Box B)) \supset (C \vee D).$$

Naturally, occurrences of parentheses override these conventions.

We refer to propositional valuations loosely as possible worlds or states of affairs. We write $w \models \alpha$ if w is a world that assigns 1 (or truth) to α, and $\models \alpha$ if α is valid. When $w \models \alpha$ we call w an α-world. We also use $\models$ for satisfaction more generally (see Section 2.3) and for semantic entailment, subscripting $\models$ as we do $\vdash$.

If X is some set of elements, a *partial order* $\leq$ is a binary relation on X that is reflexive, transitive and antisymmetric. That is, for all $x, y, z \in X$:

(a) $x \leq x$,

(b) $x \leq y$, $y \leq z$ implies $x \leq z$,

(c) $x \leq y$, $y \leq x$ implies $x = y$.

A *partially ordered set* (or poset) is a pair $\langle X, \leq \rangle$ where $\leq$ is a partial order on X. A relation $<$ is a *strict partial order* if it is transitive and antisymmetric as well as *irreflexive*:

(d) $x \not< x$.

A relation $\leq$ is a *preorder* if it is reflexive and transitive, and it is a *total preorder* if it is, in addition, *totally ordered*:

(e) $x \leq y$ or $y \leq x$.

A *total order* is a totally ordered partial order.

2.2 Default Reasoning

Because of the ubiquity of default reasoning, any general theory of AI must include some component that specifies the manner in which default conclusions can be made and justified. In this section we will examine a number of formal, logic-based systems for default reasoning.

It should be noted that material implication is not adequate as a representation of default rules. There are a number of inference patterns that hold for classical implication that are undesirable for any sort of default rule or prototypical statement. Consider the following rules.

Strengthening From $A \supset B$ infer $A \wedge C \supset B$.

Transitivity From $A \supset B$ and $B \supset C$ infer $A \supset C$.

Each of these is valid for material implication, but should not be for our default rules. Consider the default "birds fly." Taking A, B and C to stand for the propositions "bird", "flies" and "penguin", respectively, we see that if strengthening were valid, we should be able to conclude that penguins (which are birds) fly as well. But this runs contrary to the type of nonmonotonicity or nonadditivity that is a hallmark of default reasoning. Similarly, if transitivity were to hold, we would be forced to conclude that penguins fly because penguins are birds and birds fly. Thus, many people have attempted to develop systems that treat default rules in a systematic, principled, and logical way, yet do not fall prey to these "fallacies".

As well as providing a general account of default reasoning in commonsense domains, a number of these systems can be applied to more specific expert or practical areas of AI research that have a default character (Reiter 1987a). For example, semantic accounts of logic programs, deductive databases and inheritance networks often appeal to default reasoning systems. Diagnostic theories can adopt these formalisms, and the frame problem (see (Shoham 1988) for a survey of this problem) provides a useful testing ground as well. Default reasoning systems have also found use in computational linguistics (Mercer 1988). Another recent area of research in which default reasoning might be of integral importance is that of *vivid knowledge* (Levesque 1986a). Often the computational cost of deductive inference is prohibitive, but "complete" knowledge might reduce the operation to that of database lookup. Such completeness may be gained by making default conclusions, filling in gaps in the knowledge base. While these conclusions are defeasible (and hence, might be incorrect), they will speed up inference considerably (see also Selman (1990)); and if the validity of these facts is irrelevant to the queries we intend to ask, this unsoundness might not be a problem. Vivid knowledge has also been addressed to some extent in psychological circles (Johnson-Laird 1983).

A number of informal ideas involving default reasoning have been forthcoming in the areas of AI, psychology and philosophy. The notion of *frames* was proposed by Minsky (1974) to represent information about objects or situations. A frame corresponding to a situation is activated when an instance of that situation is encountered. These frames have *roles* that are often filled with default values. These values are "attached loosely ...so that they can easily be displaced by new items that fit better the current situation" (Minsky 1974, p.247). This seems to fit the reasoning pattern "Assume A in the absence of contradictory evidence."

The concept of *prototypes* was developed in psychology as a method of representing categories (Rosch 1978). Since categories are continuous, necessary and sufficient conditions for category membership, the demarcation of category boundaries, might not exist. Prototypes are the "best examples" (those that best reflect the redundancy structure) of a category, but whose prototypical properties are, at best, default conclusions to reached about arbitrary members. A similar notion of *stereotype* occurs in philosophy in an account for the meaning of natural kind terms, which

also resist definition by classical means (Putnam 1970). Formal default reasoning schemes might provide significant insight regarding the character of these three notions (Reiter 1987a), and these also suggest properties that ought to be accounted for in any reasoning scheme.

2.2.1 Defeasible Reasoning

Default reasoning has not been addressed exclusively by AI researchers. It has generally been acknowledged in contemporary epistemology that reasons for a belief need not logically entail their conclusion (Pollock 1986). Such reasons are *defeasible*, and since much of a person's knowledge is based on defeasible reasons, an account of this phenomenon is critical to any theory of knowledge.

For some time, the definition of knowledge as justified true belief was widely accepted. Recently, this definition has been rejected as too weak;[1] however, any subsequent characterization will still require an account of justified belief. It is this that may be construed as central to the study of default reasoning. The other components of knowledge will be of less relevance, for a default reasoner will in general have no way of determining if its beliefs are actually true, or if there is some *unknown* fact that renders the belief unjustified. In designing a system (or even in our own reasoning), the best we can expect is that the system's beliefs be justified.

Pollock (1986; 1987) argues that beliefs are based on *foundational states*, original input states requiring no justification. These justify all subsequent beliefs. P is a *reason* for agent A to believe Q (written $P \rightarrow Q$) iff it is logically possible for A to become justified in believing Q on the basis of P. Here Q is some proposition and P is a set of propositions representing other beliefs or foundational states. This reason P is defeasible or *prima facie* if there is an R such that R is logically consistent with P and $P \wedge R$ is not a reason for Q; and in this case, R is a *defeater* for reason P. Nondefeasible reasons are called *conclusive*. For instance, "Tweety is a bird" is a conclusive reason for "Tweety is a vertebrate," but a prima facie reason for "Tweety flies." Pollock also distinguishes two types of defeaters for $P \rightarrow Q$, namely *rebutting defeaters*, where defeater R is a reason for denying Q (i.e. a reason for $\neg Q$), and *undercutting defeaters*, where R is a reason to deny that P wouldn't be true unless Q were (i.e. a reason for $\neg(P \rightarrow Q)$). In the previous example, "Tweety is a penguin" is a rebutting defeater for the flying conclusion, while the conclusion that an object is red, based on the fact that it appears red, is undercut by the fact that the object is under a red light. Types of reasons include deductive reasons (which are conclusive), perceptual reasons, recollection (of beliefs), and inductive reasons, all of which are prima facie.[2]

Pollock (1987) provides a detailed account of the structure of defeasible reasoning. Such reasoning proceeds from an *epistemic basis*, consisting of a collection of foundational states and reasons. A *warranted* belief is one that is supported by an *ultimately undefeated argument*. Though we provide no details, this notion is reminiscent of consistency-based approaches to default reasoning (see, e.g., default logic in the next section) and has a built-in account of specificity (cf. (Poole 1985)), a concept we discuss later.

Pollock's account is rather syntactic and reasons are extremely rule-like. Reasons are used only to derive new beliefs from old ones, and cannot themselves be reasoned about. For instance, if $P \rightarrow Q$ and $P \rightarrow \neg Q$ are reasons, one cannot derive $\neg P$. Reasons are meant to represent epistemic

[1] The Gettier problem (Gettier 1963) marked a turning point in epistemology. In his paper, Gettier gives two simple examples where this definition of knowledge is insufficient. Since that time, much effort has been applied to augmenting this definition with new conditions to provide an intuitive account of knowledge.

[2] Others have taken a somewhat different approach to reasons. For instance, while Pollock's view requires that reasons and justification of a belief are intimately related, Swain (1981) holds the view that reasons are states that *cause* a person to believe something, regardless of the justification for that belief.

norms that, Pollock argues, are more like production rules than explicitly stored conditionals. Without making any claims about how reasoning is performed, we remark that such a theory of warrant is hard to evaluate without some semantic account and analysis of the "conditional" corresponding to these rules using $\rightarrow$.

2.2.2 Default Logic

Reiter (1980) introduces *default logic* as a method of default reasoning. Given a set of first-order sentences representing knowledge of the world or domain, the idea is that there exist various plausible (but fallible) ways in which this knowledge can be extended. These possible extensions are represented by *default rules* of the form $\alpha(\overline{x}) : \beta(\overline{x})/\gamma(\overline{x})$, where $\alpha(\overline{x})$, $\beta(\overline{x})$ and $\gamma(\overline{x})$ are first-order formulae.[3] A rule of this form is intended to express something like "If α is known and β is consistent with what is known, then assume γ is true." For example, the assumption that Tweety flies could be represented as

$$\frac{\texttt{bird(Tweety)} : \texttt{fly(Tweety)}}{\texttt{fly(Tweety)}}.$$

A default theory is an ordered pair $\langle D, W\rangle$ where W is a set of first-order sentences and D is a set of closed default rules (in some fixed language $\mathcal{L}$).

In order to capture the fashion in which default rules induce completion of W, Reiter defines the notion of an *extension* of a default theory by appealing to the operator Γ_T (for any default theory $T = \langle D, W\rangle$). The idea is to apply as many default rules as possible to theory W. For any $S \subseteq \mathcal{L}$, we define $\Gamma_T(S)$ to be the smallest set such that

1. $W \subseteq \Gamma_T(S)$,
2. $\Gamma_T(S)$ is deductively closed, and
3. If $\alpha : \beta/\gamma \in D$, $\alpha \in \Gamma_T(S)$, and $\neg\beta \notin S$, then $\gamma \in \Gamma_T(S)$.

An extension of T is any $E \subseteq \mathcal{L}$ such that $\Gamma_T(E) = E$. In other words, E is an extension if for any default rule whose prerequisite is in E and whose justification is consistent with E, the consequent of that rule is also in E.

Clearly, this characterization is nonmonotonic, for if D contains only the default rule about Tweety (above), and $W = \{\texttt{bird(Tweety)}\}$, then the only extension of T contains `fly(Tweety)`. However, if W is augmented with $\neg$`fly(Tweety)`, this extension is disallowed. Extensions can be thought of as plausible ways (according to the default rules) in which an agent can extend its initial set of beliefs W.

Default theories can have more than one extension. For instance, the theory

$$\left\langle \left\{ \frac{\top : c}{\neg d}, \frac{\top : d}{\neg c} \right\} \{\emptyset\} \right\rangle$$

has two extensions, one containing $\neg c$, the other $\neg d$. Reiter's original idea was that an agent should pick one extension as a "working" model of the world and reason from that extension,[4]

[3] In the sequel, we will assume that α, β, γ are sentences, called the *prerequisite*, *justification* and *consequent*, respectively. The results below hold only for *closed* default theories. Reiter (1980) deals with open default rules through Skolemization and uses all ground instances of the open rules. Also, a more general form of default rule allows the use of multiple justifications (with the obvious interpretation).

[4] This is related to the use of vivid knowledge.

although consequences of a default theory could be viewed as those sentences in the intersection of all extensions. We will adopt the latter view, assuming if there were some criteria by which an appropriate extension could be chosen (in preference to others), this could be represented in the theory itself, thereby eliminating the unwanted extensions.

Unfortunately, some theories have no extensions, for example $\langle\{\top : p/\neg p\}, \emptyset\rangle$. However, Reiter identifies the class of *normal default theories* as those whose default rules have the form $\alpha : \beta/\beta$. These types of theories have a number of desirable properties, among them the existence of extensions for any such theory, and *semi-monotonicity* (whereby adding new normal default rules will ensure new extensions subsume old ones). This semi-monotonicity also leads to a proof theory[5] for normal theories that is constructive in the sense that it does not rely on fixed points. Since normal default rules have an identical justification and consequent, by "applying" a default rule, we guarantee that its justification remains consistent, regardless of which rules are applied subsequently. This is not the case for nonnormal defaults. Of course, even normal defaults can be applied only when the justification is consistent with what is known; hence, this reliance on consistency prevents extensions of a default theory from being recursively enumerable.

The class of normal defaults is very encompassing. In fact, Reiter (1980) claimed that all naturally occurring defaults might be normal. However, due to interaction of rules, anomalous extensions can arise (Reiter and Criscuolo 1981), indicating that normal defaults are not adequate for all default reasoning. For instance, consider a representation of the facts that adults are typically employed, while students are not, and that most students are adults. The intuitive representation uses three normal defaults, namely, $A : E/E$, $S : \neg E/\neg E$, and $S : A/A$. However, if we know that S (student) is true, we get an undesirable extension where E (employed) holds (along with the proper extension containing $\neg E$). This can be resolved by the use of nonnormal defaults, for example, by replacing the first rule with $A : \neg S \wedge E/E$. Of course, using semi-normal defaults destroys Reiter's "constructive" characterization; but Reiter and Criscuolo also suggest that if adults are normally not students, normal defaults can be used again.[6]

Lukaszewicz (1988) suggests a reconstruction of default logic that has the attractive property that all default theories have extensions. In essence, the theory of Lukaszewicz requires one to keep track of the justifications used by applied defaults. This is not required in default logic, so if P and $\neg P$ are justifications for distinct rules, both rules may be activated when these are consistent with the theory. This is forbidden by the extended version of default logic.

There are several difficulties with default logic as an account of default reasoning. Because defaults are expressed as rules outside the logical language, they cannot interact with one another, nor with (first-order) facts. Thus new rules and facts cannot be derived in circumstances where they should be. For instance, if $p : \top/q$ and $p : \top/\neg q$ are default rules, it might be expected that $\neg p$ is derivable, which it is not. This problem is also reflected in the student example, where the intended interaction of rules necessitates the use of seminormal defaults (to "mimic" such interaction).

Finally, default logic is defined purely proof-theoretically. Several semantic accounts have been proposed since Reiter's initial presentation, among them the work of Etherington (1987b) and Lukaszewicz (1988); but neither account is particularly compelling, and while describing the inference process of default logic, neither provides an interpretation of the default rules themselves.

---

[5] An algorithm to decide whether β is in some extension of T.

[6] This could be solved by using $S : A/A$, $A : \neg S/\neg S$, $S : \neg E/\neg E$, and $A \wedge \neg S : E/E$. In fact, this might be a more appropriate course (see Section 2.3.3).

2.2.3 Circumscription

Given a theory about some domain, we might want to infer, unless contrary evidence exists, that objects in our domain do not possess a certain property, for example, the ability to fly. Adding an axiom $\forall x \neg \texttt{fly(x)}$ might be too strong however, because we may also have $\texttt{bird(Tweety)}$ and $\forall x \texttt{bird(x)} \supset \texttt{fly(x)}$ in our theory as well. McCarthy (1980) observes that many forms of default reasoning can be viewed as minimizing the extensions of various predicates. For instance, in the above example, if we only consider models of our theory in which the extension of $\texttt{fly}$ is minimal (with respect to set inclusion), this has the effect of asserting $\neg\texttt{fly(x)}$ only for those individuals that can't be shown to fly.

In a slight generalization of McCarthy's original formulation, we will first consider *parallel circumscription* (McCarthy 1986; Lifschitz 1985b). Given a tuple of predicate symbols $\mathbf{P}$, and a first-order sentence $A(\mathbf{P})$, we want to consider only those models of $A(\mathbf{P})$ in which extensions of predicates in $\mathbf{P}$ are minimal. Models M_1 and M_2 will be comparable only if M_1 and M_2 have the same domains, and interpret all nonlogical symbols other than those in $\mathbf{P}$ similarly. We write $M_1 \leq^{\mathbf{P}} M_2$ iff for each $P_i \in \mathbf{P}$, the extension of P_i in M_1 is no bigger than that in M_2. The (nonmonotonic) conclusions we draw on the basis of *circumscribing* $\mathbf{P}$ in A are those sentences true in all $\leq^{\mathbf{P}}$-minimal models of A.

McCarthy (1980) describes the *circumscription* of $\mathbf{P}$ in A, denoted $Circ(A;\mathbf{P})$, to be an (infinite) set of first-order sentences of the form

$$(A(\Phi) \wedge \forall \overline{x}[\Phi\overline{x} \supset \mathbf{P}\overline{x}]) \supset \forall \overline{x}[\mathbf{P}\overline{x} \supset \Phi\overline{x}],$$

for each tuple of predicate symbols Φ similar to $\mathbf{P}$, where $\Phi\overline{x} \supset \mathbf{P}\overline{x}$ stands for the conjunction of $\Phi_i\overline{x} \supset P_i\overline{x}$. This schema is intended to express the fact that if predicates Φ satisfy A and are "smaller" than predicates $\mathbf{P}$ (the antecedent), then $\mathbf{P}$ and Φ have identical extensions (the consequent). By having Φ range over all tuples of predicate symbols,[7] the desired effect is to make the extension of $\mathbf{P}$ as small as possible while still satisfying A, since any model of $Circ(A;\mathbf{P})$ must satisfy each sentence of this schema.

McCarthy shows circumscription is sound; that is, consequences of $Circ(A;\mathbf{P})$ are true in all $\leq^{\mathbf{P}}$-minimal models of A. However, Davis (1980) shows that circumscription is not complete, so that some sentences are true in all minimal models of certain theories A, yet are not derivable from $Circ(A;\mathbf{P})$.

A second-order version of circumscription is provided by McCarthy (1986) (see also (Lifschitz 1985b)). $Circ(A;\mathbf{P})$ now refers to the single second-order sentence

$$A(\mathbf{P}) \wedge \neg\exists \mathbf{p}(A(\mathbf{p}) \wedge \forall \overline{x}[\mathbf{p}\overline{x} \supset \mathbf{P}\overline{x}] \wedge \neg\forall \overline{x}[\mathbf{P}\overline{x} \supset \mathbf{p}\overline{x}])$$

where $\mathbf{p}$ is now a tuple of predicate *variables* similar to $\mathbf{P}$. Fortunately, second-order parallel circumscription is both sound and complete with respect to the minimal model semantics, as shown by Lifschitz (1985b) and Etherington (1988). In the sequel, circumscription will refer to the second-order formulation, unless explicitly stated otherwise.

While parallel circumscription is a powerful formalism, it does have certain drawbacks. A crucial weakness is pointed out by Etherington, Mercer and Reiter (1985), namely that one cannot derive various kinds of information using parallel circumscription. In particular, for any symbol $P \in \mathbf{P}$, no new positive information can be derived (i.e., $Circ(A;\mathbf{P}) \vdash P\overline{a}$ iff $A \vdash P\overline{a}$), and for any

[7] In fact, all possible predicates must be represented, so we must allow Φ to range over, say, all λ-expressions of similar type.

predicate not in $\mathbf{P}$, no new information, positive or negative, can be derived. For instance, taking an example from Etherington, Mercer and Reiter (1985), consider the theory with axioms:

$\forall x(\texttt{bird(x)} \wedge \neg\texttt{ab(x)} \supset \texttt{fly(x)})$
$\forall x(\texttt{penguin(x)} \supset \texttt{ab(x)})$
$\forall x(\texttt{ostrich(x)} \supset \texttt{ab(x)})$
$\texttt{bird(Tweety)}.$

By circumscribing ab (standing for "abnormal"), we might expect to derive `fly(Tweety)`. However, since the interpretation of predicates other than ab must remain fixed, certain models where, say, `ostrich(Tweety)` holds will be minimal; so we can't conclude Tweety flies.

To remedy this problem, *variable circumscription* has been proposed (McCarthy 1986; Lifschitz 1985b). This allows the minimization of some predicates while allowing certain others to vary in their interpretation. The variable circumscription of the sentence $A(\mathbf{P}, \mathbf{Z})$ (denoted $Circ(A; \mathbf{P}; \mathbf{Z})$) is the second-order sentence

$$A(\mathbf{P}, \mathbf{Z}) \wedge \neg\exists \mathbf{p}, \mathbf{z}(A(\mathbf{p}, \mathbf{z}) \wedge \forall \bar{x}[\mathbf{p}\bar{x} \supset \mathbf{P}\bar{x}] \wedge \neg\forall \bar{x}[\mathbf{P}\bar{x} \supset \mathbf{p}\bar{x}])$$

$\mathbf{Z}$ is now a tuple of predicate symbols allowed to vary in interpretation, while $\mathbf{z}$ is a tuple of predicate variables of similar type. Lifschitz (1985b) shows this form of circumscription is sound and complete with respect to the minimal model characterization given by the relation $\leq^{\mathbf{P};\mathbf{Z}}$, which is identical to $\leq^{\mathbf{P}}$ except that models need not agree on the interpretation of predicate symbols from $\mathbf{Z}$ in order to be comparable. In the above example, by circumscribing ab with `Fly`, `Penguin` and `Ostrich` varying, we obtain the desired result `fly(Tweety)`.

A noticeable aspect of circumscription is the difficulty we have expressing simple commonsense knowledge in a manner amenable to circumscription, and choosing which predicates to minimize and vary. McCarthy (1986) addresses this problem by suggesting a uniform approach to representing such knowledge. He advocates the use of the abnormality predicate, that is minimized (while most or all other symbols are allowed to vary), in effect asserting that most individuals are not abnormal. Default assumptions are true of those things that are not abnormal. Since individuals may be abnormal in a variety of ways, a number of *aspects* of abnormality are introduced. For instance, if a bird is abnormal with respect to, say, `aspect2` (e.g., it can't fly), then we don't want to conclude it is abnormal in other aspects as well (e.g., it has three legs). Hence, we write of a bird that can't fly `ab(aspect2, tweety)`. The use of abnormality is illustrated by the following theory:

$$\forall x(\neg\texttt{ab(aspect1,x)} \supset \neg\texttt{fly(x)}) \tag{2.1}$$
$$\forall x(\texttt{bird(x)} \supset \texttt{ab(aspect1,x)}) \tag{2.2}$$
$$\forall x(\texttt{bird(x)} \wedge \neg\texttt{ab(aspect2,x)} \supset \texttt{fly(x)}) \tag{2.3}$$
$$\forall x(\texttt{penguin(x)} \supset \texttt{ab(aspect2,x)}) \tag{2.4}$$
$$\forall x(\texttt{ostrich(x)} \supset \texttt{ab(aspect2,x)}) \tag{2.5}$$

By minimizing ab most things will be inferred not to fly; but birds will be abnormal in `aspect1`, disabling that conclusion. Similarly, birds will be assumed to fly, other than those abnormal in `aspect2`, such as ostriches and penguins.

Another version of circumscription, called *prioritized circumscription*, is suggested by McCarthy (1986) to reduce the complexity of abnormality theories. By allowing certain predicates to be minimized in preference to others (at a higher *priority*), axioms such as (2.2) above can be eliminated.

Given the fact `bird(Tweety)` together with axioms (2.1) and (2.3), minimizing `ab` results in the weak conclusion `ab(aspect1,tweety)` $\vee$ `ab(aspect2,tweety)`. By giving priority to the minimization of `ab(aspect2,x)`, the more specific information, we can conclude that Tweety flies without the need for axiom (2.2). Lifschitz (1985b) shows prioritized circumscription to be equivalent to a conjunction of variable circumscriptions. Lifschitz (1986b) also introduces *pointwise circumscription* in which predicates can be minimized at certain points in their extensions, and ignored at others.

Because the second-order circumscription axiom is complete with respect to predicate-minimal models and various theories do not possess such models (Davis 1980), the circumscription of certain consistent theories can be inconsistent (even using the first-order schema). Some work has been done on identifying satisfiable classes of theories (Etherington, Mercer and Reiter 1985; Lifschitz 1986a). The theorems of a circumscriptive theory are not generally recursively enumerable due to its second-order nature.[8] Perlis and Minker (1986) identify several useful classes of theories that are r.e. and Lifschitz (1985b) shows a class of theories whose second-order circumscription axiom is equivalent to a first-order formula; Kolaitis and Papadimitriou (1988) demonstrate that determining if the circumscription of a formula is first-order definable is an undecidable problem.

Minimal Models

Shoham (1988) has suggested that nonmonotonic inference can be viewed as selecting from models of a theory those that are in some sense *preferred*. Those facts true in all preferred models are the nonmonotonic conclusions to be derived from the theory. Shoham gives this notion of preference a precise technical interpretation. For any standard (monotonic) logic L, a *preference relation* is a preorder $\leq$ on the interpretations suitable for L. $M_1 \leq M_2$ means M_1 is as preferable as M_2, and $M_1 < M_2$ means M_1 is preferred over M_2.[9] The *preferred models* of theory T are those models that are minimal in the preorder (restricted to models of T), and the defeasible consequences of T are those sentences true in all preferred (or minimal) models of T.

It can be seen that this formulation is indeed nonmonotonic, for the preferred models of some superset of T need not be included in the set of preferred models of T. This framework is extremely general, Shoham's idea being that the details of most nonmonotonic reasoning systems can be captured by using specific instances of his *nonmonotonic logics* (a logic together with some preference relation). It has a certain intuitive appeal, and perhaps a certain psychological plausibility; it seems quite natural to restrict attention to likely or plausible models of the world rather than to consider all *possible* models of some theory. Besides accounting for novel forms of default reasoning, minimal models can also provide a semantic basis for existing techniques.

Of course, the most studied minimal model approach has been circumscription. A related approach is that of Bossu and Siegel (1985). Motivated by the closed-world assumption for databases, the idea is to minimize positive information. Semantic characterizations of default logic also appeal to minimal models. Etherington's (1987b) and Łukaszewicz's (1988) semantics for default logic can be viewed as preferring models of a default theory that violate as few default rules as possible. Conditional logic approaches to default reasoning (Section 2.3.3) such as those of Kraus, Lehmann and Magidor (1990), Pearl (1990), and the approach espoused in this thesis, can be viewed as enforcing a form of model preference, but they do not require an extra-logical notion of preference. Rather such assumptions are embedded within the logic through some type

[8] In fact, Davis (1980) uses this fact to show the incompleteness of the (r.e.) first-order version of circumscription.

[9] This is, in fact, a slight generalization of Shoham's approach, where $\leq$ is required to be a partial order. Using a preorder allows distinct models to be equally preferable.

of possible worlds semantics.

Besides default reasoning, minimal model semantics can account for other forms of nonmonotonic reasoning. In particular, various forms of temporal reasoning have been successfully defined using this technique. Indeed, Shoham's original motivation for developing his thesis of preferred models was the need for a theory of temporal reasoning. His *chronologically maximally ignorant* models attempt to deal with certain problems in temporal domains.

2.2.4 Autoepistemic Logic

Often default statements can be thought of as expressing rules or sentences involving the consistency of certain beliefs with an agent's current knowledge, for example, "If it's consistent with your beliefs that Tweety flies, then assume it's true." McDermott and Doyle (1980) attempt to capture this notion directly by augmenting the predicate calculus with a consistency operator. While their *nonmonotonic logic* has been found to be somewhat problematic, it gave rise to a number of more promising approaches. *Autoepistemic logic*, due to Moore (1985), attempts to alleviate some troublesome aspects of nonmonotonic logic by appealing to the notion of belief. In a related vein, nonmonotonic reasoning has been analyzed through the use of more traditional logics of knowledge and belief (e.g., by Levesque (1981; 1984a)).

Nonmonotonic Logic

McDermott and Doyle (1980) introduce nonmonotonic logic in an attempt to capture the consistency interpretation of default statements. The predicate calculus is extended with a modal operator M, the desired interpretation of $M\alpha$ being "α is consistent." The flying bird default, for instance, can be represented as $\texttt{bird} \wedge M\texttt{fly} \supset \texttt{fly}$. The difficulty with the consistency interpretation of M is deciding when a formula $M\alpha$ should be derivable and what consequences it should have. In general, a theory should entail $M\alpha$ exactly when it does not entail $\neg\alpha$. This definition is circular however, so consequences of a nonmonotonic theory are captured via a fixed-point construction. T is a *fixed point* of a set of sentences A iff

$$T = Cn(A \cup \{M\alpha : \neg\alpha \notin T\})$$

where $Cn(S)$ is the set of (first-order) consequences of S. The nonmonotonic consequences of a theory A are those sentences true in all fixed points of A.

A nonmonotonic theory, as with default theories, can have one, many, or no fixed points. The theory consisting of the flying bird default together with the fact $\texttt{fly}$ has just one fixed point, containing the appropriate conclusion $\texttt{fly}$. The theory

$$\{Mc \supset \neg d, Md \supset \neg c\}$$

has two fixed points (one with $\neg d$, the other $\neg c$), while $\{Mc \supset \neg c\}$ has no fixed points.

We notice immediately that, while nonmonotonic logic deals with a modal language, the only inferences it allows (other than first-order) using modal sentences is the addition of a sentence $M\alpha$ when α is not "known." This weak interpretation of consistency leads to undesirable consequences. For instance, the theory $\{Mc, \neg c\}$ is consistent, as is $\{M(c \wedge d), \neg Mc\}$. This difficulty can be attributed to a lack of semantics for nonmonotonic logic (or correspondingly, the lack of axioms restricting the interpretation of the M operator).

In (McDermott 1982), the semantic difficulties surrounding the modal connective are addressed. The language of nonmonotonic logic remains the same, and the definition of a fixed point of theory

A is as above, except now $Cn(S)$ refers to the closure if S under deducibility in some modal logic (McDermott discusses T, S4, and S5). The consistency operator is restricted in its interpretation by the axioms and inference rules of the appropriate modal logic. McDermott also provides something of a semantic account for this version of nonmonotonic logic.

Unfortunately, for the most reasonable version of modal nonmonotonic logic, NMS5, the formulation proves to be equivalent to standard, monotonic S5, and hence useless as far as nonmonotonic inference is concerned. Conversely, the other candidates are too weak to capture the intuitive interpretation of the consistency operator (although, see (Marek, Shvarts and Truszczyński 1991)).

Autoepistemic Logic

Moore (1985) analyzes the difficulties with nonmonotonic logic and proposes an alternative *autoepistemic logic*. He begins by observing that rather than default or defeasible reasoning, nonmonotonic logic is better suited to *autoepistemic* reasoning, or to an agent reasoning about its own beliefs. With this in mind, he augments classical propositional logic with the modal connective L, $L\alpha$ being interpreted as "α is believed."[10] Moore claims that autoepistemic reasoning is nonmonotonic, not because of any unsoundness of inference, but rather due to its indexical nature: $\neg L\neg\alpha$ refers to the consistency of α with an agent's *current* set of beliefs.

Given an initial set of beliefs or premises A, Moore characterizes those beliefs that are reasonable for an agent to hold. Call an autoepistemic theory T *stable* iff

1. T is closed under tautological consequence,
2. If $p \in T$, then $Lp \in T$, and
3. If $p \notin T$, then $\neg Lp \in T$.

Stable sets of belief are semantically complete in the sense that an agent knows all consequences (introspective and tautological) of its initial knowledge. Given initial premises A, stability means a complete theory T includes

$$Cn(A \cup \{Lp : p \in T\} \cup \{\neg Lp : p \notin T\})$$

where Cn refers to tautological consequence. In order to capture the notion of soundness, T is said to be *grounded* in A iff T is identical to this set, rather than merely including it. *Expansions* of A are just those autoepistemic theories that are stable and grounded in A. Unsurprisingly, a theory may possess one, many or no expansions.

In contrast to the original version of nonmonotonic logic, autoepistemic logic expansions include the set $\{Lp : p \in T\}$ (not used in nonmonotonic logic), along with the set $\{\neg Lp : p \notin T\}$ (which *is* used by McDermott and Doyle). This provides a stronger interpretation of "consistency" or "belief," but one that is not as strict as that of McDermott's NMS5 (since tautological consequence is used, rather than S5-provability).

Konolige (1987) introduces an alternative characterization of autoepistemic logic and defines stronger notions of inference to deal with the problem of self-justified expansions. Konolige also investigates the relationship between default logic and autoepistemic logic, and this is pursued further in (Marek and Truszczyński 1989) and (Truszczyński 1991). Levesque (1990) provides an

[10] This can be viewed as the dual of McDermott and Doyle's M operator.

appealing semantics for autoepistemic logic in terms of *only knowing*. This will be discussed at length in Chapter 6, along with certain problems concerning the representation of default knowledge as autoepistemic statements.

2.2.5 Applications of Default Reasoning

Logic Programming and Negation-as-Failure

Logic programs provide one method of implementing the ideas upon which default reasoners are based. Systems for logic programming, such as Prolog, often use the technique of *negation-as-failure* (NF) to achieve the desired nonmonotonicity. Given a goal $\neg p$, a logic program attempts to prove the goal p. If each branch of the evaluation tree fails (i.e., goal p finitely fails) then $\neg p$ is "proven" using NF. Clearly, this approach is nonmonotonic (for adding p to a program prevents the proof of $\neg p$) and can be used for default reasoning in a number of ways. For instance, Reiter (1987a) suggests logic programs might be used to implement McCarthy's (1986) simple abnormality theories. Consider program P with clauses

$$\texttt{fly(x)} \leftarrow \texttt{bird(x)} \wedge \neg\texttt{ab(x)}$$
$$\texttt{ab(x)} \leftarrow \texttt{penguin(x)}$$
$$\texttt{bird(A); bird(B); penguin(B).}$$

Using NF, $\neg$`ab(x)` is proven whenever `ab(x)` cannot be. This has the effect of circumscribing `ab` in P, resulting in `fly(A)` being concluded, but not `fly(B)`.

In general, logic programs are "procedural" in the sense that formal characterizations of their behavior is often lacking. However, recently much effort has been applied to providing formal semantic and syntactic accounts for programs. Reiter's (1978a) *closed-world assumption* (CWA) is one such attempt. If a program is considered to be a deductive database, often negative information is not explicitly represented. For example, if A is not an employee of company X, this fact is represented by *leaving out* the entry $\langle A, X\rangle$ in the *Employee* relationship table. The CWA attempts to capture this notion formally. For a given program P, define

$$ext(P) = \{\neg A : groundatom(A) \text{ and } P \not\models A\}.$$

Then $CWA(P) = P \cup ext(P)$. The consequences of program P are then identified as those (first-order) consequences of $CWA(P)$, obtained by adding to P all negative ground atoms not entailed by P.

While this captures the notion of implicitly represented negative information, some problems exist with this formulation. Most importantly, for general logic programs,[11] $CWA(P)$ can be inconsistent when P is not, for instance, by allowing indefinite information about ground atoms (e.g., $p \vee q$). Reiter (1978a), however, shows that if P consists of Horn clauses, $CWA(P)$ is consistent whenever P is.

While the assumption of complete information is often appropriate for deductive databases, in other logic programs the CWA might be too strong. Clark's (1978) *predicate completion* is another attempt to explain the mechanism of NF. Let P be any program consisting of extended Horn clauses. For each clause $p(t_1, \ldots t_n) \leftarrow L_1, \ldots L_m$, its *general form* (cf. Shepherdson (1988)) is

$$p(x_1, \ldots x_n) \leftarrow \exists \overline{y}(x_1 = t_1 \wedge \ldots x_n = t_n \wedge L_1 \wedge \ldots L_m).$$

[11] Programs consisting of *extended* Horn clauses, i.e., Horn clauses allowing negative literals in the body.

For each predicate symbol p, let $B_1, \ldots B_n$ be the bodies of the general forms of the clauses in which p occurs in the head. The *completed* version of predicate p is $p(\overline{x}) \equiv B_1 \vee \ldots B_n$. $COMP(P)$ is the collection of completed predicates from program P. The idea is that when we write a series of clauses *about* predicate p, we intend these clauses to be the *only* ways in which the program can prove the truth of p. As with the CWA, the consequences of a program are intended to the first-order consequences of the completed program $COMP(P)$. For instance, if we complete the flying bird program above, we obtain the same results as given by Prolog's NF mechanism.

Predicate completion can be seen to give a more intuitive account of general logic programs than does the CWA. For example, the CWA can provide no new positive information, only augmenting P with new negative facts; thus, it can't derive `fly(A)` in the previous example.[12] Predicate completion is more useful in this respect, but when restricted to definite Horn programs $COMP(P)$ is similarly restricted in its ability to derive new positive facts. Furthermore, $COMP(P)$ can be inconsistent (e.g., $p \leftarrow \neg p$), but this cannot occur when P is in Horn form. It should also be noted that the CWA has the advantage of relying on the semantic content of a program, while $COMP(P)$ is sensitive to the syntax of a program as well. Shepherdson (1988) shows a number of ways in which the CWA and predicate completion differ, despite their superficial similarities.

As far as accounting for NF, neither approach is entirely adequate. However, NF on program P is shown by Clark (1978) to be sound with respect to $COMP(P)$ and, by Shepherdson, to be sound with respect to $CWA(P)$ (cf. Shepherdson (1988)). While not complete for either approach in general, Shepherdson (1988) reports on some partial completeness results for NF. The CWA has also been extended and refined by several people (Minker 1983; Gelfond, Przymusinska and Przymusinski 1989) to the point of being almost a form of circumscription.

Diagnosis from First Principles

Default reasoning and nonmonotonicity is not a quality unique to "commonsense" reasoning. Diagnostic reasoning is the task of determining an explanation for the aberrant behavior of some system (say, a physical device). As opposed to the *experiential* or *rule-based* approach, diagnosis from first principles uses structural and behavioral knowledge of a system to derive diagnoses, rather than heuristic information. Reiter (1987b) presents a theory of diagnosis based on these ideas. A *system*, for which a diagnosis will be made, is a set of first-order sentences SD (the *structural description*), and a set of constant symbols $COMP$ (the *components* of the system). The structural description is intended to describe the behavior of the system when it is functioning properly, how it can fail, and how such failures manifest themselves in system behavior. For instance, a typical sentence might take the form "If condition X holds and component c_1 is functioning properly, then observation Y should be true." Reiter suggests the use of the distinguished predicate ab, standing for abnormality (or failure) of components (cf. McCarthy (1986)). The above sentence could be represented as $X \wedge \neg \mathrm{ab}(c_1) \supset Y$. Given a set of observations OBS, a *diagnosis* is some minimal set $\Delta \subseteq COMP$ such that $SD \cup OBS \cup \{\neg \mathrm{ab}(c) : c \in COMP - \Delta\}$ is consistent. Hence, if $\{\neg \mathrm{ab}(c) : c \in COMP\}$ is consistent with $SD \cup OBS$, the empty set is the only diagnosis, since nothing in the observations indicates deviant behavior.

This formulation is nonmonotonic in the sense that the (minimal) diagnoses of some set of observations need not be the diagnoses for some superset of those observations. If a diagnosis were *any* set of abnormal components that explains system behavior, nonmonotonicity could not arise; new observations could only remove candidate diagnoses. However, we can view this theory

[12] In fact, it *does* entail `fly(A)`, but only because $CWA(P)$ is inconsistent in this case.

of diagnosis as using preferred theories or models of the system behavior, those in which as few components as possible have failed. This nonmonotonicity suggests a relation to default reasoning and Reiter (1987b), in fact, shows how to model this system using default logic. The system representation also suggests the possibility of implementation using circumscription, by minimizing the predicate ab.

Poole (1989) provides a similar treatment of diagnosis, but allows candidate diagnoses to be arbitrary sets of sentences (taken from a candidate set of hypotheses) that explain the observed behavior, rather than a set of faulty components. Poole (1988) also claims that this framework, upon which the Theorist system is based, provides a coherent semantical basis for default reasoning. We will examine the Theorist framework in Chapter 7.

2.3 Modal and Conditional Logics

2.3.1 Modal Logics

We've seen that material implication is not adequate for representing default rules. Similar deficiences motivated the development of modern modal logic. In particular, implication is inadequate as an account of entailment due to the *paradoxes of material implication*:

$$p \supset (q \supset p)$$
$$\neg p \supset (p \supset q).$$

The first says that p is implied by anything if p is true; the second that p implies anything if p is false. While perfectly reasonable truth-functional principles, they are not acceptable properties of entailment. For instance, if p is false, it does not mean that anything at all *follows from* p. To capture this stronger sense of implication it was suggested (e.g., by MacColl and C.I. Lewis, cf. (Hughes and Cresswell 1968)) that $p \supset q$ should be (logically) *necessary* if p entails q. If we use $\Box$ to stand for necessity, entailment corresponds to *strict implication* $\Box(p \supset q)$.[13] Modal logics provide accounts of just such a necessity operator (though $\Box$ is often given other interpretations, and will be given two distinct readings in this thesis).

The presentation and terminology of propositional modal logic we give here is derived from that of (Hughes and Cresswell 1984), together with certain ideas from (Segerberg 1971). The reader is referred to these works along with (Hughes and Cresswell 1968) and (Chellas 1980) for further details.

A modal language $\mathbf{L}_M$ is any propositional language augmented with the unary connective $\Box$, so that $\Box\alpha$ is well-formed just when α is. $\Box\alpha$ is typically read as "α is necessary." The connective $\Diamond$ is introduced by definition as $\neg\Box\neg$, and $\Diamond\alpha$ is read as "α is possible." A *modal system* or *modal logic* S is any set of sentences $S \subseteq \mathbf{L}_M$[14] that includes all propositional tautologies and the axiom **K**, and is closed under modus ponens, uniform substitution and the rule of *necessitation*, **Nec**:

K $\Box(A \supset B) \supset (\Box A \supset \Box B)$

Nec From A infer $\Box A$

The smallest such set is the weakest modal logic K. For any modal system S we define derivability as follows:

[13] Naturally there are other opinions on nature of entailment, e.g., (Anderson and Belnap 1962).

[14] More precisely, these are *normal* modal systems. Nonnormal systems are not considered in this thesis. We use "S" to denote a logic while S is the corresponding set of sentences.

Definition 2.1 A sentence α is *provable in* S (written $\vdash_S \alpha$) iff $\alpha \in S$. α is *derivable* from a set $\Gamma \subseteq \mathbf{L}_M$ (written $\Gamma \vdash_S \alpha$) if there is some finite subset $\{\alpha_1, \ldots, \alpha_n\}$ of Γ such that $\vdash_S (\alpha_1 \wedge \ldots \wedge \alpha_n) \supset \alpha$.

We will discuss various extensions of K, but it will be instructive to motivate these systems semantically. The semantic account we adopt, taken from (Hughes and Cresswell 1984), draws its inspiration from the work of Kripke (1963).

Definition 2.2 A *modal model structure*[15] is a triple $M = \langle W, R, \varphi \rangle$ where W is a set, R is a binary relation on W and φ maps $\mathbf{P}$ into 2^W.

The members of W are referred to as *possible worlds* or states of affairs. $\varphi(A)$ is the set of worlds where A holds, but we often talk about the induced valuation associated with $w \in W$. Thus we may think of φ as assigning propositional valuations to members of W, those valuations determining the facts true at those states of affairs.

We take possible worlds to be hypothetical states of affairs or *counterfactual situations* (among whose number we count the "actual world"). Certain situations might be consistent with an agent's knowledge, or such that an agent has no reason to believe they do not, in fact, obtain; but we assume an agent can also "imagine" or conceive of situations in which a proposition A holds, even though it knows that A is actually false (see Section 2.3.2 on hypothetical deliberation and conditionals). In a quantificational setting, a model must also map a domain of individuals to each world, over which quantification takes place (Kripke 1963). This account, and, indeed, the very idea of quantified modal logic, has been subject to criticism, most notably by Quine.[16] In Chapter 8, we discuss the potential extension of our approach to the first-order case, and discuss the difficulties that might arise.

A model also contains an *accessibility relation* or *alternativeness relation* R on W. When wRv we say v is accessible to w, or w sees v, and when this is the case we intend (on one standard interpretation at least) that v is a possible alternative state of affairs given w. In other words, if w is the actual state of affairs, one is unwilling to reject v as possibly being the actual world. For this reason a sentence α is *necessary* from the perspective of w just in case α is true at all worlds accessible to w. Given the definition of $\Diamond\alpha$, we say α is possible just when it is true at some accessible world.

Definition 2.3 Let $M = \langle W, R, \varphi \rangle$ be a Kripke model with $w \in W$. The *satisfaction* of a formula α at w in M (where $M \models_w \alpha$ means α is true at w) is defined inductively as:

1. $M \models_w \alpha$ iff $w \in \varphi(\alpha)$ for atomic sentence α.
2. $M \models_w \neg\alpha$ iff $M \not\models_w \alpha$.
3. $M \models_w \alpha \supset \beta$ iff $M \models_w \beta$ or $M \not\models_w \alpha$.
4. $M \models_w \Box\alpha$ iff $M \models_v \alpha$ for every v such that wRv.

If $M \models_w \alpha$ we say that M *satisfies* α at w. α is *valid on* M ($M \models \alpha$) just when $M \models_w \alpha$ for all $w \in W$.

We denote by $\|\alpha\|$ the set

$$\{w \in W : M \models_w \alpha\}.$$

[15] Often called a Kripke model, or simply a model.

[16] See, for example, (Quine 1961).

Strictly speaking, some reference to M should be made in this notation, but the model we have in mind when writing this will always be clear from context.

Modal systems are typically associated with various classes of models via characterization results. Let $\mathcal{C}$ be some class of models. α is *valid* with respect to $\mathcal{C}$ iff $M \models \alpha$ for all $M \in \mathcal{C}$, which we write as $\models_{\mathcal{C}} \alpha$. Let S be a modal system. S is *sound* with respect to $\mathcal{C}$ iff $\vdash_S \alpha$ implies $\models_{\mathcal{C}} \alpha$. S is *complete* with respect to $\mathcal{C}$ iff $\models_{\mathcal{C}} \alpha$ implies $\vdash_S \alpha$. $\mathcal{C}$ is *characterized* by S iff S is sound and complete with respect to $\mathcal{C}$.

Most important modal systems correspond to interesting classes of models. For instance, the class of all models is characterized by the logic K. If we restrict the relation R to be reflexive we determine the logic T, which is K plus the axiom

T $\Box A \supset A$.

This corresponds to the intuition that all worlds consider themselves possible, or that the actual world is possible. Adding to T the axiom

4 $\Box A \supset \Box\Box A$

we get the logic S4, which corresponds to the class of reflexive and transitive models. This system is especially important when we use accessibility to represent some *ordering* on possible worlds (e.g., a temporal ordering, see footnote below). Most well-behaved notions of ordering are at the very least reflexive and transitive, and, as pointed out by Segerberg (1971), this class of models is especially nice to study.

Given any model for which R is transitive (and we will always assume reflexive), a *cluster* is defined to be any subset $U \subseteq W$ such that each member of U is mutually accessible (i.e. if $u, v \in U$ then uRv and vRu), and no proper superset of U has this property. In other words, R as restricted to U forms a maximal subrelation of R that is an equivalence relation. Given this, any such accessibility relation can be viewed as a partial order on the set of clusters in M. With an ordering interpretation of R, clusters are the maximal sets of worlds all members of which have the same "rank." Worlds are comparable in R iff their containing clusters are comparable in the induced partial order (see Figure 2.1).

Another important system is S5, which imposes an extreme ordering on worlds. It adds to S4 the axiom

5 $\Diamond A \supset \Box\Diamond A$

and characterizes the class of models formed from equivalence relations (reflexive, transitive and symmetric). Such a model consists of a set of mutually inaccessible clusters, so the induced partial order is empty (see Figure 2.2). Between S4 and S5 lies the logic S4.3, which we will have occasion to use in later chapters. This logic is especially important given the ordering interpretation of R, and consists of S4 plus the axiom

D $\Box(\Box A \supset B) \vee \Box(\Box B \supset A)$.

This adds *total connectedness* to reflexivity and transitivity; that is, for all v, w either vRw or wRv. Such structures consist of a set of clusters that is totally ordered, so R is a *weak linear ordering* on W (see Figure 2.3).[17] Because of the implicit ordering imposed on W by R, S4.3 will find

[17] S4.3 also characterizes other classes of models including the larger class of *connected* models (if uRv and uRw then either wRv or vRw). Furthermore, the "bulldozing" technique of Segerberg (1971) shows S4.3 to correspond to the class of linearly ordered models (a totally ordered set of clusters each of which is a singleton). This has important implications for viewing S4.3 as a temporal logic (Prior 1967), where wRv means v is some moment as late or later in time than w. Thus $\Box A$ is read as "At all later moments in time (including the present) A is true."

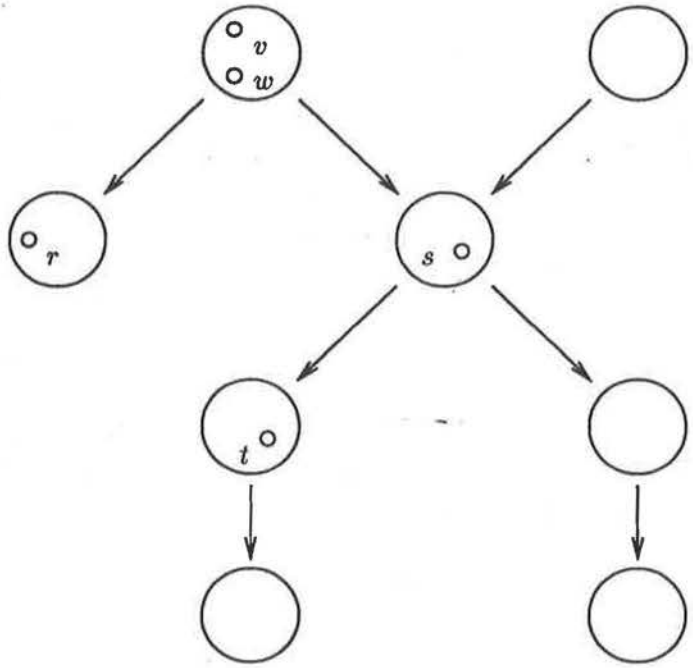

Figure 2.1: A typical S4-model. Each large circle forms a cluster whose elements (represented by points) are mutually accessible. So wRv and vRw. The transitive closure of the arrows determines accessibility outside cluster; thus wRr and wRt, but neither rRw nor rRs, hold.

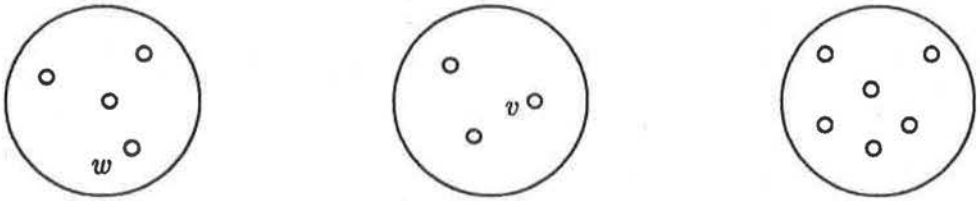

Figure 2.2: A typical S5-model. It consists of a set of clusters that are mutually inaccessible. So neither wRv nor vRw holds.

Figure 2.3: A typical totally connected model. It consists of a totally ordered set of clusters. So wRv or vRw for all v and w.

application in the interpretation of default rules.

Modal logics can also be generalized to include more than one modal operator. We will use a series of bimodal logics in Chapters 5 and 6, with two modal operators corresponding to accessibility and inaccessibility. Typically, such logics have model theories with separate accessibility relations for each modal connective. Bimodal logics are often important for dealing with temporal interpretations of accessibility, and in this context constrain one relation to be the inverse of the other (Segerberg 1970). This contrasts with our treatment where we constrain a second (implicit) relation to be the complement of the first. Multimodal logics with a finite number of boxes are often used for representing the knowledge of multiple agents in an epistemic setting (Halpern and Moses 1985), and infinite families of operators find applicability in dynamic logic (Pratt 1976).

2.3.2 Conditional Logics

While the modal concept of strict implication is an adequate account (for some) of entailment, it still has essentially the same drawbacks as material implication when viewed as a representation of typical linguistic usage of conditional constructions, or of default rules. The paradoxes have counterparts for strict implication certainly, but more discouraging is the fact that strict implication is still "monotonic"; we have the strengthening rule of inference:

$$\Box(A \supset C) \supset \Box(A \wedge B \supset C).$$

Conditional logics have appeared over the last number of years to account for the properties of conditional statements in natural language, statements of the form "If ...then ..." (or some paraphrase thereof). Focusing on propositional logics, conditional logics are generally based on the language of CPL augmented with a conditional connective, variously denoted $\Rightarrow$, $>$, $\Box\!\!\rightarrow$, or otherwise. The convention here shall be to use $>$ for arbitrary, abstract conditionals, and reserving $\Rightarrow$ for the conditional we develop here.

The two main types of conditional statements are *indicative conditionals* and *subjunctive conditionals.* Indicative conditionals are generally intended to be statements about how the world actually is, for example,

"If it rained, Pete didn't show up for our match."

Subjunctive conditionals usually express how the world might have been or could be, for example,

"If it had snowed, we would have gone skiing," or

"If it should snow, we would go skiing."

While definitive boundaries between the two classes of statements are rarely proposed, or even thought to exist, there is little doubt that the classes must be treated distinctly (Jackson 1987; Appiah 1985).

The general approach toward the treatment of subjunctive conditionals is (relatively) uncontroversial, while it is less so for indicative conditionals. Subjunctive conditionals are generally viewed as having truth conditions based on some possible worlds semantics; the controversy surrounding indicative conditionals is due to the debate over whether or not indicative conditionals *have* truth conditions. An example illustrates their problematic nature. Clearly, normal conversational conventions do not allow one to assert "If it rained, Pete didn't show up for our match" merely because it is known that no rain had fallen. Yet one paradox of material implication allows such an inference, namely the conclusion of $A \supset B$ from $\neg A$. This is evidence supporting the claim

that indicative conditionals do not have truth conditions corresponding to those of the material conditional. This is just one of many problems that arise in trying to determine a set of truth conditions for indicative statements. There are various accounts of indicative conditionals suggesting alternatively that indicatives have no truth conditions but instead have assertibility conditions (e.g., "If A then B" is *appropriate* when $P(B|A)$ is high), that they have the truth conditions of the material conditional, or both. We will not discuss indicatives further here for it is subjunctives that are of more interest in this thesis. We refer the reader to, for example, (Adams 1975; Appiah 1985; Jackson 1987; Lewis 1976).

Subjunctive conditionals are generally thought to have truth values determined by hypothetical possible worlds, following (Stalnaker 1968). While specific details differ, this thesis is maintained in most proposals dealing with subjunctive conditionals and discrepancies exist only over relatively minor details (compared to indicatives). Some basic considerations for any semantic account of conditionals are provided by Stalnaker and are recalled here.

First, a truth-functional analysis of conditionals (i.e., as material conditionals) is inappropriate. For instance, the falsity of the antecedent of a conditional is insufficient reason to affirm the truth of the conditional. Consider Stalnaker's (now dated) example:

> "If the Chinese enter Vietnam, the U.S. will use nuclear weapons."

Even if one believes the Chinese will stay out of Vietnam no matter what events occur, this is no reason to affirm the conditional. This suggests a second consideration, namely that a "connection" should exist between antecedent and consequent. However, a connection isn't always required for a conditional to be true: consider one who believes that (for whatever reason) the U.S. will definitely use nuclear weapons in Vietnam. In this case, one should assent to the truth of the conditional, even if no opinion about the Chinese is forthcoming. Stalnaker's final (rough) suggestion for determining the truth of a conditional is known as the *Ramsey test* (see Chapter 6). Briefly, one should adopt a hypothetical belief in the antecedent of the conditional, make some *minimal* changes in old beliefs to accommodate this, and finally consider whether the consequent follows in this new belief state. This view is also detailed in Nute's (1980) account of *hypothetical deliberation.*

Stalnaker (1968) develops a possible worlds semantics for such a conditional logic. Formulae are interpreted with respect to a *model structure* and a *selection function*, where a model structure consists of a set of possible worlds (including an absurd world) and an accessibility relation, and a selection function takes a proposition and a possible world as arguments and has as a value a possible world. The selection function f is intended to represent the selection of the most reasonable or *closest*[18] possible world in which to consider the truth of the antecedent of a conditional, in the sense discussed above. The conditional sentence $A > B$ is true at a possible world w iff B is true at $f(A, w)$. A number of restrictions are placed on the selection function, all reflecting reasonable intuitions regarding hypothetical deliberation. For all antecedents A and B and worlds w and v, the following conditions should hold:

(a) A is true at $f(A, w)$.

(b) $f(A, w) = \lambda$ (the absurd world satisfying all sentences) only if v is inaccessible from w for any A-world v.

(c) If A is true at w, $f(A, w) = w$.

(d) If A is true at $f(B, w)$ and B is true at $f(A, w)$ then $f(A, w) = f(B, w)$.

[18] For instance, the world selected should differ minimally (in some pragmatic sense) from the actual world.

Lewis (1973a) presents an alternative semantics for (counterfactual) conditionals that deals with some problems occurring in Stalnaker's logic. Lewis abandons two assumptions: the Limit Assumption, which states that a closest (most similar to the actual) possible world in which the antecedent of a conditional holds must exist; and the Uniqueness Assumption, that exactly one closest antecedent-world need exist.

The Uniqueness Assumption is due to Stalnaker's view that exactly one antecedent-world is closest to the actual world. The logic presented by Stalnaker has as a theorem the Law of the Conditional Excluded Middle

CEM $\Diamond A \supset (\neg(A > B) \equiv (A > \neg B))$.

This theorem entails the (objectionable, according to Lewis) theorem

$$(A > B) \vee (A > \neg B).$$

Lewis disputes the validity of such a theorem with the classic pair of conditionals (due to Quine)

"If Bizet and Verdi had been compatriots, they would have been Italian" and

"If Bizet and Verdi had been compatriots, they would have been French."

Intuitively, neither of these statements is true because it seems equally likely that Verdi could have been French or Bizet Italian. Lewis proposes that hypothetical deliberation may involve a number of equally (most) similar possible worlds where the antecedent of a conditional holds.[19] We discuss the Limit Assumption and Lewis's key counterfactual logic VC in Chapters 4 and 6.

Lewis's final analysis for determining the truth of a (counterfactual) conditional is as follows: $A > C$ is true at a world w iff some world (accessible to w) in which A and C hold is closer to w than any world (accessible to w) at which A and $\neg C$ hold, if there are in fact accessible A-worlds.

In general, conditional implication is thought to lie between material implication and strict implication, being somewhat stronger than material and weaker than strict. A view of conditionality is that it expresses some sort of *relative* necessity (Chellas 1975; Nute 1980). To quote Chellas (1975, p.133): "'If A, then B' means that the proposition expressed by B is in some way necessary with respect to that expressed by A." Therefore, conditionality can be viewed as either a sententially-indexed modality or a propositionally-indexed modality. A selection function determines, given the antecedent A of a conditional $A > B$, the worlds to be considered in deciding if B is (relative to A) necessarily true.

If the selection function maps propositions[20] and possible worlds into sets of possible worlds[21], then the conditional connective can be taken as a propositionally-indexed modal operator (e.g., see the unifying approach to conditional logic found in (Chellas 1975)). This view ensures that any conditional logic based on this type of selection function will have as a rule of inference full substitution of equivalents. If the selection function maps sentences of the language and possible worlds into sets of possible worlds[22] then the conditional connective can be taken as a sententially-indexed modal operator (e.g., see the general approach to conditional logic found in (Nute 1980)). In this case, the inference rule RCEC will hold (this rule allows equivalents to be substituted in the consequents of conditional sentences), but the rule RCEA (this rule allows equivalents to be substituted in the antecedents of conditionals), in general, will not be valid. For a discussion of why RCEA might not be appropriate as a rule of inference for all conditional logics, see (Nute 1980).

---

[19] Stalnaker (1980) defends **CEM** by appeal to van Fraassen's notion of *super-valuations*.

[20] We take a proposition to be a set of possible worlds, identified with the "sentence" it makes true.

[21] That is, $f : 2^W \times W \mapsto 2^W$.

[22] That is, $f : \mathbf{L} \times W \mapsto 2^W$.

2.3.3 Conditional Logics and Default Reasoning

Conditional logics have found application for many types of conditionality. For example, counterfactual conditionals have been described by Lewis and others using conditional logic (Lewis 1973a; Lewis 1973b; Nute 1975). Standard deontic logics based on the typical unary modal operator have been found to be inadequate in certain respects, most notably by leading to certain paradoxes (von Wright 1964; Åqvist 1967; van Fraassen 1972). Conditional logics based on the notion of conditional obligation seem more promising in this respect.

Conditional logic can be seen to be generally useful for applications in default reasoning. A number of properties of conditional logics suggest that they are ideal candidates for expressing facts of the form "If ...then normally ...". Unlike corresponding sentences using the connective of material implication, conditional sentences of the type problematic in default reasoning can be asserted consistently in most conditional logics. Consider the pair of conditionals

"If a match were struck, it would light" and

"If a wet match were struck, it would not light."

While the pair of formulae $M \supset L$ and $W \wedge M \supset \neg L$ is inconsistent with the fact $W \wedge M$, the pair $M > L$ and $W \wedge M > \neg L$ is not.

A number of properties, desirable in default reasoning, are true of the conditional connective in most conditional logics. We now take the connective $\Rightarrow$ to be the specific conditional "normally implies." $A \Rightarrow B$ is interpreted as "If A then normally B." For example, consider the following "fallacies," or inference rules that hold for material implication but not for conditional connectives in most conditional logics:

Strengthening From $A \Rightarrow B$ infer $A \wedge C \Rightarrow B$.

Transitivity From $A \Rightarrow B$ and $B \Rightarrow C$ infer $A \Rightarrow C$.

Contraposition From $A \Rightarrow B$ infer $\neg B \Rightarrow \neg A$.

As well, while not true of Stalnaker's or Lewis's semantics, most general approaches to conditional logic do not require that all logics obey:

Modus Ponens From A and $A \Rightarrow B$ infer B.

Clearly, none of these inference patterns is desirable in a logic for default reasoning, as discussed previously.

There are a number of advantages to using the conditional logic approach to default reasoning. First, the language of conditional logic provides a naturalness of expression that cannot be found in most current schemes. $A \Rightarrow B$ is a concise way of expressing that B should normally follow from A. There is no need to account for qualifications to prevent the inconsistency of such an assertion, for the connective $\Rightarrow$ is "nonmonotonic." Second (once a semantics for this interpretation of the connective $\Rightarrow$ is developed), conditional logic offers a more intuitive semantics for sentences in a default theory, and one that is well-developed (since the approach developed in this thesis is based on existing modal logics). The semantics provided for the interpretation of statements and default rules in many systems of default reasoning are problematic or unintuitive, or perhaps non-existent. Finally, with conditional logic, conditional statements can be interpreted as default rules. One can reason about default rules, or their negations, deriving new ones from existing rules together with facts about the world. This is because these "default rules" are themselves merely statements of fact about the world being modeled, allowing interaction with other default rules and facts. This is not

the case with most approaches. Even with techniques such as nonmonotonic logic or autoepistemic logic, where the default rules *are* sentences in the theory, the propriety of the rules that can be derived is somewhat dubious (Reiter 1987a). So conditional logic does offer some advantages for certain aspects of default reasoning.

Conditional logics have been examined previously in the context of AI applications and nonmonotonic reasoning. For instance, Ginsberg (1986) examines the use of counterfactuals in AI, listing uses in such areas as inconsistency-detection in knowledge bases, planning, and diagnosis. But more recently, conditional logics have started to find popularity as representation systems for default rules and specifications for default reasoning. One of the earliest attempts to develop a default reasoner based on conditional logic was made by Nute (1984b; 1984a). In (Nute 1984b), he describes PROWIS, a system implemented in Prolog for reasoning with subjunctive conditionals, and based on the logic VW. The idea is to answer queries of the form "Would C be true?" given an initial database of facts, and conditionals of the form $A > B$. In general, the query "Would C?" is answered positively if there is some conditional $A > C$ that holds along with A, and there is no true A', stronger than A, such that $A' > \neg C$ holds also. This encompasses the notion of *specificity*: we accept as true the consequences of the more explicit information in the case of conflict. For example, given the conditionals above describing the behavior of matches, we would accept the conclusion $\neg L$ (rather than L) in the presence of $W \wedge M$ (which is more specific than M).

Nute (1984a) provides a formal notion of derivation of subjunctive consequences that accounts for (and generalizes) the behavior of the PROWIS system. This system, LDR1, is again based on Horn clause logic and uses the notions of *absolute* and *defeasible* rules, and *defeaters*. In this sense, it is very similar to Pollock's system of defeasible inference. Also related is the fact that it is a purely syntactic system. The defeasible rules and defeaters are intended to represent a certain type of conditionality, yet the system fails to respect the semantics of the logic on which it is based. For instance, $(A > B) \supset (A \supset B)$ is VW-valid, but neither of Nute's systems necessarily draws the conclusion B from A and $A > B$. Since the semantics of VW cannot apply to Nute's defeasible rules, these essentially have the status of unanalyzed conditional statements.

A promising approach to the use of conditional logic for default reasoning has been initiated by Delgrande (1987; 1988), among others. In this thesis we will adopt the conditional approach to default reasoning precisely because of the advantages discussed above. Along the way we will present, discuss and compare the more important conditional representations of defaults, including the work of Delgrande (1986; 1987; 1988), Lehmann (Kraus, Lehmann and Magidor 1990; Lehmann 1989; Lehmann and Magidor 1990) and Pearl (Pearl 1988; Pearl 1990; Goldszmidt and Pearl 1990).

Chapter 3

Models of Belief Revision

In the previous chapter we looked at the problem of default reasoning, the task of drawing plausible conclusions based on a static set of beliefs. We've seen that this notion of consequence is nonmonotonic in the sense that as belief sets are augmented with new information certain inferences might be deemed less plausible than they once were, and certain conclusions unacceptable. Because default reasoning is nonmonotonic and states of belief are in constant flux, it is necessary to accommodate the revision of beliefs. In this chapter we will survey various formal approaches to belief revision.

3.1 Truth Maintenance Systems

Until recently, belief revision in AI had generally only been studied under the guise of *truth maintenance systems* (Doyle 1979; de Kleer 1986). These systems are tools used to manage a changing set of beliefs in such a way that *justified* beliefs are labeled and can be identified as such, while *unjustified* facts are not accorded the status of beliefs. The complexity of the task lies in the dynamic aspect of keeping track of justifications. If a certain belief is justified by the presence of other beliefs, the removal of the justifying beliefs should cause the original belief to be discounted unless other justification can be found. This, in turn, may cause other beliefs to be disregarded, or support belief in others still.

Implicit in this approach to revision is a commitment to a *foundational* epistemology. In such a theory only facts having adequate justification are granted the status of beliefs, and a rational agent is required to keep track of such justification on the threat of being forced to give up beliefs. A justification is a set of beliefs that permit one to accept a fact as a justified belief. These justifications cannot be circular and must ultimately be grounded in *foundational states*, or basic beliefs that are self-justifying. For instance, these are often considered to be perceptual states (Quine and Ullian 1970). These theories are especially appealing as normative accounts of belief since they describe a rather intuitive conception of what an agent ought to believe (or at least what one is *permitted* to believe). If one comes to believe A on the basis of learning B, but later has reason to disbelieve B, then the belief in A ought to be suspended as well.

In the original formulation of a truth maintenance system (TMS) Doyle (1979) proposed that the TMS be used in conjunction with some reasoning program. This program transmits various beliefs and reasons for these beliefs (or *justifications*) to the TMS, which is charged with the task of recording and managing these beliefs and reasons. As beliefs change, other beliefs for which these are justifications may have to be added or retracted by the TMS.

A belief is represented as a *node* in the system, whose status is either *in* or *out*. An *innode*

represents a currently held belief while an *outnode* is not currently believed (not to be confused with belief in the negation of the node). The set of innodes at any point in time completely determines the belief set of the program at that time. A *justification* for a node (there may be several) is an ordered pair of lists of nodes, the *inlist* and the *outlist*. A justification is valid for a belief if each node on the inlist is in and each node on the outlist is out. Thus a node is believed just when it possesses some valid justification.

Because a belief can be justified by a lack of belief in other nodes, the TMS can exhibit nonmonotonic behavior. By simply adding the belief in a node on a second node's outlist, the second node can become disbelieved. Of course, this can cause belief in other nodes to be suspended, and still others added. The main task of the TMS is to propagate these changes throughout the system. Doyle presents a number of algorithms for inconsistency removal and change propagation.

Notice belief in a node can be suspended only when some node on the (justifying) inlist becomes out or some outlist node becomes in. Thus all possible qualifications for a nonmonotonic inference must be placed on the outlist in negated form (see Chapter 2). It is not enough to say `bird` is a reason for `fly`. On the outlist of the `bird`-justification for `fly` must be the qualifications `penguin`, `emu`, `brokenWing`, `dead`, etc.

One criticism of the basic justification-based TMS is its inability to consider multiple states of affairs at any point in time, and the difficulty involved in switching contexts (an especially important component in search). The *assumption-based TMS* (ATMS) of de Kleer (1986) generalizes the TMS through the use of *contexts* and *labels* for nodes. Given a set of nodes and justifications, a context for a node is a set of *assumptions* from which the node can ultimately be derived using the justifications. Assumptions themselves are typically viewed as requiring no further justification (like foundational states). The label of the node is the set of all contexts for the node and can be derived from the justifications; and in (Reiter and de Kleer 1987) the ATMS is logically reconstructed to reflect this. The task of the ATMS is to keep track of (but not completely recompute) such labels as new nodes and justifications are added to the system. de Kleer describes such methods.

Since nodes are labeled with various contexts in which they can be assumed to hold, many inconsistent states of affairs can be considered at one time. These contexts are components of various nodes and the set of beliefs (roughly) associated with a particular situation are just those nodes labeled with the appropriate context.

3.2 The AGM Theory of Revision

In foundations theories, and truth maintenance systems in particular, an agent is required to keep track of justifications for beliefs and to give up beliefs when no justification exists. While intuitively appealing, the computational cost of keeping track of justifications might be prohibitive, and certainly psychological evidence overwhelmingly supports the hypothesis that human reasoners do not do this (Harman 1986, Chapter 4). A consequence of foundations theories is the suspension of beliefs when adequate justification cannot be found, for example, when a (sole) justification is forgotten. *Coherence* theories of belief, however, do not require such bookkeeping, and revision is guided by principles that ensure one's entire belief set remains a reasonable view of the world, or *coheres*. Much work in the philosophical community on the logic of theory change is consistent with such a coherentist epistemology in the sense that belief sets are unstructured collections of sentences, and belief revision is not accomplished through the use of justifications.[1]

[1] Of course, belief revision based on unstructured sets of beliefs cannot be said to constitute an adequate coherence theory. Indeed, some degree of implication and explanation is required among beliefs in a coherence theory (Harman

Recently, work on the logic of theory change has been adopted by the AI community for use in the task of belief revision. By far the most influential approach to revision has been that of Alchourrón, Gärdenfors and Makinson (1985), which we refer to as the AGM theory of revision.[2] Most of this work assumes beliefs sets to be modeled as deductively closed sets of sentences, and for concreteness we will assume that the underlying logic of beliefs is CPL, although nothing critical hinges on this decision, the only requirement, in general, being the compactness of the underlying logic. As usual, $\models$ and Cn will denote propositional entailment and consequence respectively. We will use K to denote arbitrary belief sets, and if $K = Cn(KB)$ for some finite set of sentences KB, we say K is *finitely specified* by KB. In this case, we will often refer to the revision of K as the revision of its base set KB, and since this should cause no confusion, we will also refer to the sentence formed by conjoining its members as KB. Typically, it will only be finitely specifiable theories with which we concern ourselves. It may be useful to think of KB as the ***explicit beliefs*** of some agent, while $Cn(KB) - KB$ are the *implicit beliefs*, those inferable given these explicit beliefs (cf. (Levesque 1984b) for a more detailed account of this distinction). This distinction makes clear how to represent in a computer (or other finite agent) the infinite belief set $Cn(KB)$.

Revising a belief set K is required when new information is learned and must be accommodated with these beliefs. If A is consistent with K (that is, when $K \not\models \neg A$), learning A is relatively unproblematic, as the new belief set $Cn(KB \cup \{A\})$ seems adequate for modeling this change in theory. This process is known as *expansion*. For instance, if Craig has no opinion regarding the truth of the fact that Pete has a very expensive new tennis racquet, learning this new information is merely a matter of adding it to his belief set, together with certain implications associated with this belief, perhaps that Pete has gotten a raise or that he is taking his tennis game more seriously (though these are most certainly ***default*** implications).

More troublesome is the revision of K by A when $K \models \neg A$, A being inconsistent with the current state of belief. Simply adding A to the belief set as with expansion is not an adequate solution since an inconsistent theory will be the result. Some beliefs in the original theory must be given up before the new fact can be accommodated. Suppose now that Craig believed, before discovering Pete has a new racquet, that Pete was on a budget and would make no significant purchases. To adjust to this new information Craig must make certain changes to his original set of beliefs. However, there are a number of ways to oblige this new information; for instance, he could drop his belief that Pete is on a budget (perhaps he got a raise) and that Pete is careful with his money, or he could question his belief that Pete is not a kleptomaniac, or that Pete only plays with his own racquets.

Typically, there are a multitude of choices for massaging a theory K in order to adapt to new information. The problem lies in choosing what beliefs to give up, and is further compounded by the fact that, in general, there are no logical grounds for choosing which of these alternative revisions is acceptable (Stalnaker 1984), the issue depending largely on context. Fortunately, there are some logical criteria for reducing this space of possibilities.

The main criterion for discarding some revisions in deference to others is that of *minimal change*. Informational economy dictates that as few beliefs as possible from K be discarded to facilitate belief in A (Gärdenfors 1988), where by "few" we refer not to the number of beliefs given up (alone), but the "quality" of these beliefs (information content). While pragmatic considerations will often enter these deliberations, the main emphasis of the work of Alchourrón, Gärdenfors and Makinson

1986); that is, some "glue" is needed to make beliefs cohere.

[2] Gärdenfors's (1988) provides an excellent exposition of the AGM theory and its relationship to other theories. We will usually refer to this account.

is in logically delimiting the scope of acceptable revisions. To this end, the AGM postulates, given below, are maintained to hold for any reasonable notion of revision (Gärdenfors 1988). We use K_A^* to denote the belief set that results from the revision of K by A, assuming $*$ to be a function mapping belief set-sentence pairs to belief sets. We use K_A^+ to denote expansion.[3]

(R1) K_A^* is a belief set (that is, deductively closed).

(R2) $A \in K_A^*$.

(R3) $K_A^* \subseteq K_A^+$.

(R4) If $\neg A \notin K$ then $K_A^+ \subseteq K_A^*$.

(R5) $K_A^* = Cn(\bot)$ iff $\models \neg A$.

(R6) If $\models A \equiv B$ then $K_A^* = K_B^*$.

(R7) $K_{A \wedge B}^* \subseteq (K_A^*)_B^+$.

(R8) If $\neg B \notin K_A^*$ then $(K_A^*)_B^+ \subseteq K_{A \wedge B}^*$.

The first two postulates state that the result of revising K by A should be a belief set containing A. (R3) and (R4) taken together assert that if A is consistent with K then K_A^* should merely be the expansion of K by A. This seems to reflect our intuitions about informational economy, that beliefs should not be given up gratuitously. This makes (consistent) expansion a special case of revision. (R5) says it is possible to revise a theory consistently to include the belief of any logically consistent sentence, while (R6) asserts that only the semantic content of new beliefs should be considered when performing revision. That the semantic content of K and not its structure is significant follows from the fact that $*$ takes as (part of) its domain deductively closed sets. (R7) and (R8) are essentially (R3) and (R4) applied to conjunctions and iterated revisions. Indeed (R3) and (R4) are derivable from these under the reasonable assumption that $K_\top^* = K$, where $\top$ is the identically-true proposition.

Other than revision (which includes expansion as a special case), one may consider another form of belief change whereby certain beliefs are given up, but not replaced with others. This process is known as *contraction*, and K_A^- denoted the belief set that results from "deleting" A from K. While this appears to be a different form of theory change, it is definable in terms of revision. AGM have provided postulates for coherent contraction functions (see (Gärdenfors 1988) for details), and have shown that $*$ and $-$ are interdefinable in such a way that each set of postulates is satisfied. In particular, we can define contraction by means of the *Harper identity*

$$K_A^- = K \cap K_{\neg A}^*$$

and revision via the *Levi identity*

$$K_A^* = (K_{\neg A}^-)_A^+.$$

Thus, though we consider primarily revision functions, we can effect any form of theory change, including contraction and expansion.

[3]The postulates are taken from (Gärdenfors 1988).

3.3 Alternative Models of Revision

Some other models of belief revision have been proposed that capture roughly the same ideas as the AGM model. For instance, Grove (1988) has proposed a *system of spheres* model for revision[4] that captures a number of intuitions regarding revision. If K is the belief set of some agent then the set of possible worlds $\|K\|$ that satisfy K are precisely those the agent considers epistemically possible. In the course of revision, certain beliefs are given up in order to accommodate others, but giving up beliefs is merely expanding one's set of possibilities, considering possible more states of affairs. Grove captures the intuition that certain beliefs should be given up instead of others by imposing an ordering on possible worlds that reflects the order in which an agent prefers to add states of affairs to its set of possibilities. This ordering is reflected in a collection S of *spheres*, or sets of possible worlds. This collection must be totally-ordered under set inclusion so that S consists of an increasing sequence of sets $S_1 \subseteq S_2 \subseteq \cdots$. The minimum of this sequence S_1 must be $\|K\|$, since this is the set of possibilities an agent most prefers, in accordance with the requirement of informational economy. When revising by A, the idea is to find the minimal A-permitting sphere (any sphere containing some A-world), denoted $s(A)$, and let $K^*_A = \|A\| \cap s(A)$. This corresponds to the notion that the closest (most similar) A-worlds to K are the ones considered possible after revising by A. Grove shows the class of revision functions determined by such models is exactly the class of AGM revision functions. We can view the spheres model as imposing a total preorder on worlds by considering $w \leq v$ iff w is contained in every sphere that contains v.[5]

Another manner in which to specify belief revision is to provide an ordering on formulae reflecting the degree to which an agent is willing to give up these beliefs to oblige new information. Such relations are known as orderings of *epistemic entrenchment* and were proposed by Gärdenfors (1984) as a means of resolving conflict when deciding among different ways of giving up beliefs. Not surprisingly, just as Grove's ordering on worlds captures the AGM postulates, so too can an ordering on sentences. In fact, the natural ordering induced on sentences by Grove's model is closely related to entrenchment orderings.

From (Gärdenfors 1988), let an entrenchment ordering (for a given theory K) be any relation $\leq_E$ on $\mathbf{L}_{CPL}$ satisfying these postulates.

(E1) If $A \leq_E B$ and $B \leq_E C$ then $A \leq_E C$.

(E2) If $A \vdash B$ then $A \leq_E B$.

(E3) If $A, B \in K$ then $A \leq_E A \wedge B$ or $B \leq_E A \wedge B$.

(E4) If $K \neq Cn(\bot)$ then $A \notin K$ iff $A \leq_E B$ for all B.

(E5) If $B \leq_E A$ for all B then $\vdash A$.

The inverse of the entrenchment relation is a *plausibility* relation, which Grove (1988) relates to entrenchment. Intuitively, a sentence A is more *plausible* than B (written $A <_G B$) iff A is more acceptable than B, or would be more readily adopted as a belief if the opportunity arose. Let a *Grove ordering* be any relation $\leq_G$ on $\mathbf{L}_{CPL}$ satisfying the following postulates.

(G1) Either $A \leq_G B$ or $B \leq_G A$.

---

[4] This model is similar to Lewis's (1973a) system of spheres model for counterfactual conditionals, which we will examine in Chapter 6.

[5] This view of Grove's model is quite like the model of revision provided in (Katsuno and Mendelzon 1990). Systems of spheres will be presented more formally in Chapter 6.

(G2) If $A \leq_G B$ and $B \leq_G C$ then $A \leq_G C$.

(G3) If $\vdash A \supset B \vee C$ then $B \leq_G A$ or $C \leq_G A$.

(G4) If $\neg A \notin K$ then $A \leq_G B$ for all B.

(G5) If $\vdash \neg A$ then $B \leq_G A$ for all B.

Grove (1988) shows that such a relation is induced by a system of spheres by defining $A \leq_G B$ iff $s(A) \subseteq s(B)$, and that any such ordering has a corresponding spheres model. Hence, this model is appropriate for revision if we consider (defining $A <_G B$ to mean $A \leq_G B$ and not $B \leq_G A$):

$$B \in K_A^* \text{ iff } A \wedge B <_G A \wedge \neg B.$$

Gärdenfors (1988) shows that the ordering $\leq_E$, defined as $A \leq_E B$ iff $\neg A \leq_G \neg B$, will satisfy (E1)–(E5) iff $\leq_G$ satisfies (G1)–(G5). Hence, revision functions specified by means of epistemic entrenchment are also equivalent to AGM revision functions.

Another model for belief revision proposed by AGM defines revision using the Levi identity in terms of *partial meet contraction functions.* Consider the problem of constructing the contraction K_A^- (from which K_A^* will be defined). One approach is to consider the set of maximal subsets of K that fail to imply A (denoted $K \perp A$), that is, those beliefs sets that give up just enough to suspend belief in A. Given these sets one can define contraction in two obvious ways. The first, *maxichoice contraction*, consists of choosing some element $K_A^- \in K \perp A$. The problem with such contraction functions is evident when we try to define revision, for if we add $\neg A$ to K_A^- (according to the Levi identity) the resulting belief set will be complete. This is due to the fact that either

$$\neg A \supset B \in K_A^- \text{ or } \neg A \supset \neg B \in K_A^-$$

for all B. Hence, revising by $\neg A$ results in commitment to every proposition or its negation. The second choice is *full-meet contraction* in which K_A^- is defined as $\cap K \perp A$. The problem with revision in this case is just the opposite, for $(K_A^-)_{\neg A}^+ = Cn(\neg A)$. In other words, when revising by $\neg A$, *all* other beliefs are given up. Partial meet contraction consists of picking an arbitrary (determined by context) subset of $K \perp A$ via some selection function S, and letting $K_A^- = \cap S(K \perp A)$, hence avoiding the problem of excessively large or small revisions.

One may notice that the problem with maxichoice and full-meet revision is that certain "obscure" consequences of beliefs come into play. For instance, if $A \in K$ then $\neg A \supset B \in K$ is a (frequently trivial) consequence of this belief. Intuitively, these implicit beliefs are generated by explicit belief in, say, A, and revising by $\neg A$ should not only remove A from K, but also the implicit beliefs due to A.[6] Nebel (1989) has proposed just such an approach in which belief sets are represented as finite sets of sentences (not deductively closed) viewed as the explicit beliefs of an agent. The contraction of K by A then amounts to removing from K those explicit beliefs that imply A. Again, some choice may be involved, but whether we choose one or intersect all maximal subsets of K, the results will generally be nontrivial. Unfortunately, such an approach is extremely sensitive to the syntactic structure of the representation of K. For instance, revising K by A has different results when $K = \{A, B\}$ and $K = \{A \wedge B\}$. In the first case $K_A^* = \{B\}$, while in the second $K_A^* = \emptyset$.[7]

[6] This feels much like a foundations theory of revision.

[7] One could argue that this sensitivity to syntax might be useful in practice. For example, if $A \wedge B$ is used instead

Other approaches to revision have also used representations of belief sets that don't allow arbitrary theories, but only those representable by a sentences in some language. This has the advantage of making explicit our assumption that we will generally only be concerned with such finitely representable KBs, yet does not rely on the syntax of the KB or sentence. Katsuno and Mendelzon (1990) define a sentential connective for revision where $A \circ B$ is intended to represent the theory resulting from revising A by B. They also present postulates that characterize revision in the case of a finitary language and show these to correspond to the AGM postulates. Furthermore, they demonstrate that, in this case, the revision of KB by A coincides with imposing a total preorder on interpretations (such that all KB-worlds are minimal) and considering the models of A minimal in this ordering to represent the revised state of affairs. In a sense, this corresponds to Grove's (1988) system of spheres representation that implicitly imposes such a total preorder on possible worlds.

Katsuno and Mendelzon also generalize their version of the AGM postulates to characterize the case where revision is determined by relaxing the ordering on models to be a (simple) preorder (or a partial order). This *partial order revision* has the appealing quality that when ranking states of affairs one need not insist that all worlds be comparable. In contrast, implicit in the AGM approach is the requirement that if a world is neither more nor less similar to KB than another then these worlds are equally similar — there can be no ambivalence. This generalization will be explored further in Chapter 6.

Update

The models of belief revision thus far examined have all been quite abstract and general, stipulating very few conditions on revision, and allowing pragmatic considerations to play a dominant role. Some more concrete proposals do exist however, for instance, for revising databases (e.g., (Fagin, Ullman and Vardi 1983; Dalal 1988); see (Katsuno and Mendelzon 1990) for a brief survey). Dalal's (1988) approach deals with the problem of belief revision by requiring that the truth value of as few atomic propositions as possible change when exacting a revision. In particular, let $diff(w)$ be the minimal number of propositional atoms on which w differs in interpretation from all $v \in \|KB\|$. Then the models of the revised KB when incorporating A are just those $w \in \|A\|$ such that $diff(w)$ is minimal. It is clear that this is a specific instance of a revision operator[8] and does not permit contextual information to bias the process. It should also be evident that this form of revision is not appropriate in general, in the sense that it does not necessarily preserve "information" unless the informational content of each atom is identical, which is unquestionably not the case. For example, if an agent revises its beliefs to include "Bush pushed the button" surely we should not expect a "minimal change in atoms" (so to speak), though it would be *consistent* to maintain beliefs like "The car will start" or "The store will open at 9AM." Unless a representation is sufficiently bizarre, atomic truth values are not an adequate measure of information loss, as pragmatic factors cannot be accounted for (see (Gärdenfors 1988, pp.91–94) for a discussion of how pragmatic concerns influence revision).

Recently, a distinction has been made between the revision of a knowledge base and the *update* of a knowledge base (Winslett 1988; Winslett 1990; Katsuno and Mendelzon 1991; Grahne 1991). The problem with revision that merits this dichotomy is illustrated with the following example.

> Suppose a means the book is on the floor and b means the magazine is on the floor. Then ψ $[a \equiv b]$ states that either the book is on the floor or the magazine is, but not

---

of A and B, some dependence between A and B might be intended by the user. It is debatable, however, whether this is the most compelling or principled way to represent such dependence.

[8] But not an AGM revision operator.

> both. Now, we order a robot to put the book on the floor. The result of this action should be represented by the revision of ψ with a. After the robot puts the book on the floor all we know is a...(Katsuno and Mendelzon 1991, p.390)

However, in this case, since a is consistent with $KB = \{a \equiv b\}$ the AGM postulates assert that the revised KB should be $a \wedge \neg b$, contrary to intuition.

In response to such problems, Winslett (1988) presented the *possible models approach* to update. The idea is to consider not the models of the updated fact closest to *some* model of the original theory (as does Dalal), but rather to consider, *for each model of KB*, those models closest to it. In more detail, suppose we want to update K by A. Let $u \in \|A\|$ and $w \in \|K\|$. We say u is *closest* to w if there is no $v \in \|A\|$ such that v differs from w on the interpretation of fewer atoms than does u (in the sense of set inclusion). The models of the updated belief set K' are just those worlds closest to some world in $\|K\|$; that is

$$\|K'\| = \bigcup_{w \in \|K\|} \{v : v \text{ is closest to } w\}.$$

This notion of update seems to capture the change in belief about a changing world, since for each epistemically possible world we consider the way it might have changed. Revision on the other hand seems more suitable for changing beliefs about a static world. This difference is reflected in the way consistent changes are handled. In the case of revision consistent change amounts to the conjunction of the new information, in accordance with postulates (R3) and (R4). Update, as we've seen, need not satisfy this property.

Again, the possible models approach seems to be a concrete proposal addressing the general phenomenon of update. That minimal change in the world is reflected by minimal change in atomic truth values is in general dubious. Katsuno and Mendelzon (1991) have proposed a general definition of update and provided postulates that capture arbitrary partial orderings on worlds while respecting the intuition that each world be considered separately when updating. Grahne (1991) has axiomatized a logic for updates in which the notion is related to counterfactual implication. The relation of update and revision to conditionals will be examined in fuller detail in Chapter 6.

Chapter 4

Conditional Logics of Normality

Many aspects of reasoning, both in commonsense and expert domains, have a default component. In most circumstances, conclusions must be reached lacking complete information, and cannot be inferred with logical certainty. In Chapter 2 we saw that a multitude of formal systems have been proposed to characterize the notion of reasonable defeasible inference. Yet among the traditional systems, such as default logic, autoepistemic logic and circumscription, an entirely adequate account cannot be found. Much of our default knowledge seems to be based on statements of typicality or normality, but none of these systems can be said to be a logic of such statements.

Consider a standard default statement "birds fly." In the circumscriptive framework of (McCarthy 1986) such a statement is written as

$$\texttt{bird} \wedge \neg \texttt{ab} \supset \texttt{fly}$$

where ab is some abnormality predicate[1] intended to mean that a bird is abnormal with respect to its flying ability. Unfortunately, this statement is made vacuously true by the fact $\neg$bird, as is the circumscriptive formulation of the opposite default "birds do not fly." A similar criticism can be made of autoepistemic logic. Clearly, statements of normality should not be made true by the falsification of their antecedents. This is simply the paradox of material implication in a default setting. As pointed out by Kraus, Lehmann and Magidor (1990), the traditional default systems do not represent conditional defaults in a compelling manner.

Some notion of normality seems central to the representation of defaults. We would like to read the flying bird default as something like "Birds normally fly," or "In the most normal circumstances in which x is a bird, x flies." Recently, several conditional approaches have been proposed that can be interpreted as allowing such a reading of a default rule (Delgrande 1987; Kraus, Lehmann and Magidor 1990; Pearl 1988). Aside from (to varying degrees) some commitment to a normative interpretation,[2] a crucial aspect of all these models is the conditional aspect of the representations. A unary modality for normality (like a unary abnormality predicate) appears inadequate in general.

Suppose we have some such connective, say N, for normality, and read $N\alpha$ as "α is normally the case." Straightforward default assumptions can be encoded in this manner (e.g., "Assume your car is parked where you left it"); but conditional defaults are not susceptible to such an analysis. One obvious encoding of "birds fly" is $N(\texttt{bird} \supset \texttt{fly})$, but again such a statement is made vacuously

[1] We think of ab as a nullary predicate for this propositional exposition.

[2] We use "normative" here to refer to aspects of normality in general, and the aforementioned linguistic conditional in particular. It should not be confused with its traditional "reason-guiding" meaning, though we will have occasion to require both versions. Context should make clear the intended connotation.

true if bird is false in the most normal states of affairs (i.e. $N\neg$bird). We have moved from the paradox of material implication to the paradox of strict implication. This situation will also validate $N(\texttt{bird} \supset \neg\texttt{fly})$. The other obvious representation is something to the effect that "If x is a bird then it normally flies," or $\texttt{bird} \supset N(\texttt{fly})$, but this is subject to the same criticism as autoepistemic and circumscriptive defaults. For this reason, unary modal approaches (for instance, that of Halpern and Rabin (1987)) appear to be of limited applicability for representing normative defaults. A substantive conditional approach, in contrast, can accommodate conditional defaults as well as unary defaults (e.g., "assume α" can be encoded as $\top > \alpha$).

In this chapter, we will present several logics for conditional statements of normality using a conditional connective $\Rightarrow$. Though we have assumed $\Rightarrow$ to represent this conditional connective, we use $\Rightarrow$ in this chapter, since $\Rightarrow$ will be defined slightly differently in the next chapter. $A \Rightarrow B$ will be read as "In the most normal states of affairs in which A holds, B is true as well," or "A normally implies B." We will provide a Kripke-style possible worlds semantics for these conditional logics, but the conditional connective will not be primitive. It is defined in terms of a unary modal operator $\Box$, although, as we've seen, we cannot interpret $\Box\alpha$ as "normally α." In fact, we will show that these *conditional logics of normality* can be defined using either of $\Box$ or $\Rightarrow$ as the basic connective and that these logics are equivalent to standard modal systems extending S4. We describe a number of advantages of the conditional approach to default reasoning, and show that existing conditional logics for defaults are fragments of the logics presented here, and hence standard modal logics. The key results of this chapter are: Theorems 4.2 and 4.3, that show the completeness of our conditional logic CT4 and its equivalence to the modal logic S4; Corollary 4.5, that shows that any extension of CT4 is also equivalent to a modal system; Theorems 4.21 and 4.22, which demonstrate that two existing logics for default reasoning, P and R, are fragments of our conditional logics; and Theorems 4.26 and 4.29, that determine a simple proof procedure for conditional sentences in CT4 and show that ε-semantics can be embedded in CT4.

Before we present our account of conditionals, there are some obvious proposals we can dismiss immediately. The most basic is to use universally-quantified material conditionals for default rules, but this has the obvious drawbacks, mentioned in Chapter 2, of not allowing exceptions and suffering from the paradoxes. Doyle (1983) has argued that the material conditional can be used provided the exceptions to the rule are explicitly listed. For instance, qualifying the material implication $\texttt{bird} \supset \texttt{fly}$ with exceptions may lead us to a (non-default) rule like "Birds that are not penguins, not emus, not dead, ..., fly." This seems to ignore the qualification problem. We've seen that specifying such conditions is not an easy task.

A more appealing proposal is to interpret such statements as making statements about conditional probabilities. If we let "birds fly" mean $P(\texttt{fly}|\texttt{bird}) > c$ for some constant c, the default statement can certainly allow exceptions, for the existence of a nonflying bird need not affect the conditional probability. Given this proposal, one cannot logically infer that a certain bird flies given the default;[3] but the principle of *direct inference* (cf. Carnap (1950), Bacchus (1990)) allows us to assign the subjective probability, or *degree of belief*, c to the proposition fly if we know bird and the conditional probability is c. Together with some rule of acceptance (e.g., believe any proposition with subjective probability greater than c), direct inference can sanction belief in such default conclusions. As well, strengthening of such rules is not typically valid, allowing the desired nonmonotonic behavior. So one could consistently assert that $P(\texttt{fly}|\texttt{bird} \wedge \texttt{penguin}) \ll c$.

Unfortunately, straightforward probability theory seems quite weak in the types of inference it sanctions. Certain rules of inference such as restricted transitivity and cautious monotonicity

[3]Indeed, this would prevent the exception-allowing feature.

are not forthcoming.[4] For instance, consider facts "Most students get an A" and "Most students are male." A plausible conclusion is "Most male students get an A," but it is not a deductively sound inference on the basic probabilistic interpretation (e.g., a logic of majority). For instance, it could be that 51 of 100 students get an A and 51 students are male, yet only two male students get an A (Pearl 1988). However, Bacchus (1990) has proposed a system for probabilistic reasoning that adds to probability theory the capacity for direct inference together with the ability to make assumptions of independence, and thus reason by default. Just as classical logic can be extended, so too can probability theory.

Many have argued that we often do not want to "reason with the numbers," for reasons of efficiency, or because of the unnatural aspect of probabilities in many circumstances, or because they just don't exist (McCarthy and Hayes 1969; Doyle 1983; Loui 1987b; Halpern and Rabin 1987). It makes sense to talk about the probability of drawing a black ball from the proverbial urn, but it is more difficult to assign a conditional probability to "birds fly." Less obvious still is the actual probability to be assigned to a conditional like "If Bill and Ted come to the party, Sue will have a good time." It might be natural to suggest that Sue "probably" will enjoy herself, but it is less so to assign some meaningful number (or range of numbers) to the statement. Furthermore, if one has to assent to or deny such a statement, what conditional probability constitutes an adequate acceptance criterion? See, for example, (Halpern and Rabin 1987) for a brief discussion of this and related issues.

We would like an account of conditional defaults that doesn't require an explicit rule of acceptance or explicit conditional probabilities. This is not to say a probabilistic account is not meaningful or even required in certain circumstances, nor that we should ignore intuitions gleaned from such accounts. In fact, our conditional logics will compare favorably to, say, the system of Bacchus (1990), and to a certain extent can be viewed as a qualitative version of probabilistic inference with a built-in rule of acceptance.

Many conditional logics have just the properties we require of such normative conditionals. For instance, the paradoxes are avoided and strengthening of the antecedent is generally invalid, allowing the desired nonmonotonicity of default rules. Our conditional logic must differ from standard subjunctive logics however. Instead of evaluating the truth of the consequent at the antecedent-worlds most *similar* to the actual world, as determined by some notion of comparative similarity, we will evaluate the consequent at the *most normal* such worlds, based on some abstract ordering of normality.

4.1 The Conditional Logic CT4

Our semantics for *conditional logics of normality* (CLNs) will be based on Kripkean possible worlds structures. Roughly speaking, a conditional $A \Rightarrow B$ will be true if the most *normal* worlds that make A true also satisfy B. A standard conditional semantics would evaluate $A \Rightarrow B$ at a world w by use of a selection function f, mapping pairs consisting of a world and some proposition into sets of worlds. The set $f(w, \|A\|)$ is intended to represent the set of most normal A-worlds (from the perspective of w).[5] Obviously, $A \Rightarrow B$ will be true at w iff B is satisfied at each most normal A-world, that is if $f(w, \|A\|) \subseteq \|B\|$.

Delgrande (1986; 1987) defines a conditional logic of normality N in just such a manner. The selection function semantics is, of course, very general and requires no constraints on various values

[4] See the next section and rules RT and CM.

[5] Recall that $\|A\|$ is the set of worlds satisfying A (or, the proposition denoted by A.

of the selection function. So there need not be any relationship between, say, $f(w, \|A\|)$ and $\|A\|$. This can lead to undesirable behavior; for example, A need not be true in $f(w, \|A\|)$, so A need not normally imply itself. Delgrande places various restrictions on f to enforce reasonable behavior and axiomatizes the resulting semantics.

The approach we take here is different, more congruous with traditional modal semantics. Instead of a selection function that picks out the most normal A-worlds, we will assume all states of affairs are ordered to reflect some measure of normality. With each possible world is associated an implicit rank, and the most normal A-worlds will be those worlds with the lowest rank, or (roughly) minimal in the order. This ordering will be specified as an accessibility relation R on the set of possible worlds W, hence determining a Kripke model.

Delgrande (1986) presents similar intuitions to motivate his semantics for N, but ends up using the selection function semantics instead. As a result, the logic N is not complete with respect to the accessibility semantics, and differs from the logics we present.[6] In a sense, the ordering of normality can be construed as placing a preference on states of affairs. Given this ordering, we prefer to think of the most normal A-worlds (when only A is known) as representing the actual state of affairs. Thus, the preference logics of Shoham (1988) embody similar intuitions. However, Shoham's work determines *preferential consequence relations*, whereby $A \models_{\leq} B$ means B is true at the preferred A-models. Since $\models_{\leq}$ is a consequence relation rather than a connective like our $\Rightarrow$, the expressive power of preferential logics is much weaker in comparison.

Lehmann and his colleagues (Kraus, Lehmann and Magidor 1990; Lehmann 1989; Lehmann and Magidor 1990) have studied *nonmonotonic consequence relations* that are determined by a model theory similar to Shoham's preference semantics and our semantics. They propose a number of such relations and investigate them in proof-theoretic terms as well as semantically. They are intimately related to our CLNs and we explore the connection in detail in Section 4.3.

4.1.1 The Modal Logic S4

The concept of normality we wish to impose on states of affairs will be represented as a binary accessibility relation R on a set of possible worlds W. The interpretation of R is as follows: wRv iff v is at least *as normal as* w. That is, a world sees only those worlds that are considered to be as normal as itself. As usual, we will say v is *more normal* than w just when wRv and not vRw. Often we say v is *less exceptional* than w in this case. Worlds v and w are *equally normal* iff they are mutually accessible (wRv and vRw). This ordering of normality is a measure of the degree to which we would be willing to accept a possible world as representing the actual state of affairs given that this is consistent with our knowledge. If w and v both satisfy our set of beliefs (that is, they are *epistemically possible* states of affairs) and v is more normal than w, then we prefer to consider v as possibly representing the actual state of affairs, and are willing to consider those propositions satisfied by v that are not contained in our belief set to be more likely than those satisfied by w. This ordering will often be context- or application-dependent, a world being considered more normal just when it is less exceptional in relevant respects. We will say a few words about the nature of such orderings on possible worlds in Section 4.1.4.

There are some restrictions that must be placed on R if it is to represent a coherent notion of normality. Evidently, R must be reflexive, for any world is as normal as itself. As well, we require transitivity for R to be considered an ordering at all. We take these to be the minimal requirements of R, and insist on no other conditions, in general. Such accessibility relations are associated with

[6] We return to this point in Section 4.2.2.

the class of Kripke models that characterize the modal logic S4 (or KT4).

We take the language $\mathbf{L}_M$ to be a standard unary modal language formed from a denumerable set $\mathbf{P}$ of propositional variables, together with the connectives $\neg$ (negation), $\supset$ (material implication), and $\Box$ (necessity). All variables are well-formed formulae and $\neg\alpha$, $\alpha \supset \beta$ and $\Box\alpha$ are well-formed just when α and β are. The connectives $\wedge$ (conjunction), $\vee$ (disjunction), $\equiv$ (equivalence), and $\Diamond$ (possibility) are introduced by definition.

(a) $A \wedge B \equiv_{\text{df}} \neg(A \supset \neg B)$

(b) $A \vee B \equiv_{\text{df}} \neg A \supset B$

(c) $A \equiv B \equiv_{\text{df}} (A \supset B) \wedge (B \supset A)$

(d) $\Diamond A \equiv_{\text{df}} \neg\Box\neg A$

The following definitions are standard in the modal logic literature (Hughes and Cresswell 1984; Chellas 1980).

Definition 4.1 An *S4-model* is a triple $M = \langle W, R, \varphi\rangle$, where W is a set (of possible worlds), R is a reflexive, transitive binary relation on W (the accessibility relation), and φ maps $\mathbf{P}$ into 2^W ($\varphi(A)$ is the set of worlds where A holds).

Definition 4.2 Let $M = \langle W, R, \varphi\rangle$ be an S4-model, with $w \in W$. The *satisfaction* of a formula α at w in M (where $M \models_w \alpha$ means α is true at w) is defined inductively as:

1. $M \models_w \alpha$ iff $w \in \varphi(\alpha)$ for atomic sentence α.
2. $M \models_w \neg\alpha$ iff $M \not\models_w \alpha$.
3. $M \models_w \alpha \supset \beta$ iff $M \models_w \beta$ or $M \not\models_w \alpha$.
4. $M \models_w \Box\alpha$ iff $M \models_v \alpha$ for every v such that wRv.

If $M \models_w \alpha$ we say that M *satisfies* α at w.

It is easy to verify that the connectives introduced by definition have the following familiar truth conditions:

(a) $M \models_w \alpha \wedge \beta$ iff $M \models_w \alpha$ and $M \models_w \beta$.

(b) $M \models_w \alpha \vee \beta$ iff $M \models_w \alpha$ or $M \models_w \beta$.

(c) $M \models_w \alpha \equiv \beta$ iff either $M \models_w \alpha$ and $M \models_w \beta$, or $M \models_w \neg\alpha$ and $M \models_w \neg\beta$.

(d) $M \models_w \Diamond\alpha$ iff $M \models_v \alpha$ for some v such that wRv.

The sentence $\Box\alpha$ can be read as "In all worlds at least as normal as the actual, α holds," and $\Diamond\alpha$ as "In some world at least as normal as the actual, α holds." Validity is defined in a straightforward manner.

Definition 4.3 Let $M = \langle W, R, \varphi\rangle$ be an S4-model. A sentence α is *valid on* M (written $M \models \alpha$) iff $M \models_w \alpha$ for each $w \in W$. A sentence α is *S4-valid* (written $\models_{S4} \alpha$) just when $M \models \alpha$ for every S4-model M.

We now provide a standard axiomatization of S4 (see, e.g., (Hughes and Cresswell 1984)).

Definition 4.4 The *modal logic* S4 is the smallest set $S \subseteq \mathbf{L}_M$ such that S contains CPL and the following axioms, and is closed under the following rules of inference:

K $\Box(A \supset B) \supset (\Box A \supset \Box B)$

T $\Box A \supset A$

4 $\Box A \supset \Box\Box A$

Nec From A infer $\Box A$

MP From $A \supset B$ and A infer B

US From A infer A', where A' is a substitution instance of A

Definition 4.5 A sentence α is *provable in* S4 (written $\vdash_{S4} \alpha$) iff $\alpha \in$ S4. α is *derivable* from a set $\Gamma \subseteq \mathbf{L}_M$ (written $\Gamma \vdash_{S4} \alpha$) if there is some finite subset $\{\alpha_1, \ldots, \alpha_n\}$ of Γ such that $\vdash_{S4} (\alpha_1 \wedge \ldots \wedge \alpha_n) \supset \alpha$.

Theorem 4.1 (Hughes and Cresswell 1984) *The system* S4 *is characterized by the class of S4-models; that is,* $\vdash_{S4} \alpha$ *iff* $\models_{S4} \alpha$.

Thus, the logic S4 fully characterizes the class of models based on the minimal intuitions about normality, specifically those models consisting of a reflexive, transitive normality ordering. The question remains, however: how do we define the truth conditions for $A \Rightarrow B$? Roughly, we want B to be true at the most normal A-worlds, so if (using selection function notation) $f(w, \|A\|)$ is the set of A-worlds minimal in ordering R then $A \Rightarrow B$ holds just when $f(w, \|A\|) \subseteq \|B\|$. Unfortunately, nothing about the structure of S4-models ensures the existence of minimal A-worlds. We may have models where every A-world sees another that is more normal. For example, if A is the proposition

"Pete has height greater than seven feet"

and worlds are ranked as more normal (in this context) as a function of Pete's decreasing height, then there is no obvious minimal or most normal A-world, only an infinite sequence of more and more normal worlds approaching the seven foot limit. A selection function semantics cannot adequately represent such a situation (for example, Delgrande's (1987) logic). Even semantics based on a normality ordering cannot be applied in this case where the truth of $A \Rightarrow B$ is defined in terms of most normal A-worlds (e.g., the systems of Kraus, Lehmann and Magidor (1990) and Shoham (1988)), for such semantics must make the *Limit Assumption*, whereby in each model there exists some minimal A-world for any proposition A. Without this assumption any conditional $A \Rightarrow B$ would be made vacuously true whenever minimal worlds failed to exist. In this case, they would assent to the truth of, say,

"If Pete had height greater than seven feet he would be under six feet tall."

We will discuss the Limit Assumption in greater detail in Chapter 6.

Fortunately, we can provide meaningful truth conditions for $A \Rightarrow B$ without assuming the existence of minimal A-worlds. In the case where such worlds do exist, obviously $A \Rightarrow B$ should hold just when B holds at all such worlds, for these are the most normal A-worlds. In contrast, suppose there is some unending chain of more and more normal A-worlds. If some B-world lies in this chain having the property that B holds, whenever A does, at all (still) more normal worlds in

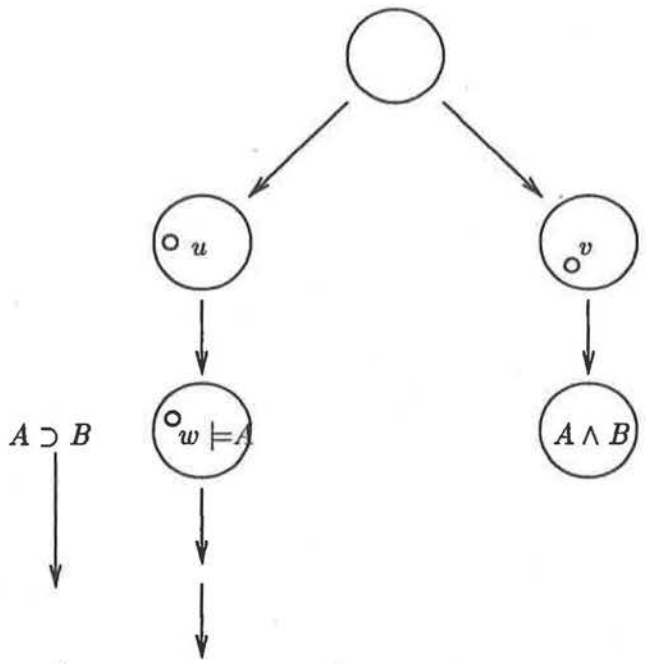

Figure 4.1: A model verifying $A \Rightarrow B$. At each world where A can be "seen" $A \wedge \Box(A \supset B)$ can also be seen. For v there is a set of minimal (most normal) A-worlds verifying B. For u there is not; but there is a point w at which A and B hold and at which $A \supset B$ holds at all lower points.

the chain, then $A \Rightarrow B$ ought to be considered true. Even though no *most* normal A-world exists, B would hold at the "hypothetical limit" of A-worlds in this chain. In other words, if $A \Rightarrow B$ is to be true it must be that whenever some more normal A-world exists, say w, there must exist an even more normal $A \wedge B$-world v (so wRv) such that $A \supset B$ holds at all worlds more normal still (all worlds u such that vRu). Another phrasing of this: for every $A \wedge \neg B$-world there exists a more normal $A \wedge B$-world w such that no $A \wedge \neg B$-world is as normal as w (see Figure 4.1).

We want to express these truth conditions within our modal language $\mathbf{L}_M$. These considerations suggest that for each state of affairs there should be a more normal state where both A and $\Box(A \supset B)$ hold, for this will mean $A \supset B$ holds at all more normal worlds. Thus an initial attempt to define $A \Rightarrow B$ would be

$$\Box\Diamond(A \wedge \Box(A \supset B)).$$

However, this can't be right because it fails to account for worlds where $\Box\neg A$ holds. It may be that $\Box\neg A$ holds because A is some exceptional property, yet we still want to allow conditional $A \Rightarrow B$ to be true. For instance, animals might normally not be birds, yet birds still normally fly. We want to discount these states of affairs as having no influence on the truth of the conditional. Thus we are lead to the following definition.[7]

[7]David Makinson (personal communication) has pointed out that this definition is equivalent (in S4) to the simpler $\Box(A \supset \Diamond(A \wedge \Box(A \supset B)))$. This definition was also discovered independently by Lamarre (1991). We retain our formulation since a simpler definition (not equivalent in S4) for the stronger logic CT4D or S4.3 is just our definition with the outer $\Box$ dropped (see Proposition 4.8 and Chapter 6).

Definition 4.6 The *normative conditional* $A \Rightarrow B$ is defined in $\mathbf{L}_M$ as

$$(A \Rightarrow B) \equiv \Box(\Box\neg A \vee \Diamond(A \wedge \Box(A \supset B))).$$

As this makes some formulae easier to read we introduce the following abbreviation:

$$A \not\Rightarrow B \equiv_{\text{df}} \neg(A \Rightarrow B).$$

Before examining properties of this conditional, we will examine an equivalent formulation of the logic in which the conditional connective is primitive.

4.1.2 Equivalence to S4

We have defined the conditional connective in terms of a unary modal operator, but an interesting question is whether we can use the language in which the conditional is primitive and determine the "same" logic. The fact that we can express the conditional in terms of modal operators contrasts our logic sharply with traditional conditional logics. Lewis (1973a) has suggested that the subjunctive cannot be defined with necessity and standard truth functions. However, conditional logics often express necessity and possibility in terms of the conditional so, in fact, this should be an easier task.

Let $\mathbf{L}_C$ be a conditional language similar to $\mathbf{L}_M$ except that it adds to $\mathbf{L}_{CPL}$ the binary connective $\Rightarrow$ instead of the unary operator $\Box$. We wish to define a conditional logic that ensures the same treatment of $\Rightarrow$ as in the previous section. Though based on the class of S4 (or KT4) models, we denote the new logic CT4 to indicate that the conditional connective is primitive. A CT4-model is just an S4-model, but the truth conditions are modified to capture the new connective. The only new clause is the following

Definition 4.7 Let $M = \langle W, R, \varphi\rangle$ be a CT4-model, with $w \in W$. The truth of a formula $A \Rightarrow B$ at w in M is defined as:

1. $M \models_w \alpha \Rightarrow \beta$ iff for each w_1 such that wRw_1 either
   (a) there is some w_2 such that w_1Rw_2, $M \models_{w_2} \alpha$, and for each w_3 such that w_2Rw_3, $M \not\models_{w_3} \alpha$ or $M \models_{w_3} \beta$; or
   (b) for every w_2 such that w_1Rw_2, $M \not\models_{w_2} \alpha$.

We define the unary modalities in the standard manner (Stalnaker 1968; Lewis 1973a) as

$$\Box\alpha \equiv_{\text{df}} \neg\alpha \Rightarrow \alpha \quad \text{and}$$

$$\Diamond\alpha \equiv_{\text{df}} \neg(\alpha \Rightarrow \neg\alpha).$$

It is easy to verify that these connectives, introduced by definition, have the following familiar truth conditions:

1. $M \models_w \Box\alpha$ iff $M \models_v \alpha$ for each v such that wRv.
2. $M \models_w \Diamond\alpha$ iff $M \models_v \alpha$ for some v such that wRv.

Validity in CT4 is defined in the usual way as truth at all worlds in all models and $\models_{CT4} \alpha$ means α is CT4-valid. In order to axiomatize CT4, it should be evident from the truth conditions of the defined $\Box$ that the axioms of S4 (taking $\Box$ and $\Diamond$ to be the appropriate abbreviations in $\mathbf{L}_C$) are valid. In fact, a complete axiomatization of CT4 requires in addition only the "characteristic" axiom for the conditional connective

C $(A \Rightarrow B) \equiv \Box(\Box\neg A \vee \Diamond(A \wedge \Box(A \supset B)))$.

Completeness of CT4 follows quite readily from the interdefinability of $\Box$ and $\Rightarrow$.[8]

Definition 4.8 The *conditional logic* CT4 is the smallest $S \subseteq \mathbf{L}_C$ such that S contains CPL and the following axioms, and is closed under the following rules of inference:

K $\Box(A \supset B) \supset (\Box A \supset \Box B)$

T $\Box A \supset A$

4 $\Box A \supset \Box\Box A$

C $(A \Rightarrow B) \equiv \Box(\Box\neg A \vee \Diamond(A \wedge \Box(A \supset B)))$

Nec From A infer $\Box A$

MP From $A \supset B$ and A infer B

US From A infer A', where A' is a substitution instance of A

Provability and derivability, denoted by $\vdash_{CT4}$, are defined as usual. Completeness is given by

Theorem 4.2 *The system* CT4 *is characterized by the class of CT4-models; that is,* $\vdash_{CT4} \alpha$ *iff* $\models_{CT4} \alpha$.

It should be clear that CT4 and S4 are definitional variants of each other and are, in a strong sense, equivalent systems. We can freely translate between the languages $\mathbf{L}_C$ and $\mathbf{L}_M$ in a manner that preserves consequence in the logics S4 and CT4. Of course, in itself, this cannot be a claim of equivalence.[9] But the nature of the translations provided by the definitions is such that they induce isomorphisms between the Lindenbaum algebras of the logics CT4 and S4, and each induces the inverse of the other. In other words, if we consider the set of provably equivalent sets of sentences in each logic, the structure of both sets under implication will be identical, with the translation induced by the definitions mapping any element (set of equivalent sentences) to the corresponding element in the other algebra. Thus we are able to show that either CT4 or S4 can be taken as primitive, by demonstrating their essential equivalence.

More precisely:

Definition 4.9 For $\alpha \in \mathbf{L}_C$, the *translation* of α into $\mathbf{L}_M$ (denoted α°) is defined inductively as follows:

1. α, if α is atomic.
2. $\neg\beta^\circ$, if α has the form $\neg\beta$.

[8] Keep in mind that $\Box$ and $\Diamond$ in the axiom system for CT4 are abbreviations in $\mathbf{L}_C$. So, for example, **T** actually stands for $(\neg A \Rightarrow A) \supset A$.

[9] For example, each of intuitionistic and classical propositional logic can be embedded in the other in a like fashion (cf. (Achinger and Jankowski 1986)) yet they are not considered equivalent logics.

3. $\beta^\circ \supset \gamma^\circ$, if α has the form $\beta \supset \gamma$.
4. $\Box(\Box\neg\beta^\circ \vee \Diamond(\beta^\circ \wedge \Box(\beta^\circ \supset \gamma^\circ)))$, if α has the form $\beta \Rightarrow \gamma$.

Definition 4.10 For $\alpha \in \mathbf{L}_M$, the *translation* of α into $\mathbf{L}_C$ (denoted α^*) is defined inductively as follows:

1. α, if α is atomic.
2. $\neg\beta^*$, if α has the form $\neg\beta$.
3. $\beta^* \supset \gamma^*$, if α has the form $\beta \supset \gamma$.
4. $\neg\beta^* \Rightarrow \beta^*$, if α has the form $\Box\beta$.

Theorem 4.3 $\vdash_{CT4} \alpha \equiv (\alpha^\circ)^*$ *and* $\vdash_{S4} \alpha \equiv (\alpha^*)^\circ$. *Also,* $\vdash_{CT4} \alpha \supset \beta$ *iff* $\vdash_{S4} \alpha^\circ \supset \beta^\circ$. *In other words,* CT4 *and* S4 *are equivalent.*

Hence reasoning performed in one logic can be just as easily accommodated in the other. For this reason, we take the modal formulation to be basic and let the connective $\Rightarrow$ be defined within *S4*. However, we continue to refer to the logic as CT4 to emphasize our interest in the conditional aspect of the logic, and it should be kept in mind that this is just S4.

4.1.3 Properties of CT4

A number of derived rules of inference and theorems of CT4 suggest it is an appropriate logic for representing and reasoning with exception-allowing generalizations and statements of normality and typicality. For instance, the following theorems are expected of a logic of normality.

Proposition 4.4 *The following are valid in* CT4.[10]

ID $A \Rightarrow A$

LLE $\Box(A \equiv B) \supset ((A \Rightarrow C) \equiv (B \Rightarrow C))$

And $((A \Rightarrow B) \wedge (A \Rightarrow C)) \supset (A \Rightarrow B \wedge C)$

RT $(A \Rightarrow B) \supset ((A \wedge B \Rightarrow C) \supset (A \Rightarrow C))$

Or $((A \Rightarrow C) \wedge (B \Rightarrow C)) \supset (A \vee B \Rightarrow C)$

RCM $\Box(B \supset C) \supset ((A \Rightarrow B) \supset (A \Rightarrow C))$

CM $((A \Rightarrow B) \wedge (A \Rightarrow C)) \supset (A \wedge B \Rightarrow C)$

The theorem **ID** asserts that A normally implies A while **And** and **RCM** ensure that the set of facts normally implied by A is closed under logical consequence (and, even stronger, strict implication).

We have seen that most conditional logics will not validate the the following theorems. CT4 is no exception, as the following are not valid (using the names of inference rules for the corresponding sentences).

[10] Many of the names of rules and theorems are taken or adapted from (Nute 1980; Delgrande 1987; Lehmann 1989).

Strengthening $(A \Rightarrow B) \supset (A \wedge C \Rightarrow B)$

Transitivity $(A \Rightarrow B) \wedge (B \Rightarrow C) \supset (A \Rightarrow C)$

Modus Ponens $A \wedge (A \Rightarrow B) \supset B$

It is interesting to note, however, that weaker versions of these rules hold. The following theorems of CT4 are the "normal" counterparts of these rules.

Weak Strengthening $(A \Rightarrow B) \Rightarrow (A \wedge C \Rightarrow B)$

Weak Transitivity $(A \Rightarrow B) \wedge (B \Rightarrow C) \Rightarrow (A \Rightarrow C)$

Weak Modus Ponens $A \wedge (A \Rightarrow B) \Rightarrow B$

Therefore, while one cannot infer B from A and $A \Rightarrow B$, it is reasonable to conclude that normally B would hold. This suggests that default reasoning can be modeled as the process of asking what normally follows from a knowledge base that includes conditional statements, appealing to Weak Modus Ponens. In Chapter 5 we will see that this proposal is inadequate.

Full transitivity is not a desirable rule of inference in general. For instance, if we know

$$\texttt{penguin} \Rightarrow \texttt{bird} \quad \text{and} \quad \texttt{bird} \Rightarrow \texttt{fly}$$

the conclusion $\texttt{penguin} \Rightarrow \texttt{fly}$ should not be forthcoming. A restricted form of transitivity is, however, ensured by **RT**. If

$$\texttt{canary} \Rightarrow \texttt{bird} \quad \text{and} \quad \texttt{canary} \wedge \texttt{bird} \Rightarrow \texttt{fly}$$

hold, the conclusion $\texttt{canary} \Rightarrow \texttt{fly}$ is valid.

Strengthening of the antecedent is not valid in CT4 either. It is consistent to assert

$$\{A \Rightarrow D\ , A \wedge B \Rightarrow \neg D\ , A \wedge B \wedge C \Rightarrow D\ , \cdots\}.$$

For example, we can represent the facts contained in the following account.

> If the U.S.A. threw its weapons into the sea tomorrow, there would be war; but if the U.S.A. and the other nuclear powers all threw their weapons into the sea tomorrow there would be peace; but if they did so without sufficient precautions against polluting the world's fisheries there would be war; but if, after doing so, they immediately offered generous reparations for the pollution there would be peace .... (Lewis 1973a, p.10)

Of course, the conditional is exception-allowing for conditional modus ponens is not valid. We can assert that birds normally fly while Tweety the bird does not. The conditional law of excluded middle

CEM $(A \Rightarrow B) \vee (A \Rightarrow \neg B)$

is not valid in our logic. In contrast with Stalnaker's (1968) conditional semantics, a proposition A need not normally indicate an attribute or its negation. This is because our models do not insist that most normal A-worlds be unique. CT4 is more closely related to Lewis's (1973a) conditional logic VC than Stalnaker's logic C2.[11]

[11] This is so regarding the Limit Assumption as well, which is a property of Stalnaker's but not Lewis's semantics.

The definition of the normative conditional allows the derivation of $A \Rightarrow B$ from $\Box\neg A$. For instance, if it is known that it never snows it is consistent to assert "We normally go skiing when it snows" (as well as "We normally do not go skiing when it snows"). This seems to contradict linguistic usage of normative statements, where the conditional "When it snows, we ski" generally conveys the possibility of snow, or the existence of occasions of snowing. We view this as a matter of the conversational *implicature* of normatives, relegating it to the pragmatics of linguistic usage rather than the semantics of such statements. The normative $A \Rightarrow B$ pragmatically *implicates* the possibility of its antecedent $\Diamond A$. The negation of the implicated statement is logically stronger than the conditional (that is, $\neg\Box A$ semantically *implies* $A \Rightarrow B$), so if it were known, it would be a more appropriate or informative assertion. This can be viewed as a *scalar Quantity implicature*, whereby statements of certain types implicate the negation of semantically stronger counterparts (Levinson 1983). For example, "possibly A" implicates "not necessarily A," while "some B's have property P" implicates "not all B's have P."

Another implicature of $A \Rightarrow B$ is $\neg\Box(A \supset B)$, which semantically entails the conditional. It is (semantically) consistent, but (pragmatically) inappropriate, to say that "penguins are normally birds" when in fact penguins are necessarily (or by definition) birds.

Other properties of the conditional are illustrated by the following examples.

Example 4.1 Suppose we know both

$$\texttt{obsD1} \not\Rightarrow \texttt{faultyC1} \quad \text{and} \quad \texttt{obsD1} \not\Rightarrow \neg\texttt{faultyC1},$$

meaning if device D1 fails in some specified way (observation obsD1) that the diagnosis should favor neither the inclusion nor the exclusion of component C1 among the set of faulty components. This situation would probably lead to further testing. However, if $\texttt{obsD1} \Rightarrow \texttt{faultyC1}$ holds, then it cannot also be the case that $\texttt{obsD1} \Rightarrow \neg\texttt{faultyC1}$.

■

Example 4.2 CT4 can deal with simple cases of inheritance with exceptions (see, e.g. (Touretzky 1986)). Let KB contain

$$\texttt{bird} \Rightarrow \texttt{fly}, \quad \texttt{penguin} \Rightarrow \neg\texttt{fly},$$

$$\texttt{penguin} \Rightarrow \texttt{bird}, \quad \texttt{penguin} \Rightarrow \texttt{wings}.$$

If something is both a penguin and a bird inheritance of properties is derived from the more specific superclass penguin in the case of conflict. Hence, by CM both

$$\texttt{penguin} \wedge \texttt{bird} \Rightarrow \neg\texttt{fly} \quad \text{and}$$

$$\texttt{penguin} \wedge \texttt{bird} \Rightarrow \texttt{wings}$$

are derivable. Now suppose we want to assert that penguins are (or must be) birds, rather than that they are typically birds. This *strict* information is captured by replacing the third premise with $\Box(\texttt{penguin} \supset \texttt{bird})$. The same conclusions are still derivable from KB.

Adding the fact that emperor penguins are penguins

$$\Box(\texttt{emp-penguin} \supset \texttt{penguin})$$

allows the derivation of $\texttt{emp-penguin} \Rightarrow \neg\texttt{fly}$, even though we can also derive $\Box(\texttt{emp-penguin} \supset \texttt{bird})$. This holds even when the relationship between penguins and birds is not strict. Fur-

thermore, we can derive

$$\texttt{bird} \Rightarrow \neg\texttt{penguin}.$$

Since penguins are exceptional birds, birds should typically not be penguins.

■

This example illustrates the semantics underlying simple inheritance with exceptions (Touretzky 1986; Touretzky, Horty and Thomason 1987; Horty, Thomason and Touretzky 1987). Since penguin is a specific subclass of bird, we conclude that emperor penguins do not fly even though they are both birds and penguins. In the case of conflict, properties are inherited from the more specific (the exceptional) superclass in this simple case. This can be explained as follows (using the obvious abbreviations): if $P \wedge B \not\Rightarrow \neg F$ is true, some most normal $P \wedge B$-worlds must satisfy F. Since $P \Rightarrow \neg F$ there must also be a $P \wedge \neg F$-world more normal than any such $P \wedge B$-worlds. As this $P \wedge \neg F$-world must satisfy $\neg B$, this contradicts the truth of $P \Rightarrow B$. So $P \wedge B \Rightarrow \neg F$ must be true.

In this example, we also have that birds must typically not be penguins. However, suppose all we know is that penguins are *not typically fliers* (rather than *typically not fliers*). In CT4 we can make such a distinction since we can negate defaults. This ability is useful when we want to assert an exception to a default without committing to the opposite conclusion, a task not achievable in most default systems (at least not in such a straightforward fashion), for instance, in Pearl's (1988) ε-semantics or even inheritance networks. Thus we can state that penguins are exceptional birds, in that they can't be inferred to fly, without requiring that they typically do not fly. This is the distinction between

$$\neg(\texttt{penguin} \Rightarrow \texttt{fly}) \quad \text{and}$$

$$\texttt{penguin} \Rightarrow \neg\texttt{fly}.$$

We call this first sentence the *weak negation* and the second the *strong negation* of the conditional. Assuming the antecedent is possible, the strong negation entails the weak negation. We can think of the weak negation as an undercutting defeater for the conditional default $\texttt{bird} \Rightarrow \texttt{fly}$, while the strong negation is a rebutting defeater (see Section 2.2.1).

Intuitively, if we assert the weak negation of "penguins fly" we'd still like to conclude that birds are typically not penguins, since penguins are still exceptional birds, if only in the weaker sense. Unfortunately, such an inference is not valid in CT4, for the appropriate inference rule, *Rational Monotonicity*, is not valid. We can extend CT4 to include such a rule (see the next section).

More complex inheritance, including straightforward chaining is not sanctioned (logically) within CT4 (for instance, simple transitivity we've seen is invalid and logically undesirable). To remedy this, approaches to inheritance based on normality augment logical inference with extra-logical principles. In (Boutilier 1989; Boutilier 1991d) such a model is based on an extension of CT4, discussed in the next section, to which is added a preference relation on interpretations corresponding to some intuitions about inheritance. Delgrande (1990) provides a similar treatment based on the logic N. We will examine more general approaches to default reasoning using our conditional logics in the next chapter.

Example 4.3 In (Goldszmidt and Pearl 1989) it is argued that inconsistency of default rules has been ignored for the most part. For example, in default logic or circumscription the obvious encodings of the following three sentences is consistent: "All birds fly"; "Typically penguins are birds"; "Typically penguins do not fly." Assuming that some penguins exist, intuitively these sentences form an inconsistent set: there is no state of affairs that can satisfy the set.

In CT4, the expected contradiction is derivable, for the set

$$\{\Box(\texttt{bird} \supset \texttt{fly}), \texttt{penguin} \Rightarrow \texttt{bird}, \texttt{penguin} \Rightarrow \neg\texttt{fly}\}$$

is inconsistent (given $\Diamond\texttt{penguin}$).[12] This is similar to the treatment of such rules in (Goldszmidt and Pearl 1989), but the correspondence is not coincidental (see Section 4.3.2).

■

4.1.4 Normality Orderings

The use of S4-models and the connective $\Rightarrow$ to represent normality presumes that possible worlds can be ranked according to their degree of normality. But what does it mean for one world to be *more normal* than another? Why is it that a world where all birds fly is less exceptional than one where some birds (or all birds) do not fly? Surely there is no *inherent* property of possible worlds that allows them to be judged to be more or less normal. Does this make the notion of comparing worlds in this way ill-conceived?

We do claim that such an ordering is meaningful. However, we do not require that there exist something intrinsic about the concept of a possible world, or counterfactual situation, that makes one world more normal than another. These rankings are purely subjective and the (space of) rankings deemed plausible by a (supposedly) rational agent will typically be determined by empirical data. Possible worlds are used in this context in much the same manner that they are used to represent epistemic states. To criticize possible worlds accounts of knowledge because no intrinsic property of possible worlds determines which worlds are epistemically possible is misguided. Of course there is no *a priori* property of worlds that determines that, say, world w_1 is "believable" while w_2 is not. This is a subjective matter. World w_1 is considered epistemically possible *by a particular agent* if w_1 is consistent with that agent's beliefs. These beliefs should be grounded by an agent's experience, so it is entirely possible that a different agent views w_1 as epistemically impossible.

In analogous fashion, it is perfectly reasonable that one agent considers w_1 (where all birds fly) to be more normal than w_2 (where no birds fly), while another agent adopts the inverse relation. The first agent believes that birds normally fly and its ordering reflects that, while the second agent believes that birds normally do not fly. Why does the first agent believe what it does? Presumably, this belief was accepted for the same reason as most (justified) beliefs, because of empirical evidence. In the agent's experience, some statistically significant large proportion of birds were observed to fly (or perhaps it read this fact in a book, etc.). As we discussed at the beginning of this chapter, it might not be feasible to keep track of probabilities, so an agent abstracts the details of its experience, encoding the distillation as "birds normally fly."

Why do we not simply say "birds probably fly" or "birds fly with some high probability"? Sentences of this type carry certain implicature not intended by normality or typicality, especially when dealing with individuals rather than classes. If we say "The probability that ABC Dry Cleaners gets my shirt clean is .9" when we've just gotten a shirt back, we convey the false impression that we do not know whether the shirt is clean. However, once we have the shirt in our possession we either know it is clean (and the probability is 1) or we know it is not clean (and the probability is 0).[13] We cannot take the probability of the known outcome of an event to be other than 0 or 1. Of

[12] See the next section regarding the purpose of the use of $\Box$ in front of $\texttt{bird} \supset \texttt{fly}$.

[13] We ignore degrees of cleanliness, not examining the shirt, etc.

course, we can *randomize* over all similar events (see the next subsection) and say "Although they did not clean my shirt this time, the probability of of getting a clean shirt (over the space of similar — including counterfactual — cleaning situations) is .9." In order to arrive at any probability other than .5 as such a *degree of belief*, possible worlds must be weighted as either more or less probable. Of course, in a case like this, it seems more natural to say "They normally get my shirts clean." Adopting this perspective relieves us from the task of assigning actual probabilities to situations, requiring only relative rankings of normality.

By using rankings in this way, we encode our *expectations* about the world, information that enables reliable performance in (implicit of explicit) predictive tasks. We do not need to rely on the numbers, or randomize over counterfactual situations. However, we can impose a probabilistic interpretation on our ordering of normality. This is discussed in Sections 4.3.2, 5.4 and 7.2.3.

An important question we do not address here is that of learning or adopting defaults. When is it the case that an agent is justified in adopting belief in a sentence "If A then normally B" based on empirical evidence? Surely some statistical knowledge plays a role in justification, but it need not be decisive. That is, probabilities need not directly constrain the acceptance of conditionals. The second agent above might have decided that birds normally do not fly based on the observation of one bird that did not fly. The evidence is not highly significant statistically and one may be well-advised to abstain from belief in the conditional; but acceptance need not imply a degree of irrationality. Perhaps an agent is forced into action due to circumstances beyond its control, and must accept or reject this fact.[14] In any case, the representation of such defaults is crucial, even if a compelling theory of justification is lacking. Certainly much of the knowledge we wish to impart of our knowledge bases is defeasible, and the only justification for acceptance required by a database is the fact that it has been "told" something. While a fully rational agent will need to learn its own default rules (just as it will need to discover its own beliefs), this does not obviate the need of a theory of inference for such rules.

These are difficult questions, and will not be discussed further here, but it should be clear that they can be addressed within the framework of normality orderings. It is important to notice however that such orderings are not dependent on *a priori* relations among possible worlds. The set of *most normal* worlds *for a particular agent* are just those that violate none of its expectations. If an agent has no expectations (a practical impossibility certainly), all worlds are equally normal. If an agent has the (sole) expectation that birds normally fly then any world where some bird does not fly is less normal than worlds where all birds fly. More expectations impose more structure on the orderings an agent considers plausible. This structure is discussed in detail in Section 5.4 where we present Pearl's (1988) notion of Z-ranking and compare it to our interpretation of normality. In this way we can say precisely what it means for one world to be more normal than another relative to a given set of expectations or default rules.

4.1.5 Background Versus Evidence

Example 4.2 also illustrates a distinction between *background knowledge* and *evidence* provided by the current situation. This dichotomy has been deemed essential (either implicitly or explicitly) by a number of researchers (Bacchus 1990; Poole 1988; Delgrande 1988; Pearl 1988; Kraus, Lehmann and Magidor 1990; Geffner 1989), and is especially important in a probabilistic framework. Background knowledge is "generic" information, either in the form of default rules or constraints on permissible probability distributions, etc., while evidence includes information relevant to the

[14] To take Reiter's example, maybe an agent is lost in the jungle and must design and build a bird trap to survive. It has seen one instance of a bird, which cannot fly.

actual situation about which we are reasoning. Accounts of the distinction in the literature are often vague or impressionistic and provide little guidance as to what should count as background and what as evidence. For instance, in (Geffner 1989) a context for a particular reasoning situation is a pair $\langle K, E\rangle$ where K is background and E is a set of (for our purposes, propositional) sentences counting as the evidence constraining the situation at hand. Unfortunately, K itself is a pair $\langle L, D\rangle$ where D is a set of defaults and L is also a set of sentences. What exactly distinguishes L from E is not made clear, although intuitively facts about the actual world, such as "Tweety is a bird" count as evidence while facts such as "Penguins are birds" count as background.

Delgrande (1988) makes the distinction somewhat more sharply. His default theories are pairs $\langle D, C\rangle$ where C is a set of *contingent* sentences, constraining how the world is, and D is a set of *necessary* and *default statements* about how the world must or could be. C can contain only propositional sentences and elements of D must have the form $\alpha \Rightarrow \beta$ or $\neg(\alpha \Rightarrow \beta)$ (which includes sentences of the form $\Box\alpha$ and $\Diamond\alpha$). Here it is made clear that background information consists of sentences expressing *intensional information* (loosely, facts whose truth depends on states of affairs other than the actual) while evidence does not. If some fact is to be considered true no matter how the current situation might have been (or will be), then it must be expressed intensionally.

In a probabilistic system, like that of Bacchus (1990) or Pearl (1988), the distinction is similar. Background is information about (subjective) probabilities that can include necessary information (having probability 1), and evidence consists of a set of propositional sentences, a knowledge base KB on which we condition to derive the probability of new information (say some prediction we wish to make).

Example 4.2 assumes default information about the normal state of affairs is background, or intensional, as it is expressed using the conditional connective. For the proper conclusions to be reached when we know penguins are (in the strict sense) birds, this must be considered background as well, and must be expressed intensionally as $\Box(\texttt{penguin} \supset \texttt{bird})$. Indeed, if we included $\texttt{penguin} \supset \texttt{bird}$ in KB instead, we could not have inferred that penguin-birds normally do not fly, for such a sentence constrains only the actual state of affairs, and it could be that penguins are not generally birds. The difficult part of distinguishing background from evidence is deciding what "real world knowledge" counts as which, and we do not address this issue here. We claim that translating such knowledge into logical sentences is an easier task when this distinction is made logically precise, and that intensional versus extensional is a more precise discrimination than one afforded by nebulous concepts like background and evidence.[15]

The problems that arise when background and evidential "components" of a knowledge base are not distinguished are discussed at length by Poole (1991), for instance, the problem of inferring specificity of defaults. He suggests that a knowledge base be divided into two components representing background and evidence, that these be treated distinctly and kept apart, and even that background information need not be considered in a lot of reasoning. We will see that this dichotomy, in the extreme form presented, can lead to problems.

One aspect of this distinction obviated by our treatment is the division of KB into these two components before reasoning is done. A KB will typically consist of a set of facts, some of which are contingent and some of which are to count as background. To ask which sentences α (normally) followed from KB in the previous example, when $\texttt{penguin} \wedge \texttt{bird}$ formed our specific evidence, we asked if

$$\texttt{penguin} \wedge \texttt{bird} \Rightarrow \alpha$$

[15]The distinction is also clearer in probabilistic systems where background is given by probabilistic information or distributions.

was derivable from KB (the background). More generally, using Delgrande's terminology, we ask

$$D \vdash_{CT4} C \Rightarrow \alpha.$$

While this phrasing makes just the aforementioned distinction, we could equivalently ask

$$\vdash_{CT4} D \cup C \Rightarrow \alpha,$$

putting background and evidence on logically equal footing, and thus eliminating this discrimination altogether.[16] The knowledge of intensional and extensional facts need not be distinguished (other than having to be expressed appropriately).

In Delgrande's logic this mixture of components is not permitted since the logic N does not deal properly with nested conditionals (see Section 4.2.2), so $D \cup C \Rightarrow \alpha$ will not be valid in most natural cases, though it is well-formed. In the conditional logics of Pearl (1988) and Kraus, Lehmann and Magidor (1990) such nesting cannot be expressed. A similar criticism can be levied against the probabilistic system of Bacchus (1990). In general, we would like to ask for the conditional probability $P(\alpha|D \cup C)$. However, this is ill-formed when D contains statements of (subjective) conditional probability (i.e., defaults).[17]

Evidently, there is a certain methodological simplicity in blurring the distinction between background and evidence at this logical level. However, aside from any putative advantages of this homogeneity, there is a clear increase in expressive power afforded by such a conflation, for it allows one to achieve a degree of interaction between the two components otherwise unobtainable. Suppose one wishes to express the fact that if A and B hold then As are normally Bs. For instance, we might be eager to draw some conclusion about the class of As based on a particular exemplar.[18]

Example 4.4 Consider an agent that takes some shirts to ABC Dry Cleaning for the first time (I suppose we AI types expect this agent to be a robot?). If the shirts come back poorly cleaned, we might want the robot to infer that cleaners ABC normally do a poor job. We can represent this as

$$(\texttt{takeshirts(x)} \wedge \texttt{notclean(x)}) \supset (\texttt{takeshirts(x)} \Rightarrow \texttt{notclean(x)})$$

and capture some interaction between evidence and background (as well as a simple case of inductive learning).

■

It is precisely this type of interaction that cannot be accounted for in systems where background

[16] We assume D and C are finite sets and use them to represent the conjunction of their elements. We show that this query mechanism is too weak for default reasoning in most situations in the next chapter.

[17] In fact, this is a somewhat unfair criticism of Bacchus's system. While subjective probabilities cannot be conditioned on, the novel aspect of his system is that it allows degrees of belief to be inferred (by default) from objective probabilities. These objective probabilities are drawn from distributions in the "actual world" and are represented as objective facts in the evidential part of KB. Hence, one can condition on *objective* conditional probabilities. In the example below, indeed, his system could take an objective conditional probability of 1, infer a degree of belief 1 in the statement $A \supset B$ and derive the default $A \Rightarrow B$. In fact, the default would necessarily be strict, $\Box(A \supset B)$, so we still allow an expressiveness not possible in his system, where a "subjective probability" $A \Rightarrow B$ need not correspond to the "objective probability" $\Box(A \supset B)$. But the ability to link evidence and probabilities logically (see below) certainly exists in Bacchus's system.

[18] In fact, it appears people are notorious for this statistically unreliable form of inference. For instance, people seem to ignore sample size in assessing the reliability of their statistical judgements, even when the relevance of sample size is emphasized (Tversky and Kahneman 1974).

and evidence are separated in a *KB*.

4.2 Extensions of CT4

While CT4 captures many aspects of normal implication, there are some theorems, intuitively valid in many circumstances, that fail to hold. For instance, the rule of *Rational Monotonicity* (Lehmann 1989) discussed above (Example 4.2), and the related axiom **CV** (Delgrande 1987), are not valid in CT4.

RM From $A \Rightarrow C$ and $A \wedge B \not\Rightarrow C$, infer $A \Rightarrow \neg B$

CV $(A \not\Rightarrow B) \supset ((A \Rightarrow C) \supset (A \wedge \neg B \Rightarrow C))$

RM says, for instance, that if birds normally fly and penguins are exceptional birds (that do not typically fly), then it must be the case that birds are normally not penguins. An instance of **CV** says that if birds are not normally penguins and birds normally fly, then birds that are not penguins also typically fly.

We would like to extend the logic so that these and other reasonable properties of normal implication are validated. For this reason, we will allow any system in the language $\mathbf{L}_C$ that extends CT4 to be called a conditional logic of normality.

Definition 4.11 A *conditional logic of normality* (CLN) is any system $S \subseteq \mathbf{L}_C$ closed under the inference rules **Nec**, **MP** and **US**, such that CT4 $\subseteq S$.

The following obtains immediately given Theorem 4.3. This result has tremendous impact, as it indicates that for any logic in the class of CLNs, an equivalent modal formulation can be used. Thus results concerning existing modal logics can be appropriated directly for the corresponding CLNs.

Corollary 4.5 *Any modal system that extends* S4 *is equivalent to some CLN, and, conversely, any CLN is equivalent to some modal system that extends* S4.

We take the modal logic S5 to be the upper limit of useful CLNs. The corresponding conditional logic, denoted CT45, consists of CT4 and the additional axiom

5 $\Diamond A \supset \Box\Diamond A$.

This logic is characterized by the class of models in which R is an equivalence relation on W. Thus every world is as normal as every other.[19] For this reason, normal implication reduces to strict implication in this system. That is,

$$\vdash_{CT45} (\alpha \Rightarrow \beta) \equiv \Box(\alpha \supset \beta).$$

Any exception-allowing capacity or nonmonotonicity permitted by $\Rightarrow$ disappears, for strict implication has no such ability; it is monotonic as

$$\Box(\alpha \supset \beta) \supset \Box(\alpha \wedge \gamma \supset \beta)$$

[19] More exactly, each world is as normal as those it can see. Restricting attention to *cohesive* models (see Chapter 6), this means all worlds are mutually accessible.

is valid in any normal modal system. In fact, in CT45 normal implication reduces to straightforward logical implication, relative to the states of affairs contained in a model. While its usefulness as a CLN might be limited, CT45 does demonstrate that logical implication can be viewed as a form of normal implication.

Weaker extensions of CT4 appear more interesting. Let CT4G be the extension of CT4 containing the axiom

G $\Diamond\Box A \supset \Box\Diamond A$.

This logic corresponds to the modal logic S4.2 and is based on the class of reflexive, transitive, *convergent* models[20] (Hughes and Cresswell 1984). This logic fails to validate **RM** but does include a weaker version of it:

WRM $(A \Rightarrow C) \wedge \neg(A \wedge B \Rightarrow C) \supset (A \Rightarrow \neg B) \vee (\top \Rightarrow \neg A)$.

So, for instance, if we know birds normally fly and penguins (which are birds) do not normally fly, then we can cannot conclude that birds are normally not penguins, but we can infer the disjunction of this with the fact that "things" are normally not birds. It is hard to motivate the property of convergence for a normality ordering, but it might be useful in some applications.

4.2.1 The Logic CT4D

Another compelling condition we may impose on an accessibility relation is that of *connectedness*: if wRv and wRu then either uRv or vRu. This condition requires that all states of affairs (from the perspective of any particular world) be comparable with respect to normality. If neither of u nor v are more normal than the other (but both are more normal than some w) then they must be equally normal, or mutually accessible. They cannot be incomparable. The modal logic associated with connected orderings is KT4D, or S4.3 (Hughes and Cresswell 1984).

Definition 4.12 $M = \langle W, R, \varphi\rangle$ is a *CT4D-model* (or an S4.3-model) iff M is a CT4-model and R is a *connected* relation.

Definition 4.13 The conditional logic CT4D is the smallest CLN including axiom

D $\Box(\Box A \supset B) \vee \Box(\Box B \supset A)$.

Theorem 4.6 (Boutilier 1988) *The system* CT4D *is characterized by the class of CT4D-models; in other words,* $\models_{CT4D} \alpha$ *iff* $\vdash_{CT4D} \alpha$.[21]

Not surprisingly, CT4D and S4.3 are equivalent systems.

Theorem 4.7 (Boutilier 1988) CT4D *and* S4.3 *are equivalent.*

The fact that CT4D-models are connected means that the set of worlds accessible to a particular world w (the *submodel generated by* w) determines a totally-connected set. From the "point of view" of w, every conceivable world is ranked and comparable to each other world. In other words, unlike in CT4-models, there is only one "path" of normality; there are no competing alternatives to consider. This should allow us to simplify our definition of the conditional $\Rightarrow$. Indeed, we

[20] R is convergent if wRw_1 and wRw_2 implies there exists a w_3 such that w_1Rw_3 and w_2Rw_3.

[21] In (Boutilier 1988), CT4D is called E.

can simply remove the outer modal box in our earlier definition since its function was to ensure consideration of the (hypothetical) most normal antecedent world in *each* chain of increasingly more normal worlds. In CT4D (or S4.3) models, there is only one chain. For $A \Rightarrow B$ to be true either A is false at every world in this chain, or at the "most" normal A-worlds in this chain $\Box(A \supset B)$ holds. This leads us to the following reformulated definition:

$$A \Rightarrow B \equiv_{\text{df}} \Box\neg A \vee \Diamond(A \wedge \Box(A \supset B)).$$

Proposition 4.8 $\vdash_{CT4D} \Box(\Box\neg A \vee \Diamond(A \wedge \Box(A \supset B))) \equiv \Box\neg A \vee \Diamond(A \wedge \Box(A \supset B))$.

When dealing with CT4D, we will use the simpler definition of $\Rightarrow$ afforded by this sentence.

CT4D appears to be a very natural logic for representing and reasoning with default statements. The aforementioned theorems of CT4 are obviously theorems of CT4D, and the desirable nontheorems, such as conditional modus ponens, transitivity and strengthening, all remain invalid. Furthermore, CT4D validates both **RM** and **CV**.

Proposition 4.9 *The following are valid in* CT4D.

RM $((A \Rightarrow C) \wedge (A \wedge B \not\Rightarrow C)) \supset (A \Rightarrow \neg B)$

CV $(A \not\Rightarrow B) \supset ((A \Rightarrow C) \supset (A \wedge \neg B \Rightarrow C))$

Example 4.5 Suppose some diagnostic system needs to reason about the behavior of a robotic manipulator. We can assert that the robot usually grabs parts with its right arm with the sentence

$$\texttt{holding(x)} \Rightarrow \texttt{holdingright(x)}.$$

If two distinct parts (say, "widgets") are normally handled together for assembly and the robot can grab each in either hand, we can state the exceptions as

$$\neg(\texttt{holding(x)} \wedge \texttt{widget(x)} \Rightarrow \texttt{holdingright(x)}) \quad \text{and}$$

$$\neg(\texttt{holding(x)} \wedge \texttt{widget(x)} \Rightarrow \texttt{holdingleft(x)}).$$

Of course, we must constrain the relationship between the right and left arms as the background or intensional (rather than evidential) sentence

$$\Box(\texttt{holding(x)} \supset (\texttt{holdingright(x)} \equiv \neg\texttt{holdingleft(x)})).$$

In CT4 we cannot infer that widgets are atypical parts: if KB consists of the above sentences together with (evidence) `holding(P)`, then

$$\not\vdash_{CT4} KB \Rightarrow \neg\texttt{widget(P)}.$$

But in CT4D this fact is derivable since

$$\vdash_{CT4D} KB \Rightarrow \neg\texttt{widget(P)}.$$

■

4.2.2 The Conditional Logic N

As we mentioned at the outset, Delgrande's (1986; 1987) logic N was motivated using a reflexive, transitive, connected relational semantics like that of CT4D, but is characterized by a selection function semantics, which we present here.

Definition 4.14 (Delgrande 1987) An *N-model* is any triple $M = \langle W, f, \varphi \rangle$ where W is a set (of possible worlds), φ (the valuation) maps $\mathbf{P}$ into 2^W, and (the selection function) f maps $W \times 2^W$ into 2^W such that the following conditions hold:

1. $f(w, \|A\|) \subseteq \|A\|$.
2. If $f(w, \|A\|) \subseteq \|B\|$, then $f(w, \|A\|) \subseteq f(w, \|A \wedge B\|)$.
3. If $f(w, \|A\|) \not\subseteq \|B\|$, then $f(w, \|A \wedge \neg B\|) \subseteq f(w, \|A\|)$.
4. $f(w, \|A \vee B\|) \subseteq f(w, \|A\|) \cup f(w, \|B\|)$.

The truth conditions for sentences in N-models are straightforward with the conditional case given by

$$M \models_w A \Rightarrow B \quad \text{iff} \quad f(w, \|A\|) \subseteq \|B\|.$$

An axiomatization of N can be found in (Delgrande 1987).

The logic CT4D was first studied in (Boutilier 1988) as an extension of N and there it was shown that N and CT4D do not coincide precisely, for they differ on the validity of various nested conditional sentences. In particular, certain natural nested theorems of CT4D do not hold in N, for instance

$$A \wedge (A \Rightarrow B) \Rightarrow B.$$

However, it was shown that all N theorems are derivable in CT4D.

Theorem 4.10 (Boutilier 1988) N $\subset$ CT4D.

There it was conjectured that on only these nested sentences does theoremhood in the two logics differ, and that while N fails to provide a compelling account of nested conditionals, the "flat" fragments on N and CT4D are identical. We can show this is actually false. In particular, the rule of *Cautious Monotonicity*, **CM**, does not hold in N, as indicated by the following proposition.[22]

Proposition 4.11 $(A \Rightarrow B) \wedge (A \Rightarrow C) \supset (A \wedge B \Rightarrow C) \notin$ N.

Let the language $\mathbf{L}_C^-$ be the language $\mathbf{L}_C$ restricted so that no occurrence of a conditional lies within the scope of another. The flat logics CT4D– and N– are just those systems consisting of the theorems of CT4D and N, respectively, that are contained in the impoverished language $\mathbf{L}_C^-$. As an easy consequence of Proposition 4.11 we have that the flat fragment of N is *properly* contained in that of CT4D.

Theorem 4.12 N– $\subset$ CT4D–.

[22] This fact is also shown in (Lehmann and Magidor 1990) by considering the relational semantics for N. We give a proof in terms of the selection semantics.

4.3 Relationship to Other Conditionals

Several other conditional approaches to the representation of default statements have been proposed in the literature. We've briefly looked at the logics of (Bacchus 1990) and (Delgrande 1987), which share some underlying motivation with our approach. Some models can be viewed as argument systems in which a "conditional" $A \rightarrow B$ means A is an argument, or a defeasible reason, for B (Loui 1987a; Pollock 1987; Nute 1984a; Nute 1984b; Lin and Shoham 1989). Two very influential conditional representations are the *cumulative* (and especially the *preferential*) consequence relations of (Kraus, Lehmann and Magidor 1990; Lehmann 1989) and the ε-semantics of (Pearl 1988). The motivation underlying preferential relations is quite similar to ours, and that of ε-semantics is rather different; yet both approaches turn out to be equivalent to fragments of CLNs developed in our framework.

4.3.1 Preferential and Rational Consequence Relations

The original notion of a logical consequence operator as studied by Tarski and Scott (see (Czelakowski and Mrmalinowski 1985)) required a consequence relation to be monotonic. In the Tarskian notation this means $Cn(X) \subseteq Cn(X \cup Y)$ where X and Y are any sets of sentences. In a Gentzen-style sequent calculus (Gentzen 1934) monotonicity is expressed as a rule of inference

$$\frac{\alpha \vdash \gamma}{\alpha, \beta \vdash \gamma}$$

Thus, from $\alpha \vdash \gamma$ one can infer $\alpha, \beta \vdash \gamma$, if the derivability relation $\vdash$ is monotonic. Gabbay (1985) proposed a generalization of logical consequence, motivated by default reasoning, in which monotonicity is violated for the purposes of representing the behavior of nonmonotonic inference systems as consequence relations. To distinguish this notion from (monotonic) logical derivability, Gabbay uses the relation symbol $\mathrel{|\!\sim}$ to denote nonmonotonic consequence, and suggests three basic conditions that any such relation should meet:

$$\alpha \mathrel{|\!\sim} \alpha \qquad \text{(Reflexivity)}$$

$$\frac{\alpha \wedge \beta \mathrel{|\!\sim} \gamma,\ \alpha \mathrel{|\!\sim} \beta}{\alpha \mathrel{|\!\sim} \gamma} \qquad \text{(Cut)}$$

$$\frac{\alpha \mathrel{|\!\sim} \beta,\ \alpha \mathrel{|\!\sim} \gamma}{\alpha \wedge \beta \mathrel{|\!\sim} \gamma} \qquad \text{(Weak Monotonicity)}$$

The theory of nonmonotonic consequence relations has been studied by a number of people, including Makinson's (1989) semantic account of these relations, and Besnard's (1988) study of proof-theoretic properties. But the most comprehensive treatment seems to be that of Lehmann and his colleagues (Kraus, Lehmann and Magidor 1990; Lehmann 1989; Lehmann and Magidor 1990). They study a series of consequence relations satisfying increasingly stronger principles and provide a model-theoretic account of each such logic. As discussed in the introduction to this chapter, the semantics of these systems is very much like our modal relational semantics and can be interpreted as ordering states of affairs according to some measure of normality.[23] The theory

[23] In fact, for weaker relations than those we examine here the orderings are on (possibly incomplete) *states of*

of *cumulative inference* is the weakest notion they develop (see also (Makinson 1989)) but we are interested particularly in two stronger systems, the logics of *preferential* and *rational* relations.

The suggested reading of a sequent $\alpha \mathrel{|\!\sim} \beta$, dubbed a *conditional assertion*, is something like "if α were all the information about the world available to an agent then β would be a sensible conclusion" (Lehmann 1989). We assume our underlying logic and language (from which α, β are drawn) to be classical propositional. Again, this is for convenience, as Kraus, Lehmann and Magidor (1990) treat arbitrary languages, assuming only certain connectives and compactness. A *nonmonotonic consequence relation* C is any binary relation on $\mathbf{L}_{CPL}$ satisfying certain requirements. We write $\alpha \mathrel{|\!\sim} \beta$ to indicate $\langle \alpha, \beta \rangle \in C$, and $\alpha \not\mathrel{|\!\sim} \beta$ when $\langle \alpha, \beta \rangle \notin C$. Two classes of relations are given particular emphasis by Lehmann (1989), namely preferential and rational relations, which satisfy the inference rules given in the following definitions (some of which are renamed to emphasize the similarity to various theorems of CLNs).

Definition 4.15 (Lehmann 1989) A *preferential consequence relation* is a consequence relation that satisfies the following rules of inference:

$$\frac{\models_{CPL} \alpha \equiv \beta,\ \alpha \mathrel{|\!\sim} \gamma}{\beta \mathrel{|\!\sim} \gamma} \qquad \textbf{(LLE)}$$

$$\frac{\models_{CPL} \alpha \supset \beta,\ \gamma \mathrel{|\!\sim} \alpha}{\gamma \mathrel{|\!\sim} \beta} \qquad \textbf{(RCM)}$$

$$\alpha \mathrel{|\!\sim} \alpha \qquad \textbf{(ID)}$$

$$\frac{\alpha \mathrel{|\!\sim} \beta,\ \alpha \mathrel{|\!\sim} \gamma}{\alpha \mathrel{|\!\sim} \beta \wedge \gamma} \qquad \textbf{(And)}$$

$$\frac{\alpha \mathrel{|\!\sim} \gamma,\ \beta \mathrel{|\!\sim} \gamma}{\alpha \vee \beta \mathrel{|\!\sim} \gamma} \qquad \textbf{(Or)}$$

$$\frac{\alpha \mathrel{|\!\sim} \beta,\ \alpha \mathrel{|\!\sim} \gamma}{\alpha \wedge \beta \mathrel{|\!\sim} \gamma} \qquad \textbf{(CM)}$$

Definition 4.16 (Lehmann 1989) A *rational consequence relation* is a preferential consequence relation that satisfies the following rule of inference:

$$\frac{\alpha \mathrel{|\!\sim} \gamma,\ \alpha \wedge \beta \not\mathrel{|\!\sim} \gamma}{\alpha \mathrel{|\!\sim} \neg\beta} \qquad \textbf{(RM)}$$

Families of models are proposed to characterize these notions of consequence. These models determine the truth of conditional assertions. Intuitively, $\prec$ ranks worlds according to their degree of normality.

---

knowledge, or *sets* of states of affairs, rather than possible worlds.

Definition 4.17 (Lehmann 1989) Let $\langle X, \prec \rangle$ be a poset. $V \subseteq X$ is *smooth* iff for each $v \in V$, either v is minimal in V (that is, there is no $x \in V$ such that $x \prec v$) or there is some element w minimal in V such that $w \prec v$.

Definition 4.18 (Lehmann 1989) A *preferential model* (or, a P-model) M is a triple $\langle S, \varphi, \prec \rangle$ where S is a set (of possible worlds), φ maps propositional variables into 2^S ($\varphi(A)$ is the set of worlds where A holds) and $\prec$ is a strict partial order on S such that for all propositional formulae α, $\|\alpha\|$ is smooth.

Definition 4.19 (Lehmann 1989) A *ranked model* (R-model) is a preferential model $M = \langle S, \varphi, \prec \rangle$ where the relation $\prec$ is such that there exists a totally ordered set $\langle \Omega, < \rangle$ and a function $f : S \to \Omega$, where $s \prec t$ iff $f(s) < f(t)$.

Definition 4.20 (Lehmann 1989) A P-model or R-model $M = \langle S, \varphi, \prec \rangle$ *satisfies* a conditional assertion $\alpha \mathrel{|\!\sim} \beta$ (written $\alpha \mathrel{|\!\sim}_M \beta$) iff for any s $\prec$-minimal in $\|\alpha\|$, $s \in \|\beta\|$. $\mathrel{|\!\sim}_M$ is the *consequence relation defined by* M.

Intuitively, a preferential model can allow certain states to be incomparable according to the ranking $\prec$. The requirement that ranked models respects some total ordering amounts to insisting that each world be comparable in the partial order; for if each world has some rank in Ω, then it must be comparable (via $<$) with each other world. The similarity to CT4 and CT4D models possessed by such structures is readily apparent. The smoothness condition is required to ensure the truth conditions for $\alpha \mathrel{|\!\sim} \beta$ are not vacuous when α is possible, and is precisely the Limit Assumption. The following representation results are also obtained.

Theorem 4.13 (Lehmann 1989) *$\mathrel{|\!\sim}$ is a preferential consequence relation iff it is the consequence relation defined by some P-model.*

Theorem 4.14 (Lehmann 1989) *$\mathrel{|\!\sim}$ is a rational consequence relation iff it is the consequence relation defined by some R-model.*

The logic defined by preferential consequence relations is called P and we denote by R the logic of rational relations.

Let KB be some set of conditional assertions. The consequences of KB in P are just those assertions derivable using the rules of P, and these form the set of assertions *preferentially entailed* by KB, or true in all P-models of KB. Presumably, rational relations are more reasonable in many instances as they must satisfy **RM**. We would hope a similar result holds for the logic R. Unfortunately, Lehmann presents the following apparently discouraging result.

Theorem 4.15 (Lehmann 1989) *Let KB be a set of assertions. An assertion A is satisfied by all R-models of KB iff A is satisfied by all P-models of KB.*

This result seems to suggest that the logic R is no stronger than the logic P. However, on further inspection of the theorem, it should be evident that this result obtains only because of the restricted form of sentences in KB. Consider the rule **RM**, which distinguishes R from P. This rule allows the derivation of $\alpha \mathrel{|\!\sim} \neg\beta$ when $\alpha \mathrel{|\!\sim} \gamma$ and $\alpha \wedge \beta \not\mathrel{|\!\sim} \gamma$ hold. However, when our set of premises KB consists solely of conditional assertions, the second premise, $\alpha \wedge \beta \not\mathrel{|\!\sim} \gamma$ cannot be represented, for

it says a certain assertion is not derivable. The best we could hope for is to, say, assert $\alpha \mathrel{|\!\sim} \gamma$ and infer $\alpha \wedge \beta \not\mathrel{|\!\sim} \gamma$ due to its absence from KB; but of course, this is not derivable because, in general, $\alpha \wedge \beta \mathrel{|\!\sim} \gamma$ is consistent with $\alpha \mathrel{|\!\sim} \gamma$. There exist (ranked) models of KB that satisfy $\alpha \wedge \beta \mathrel{|\!\sim} \gamma$, so *all* models of KB need not satisfy the second premise of **RM**, and the conclusion of **RM** is not derivable.

The obvious solution for this specific problem is to allow for the inclusion of "negations" of assertions among our premises KB. In this case, Theorem 4.15 cannot hold (consider a generic instance of **RM**, which will not be true of all P-models of KB). In essence, the language of conditional assertions is too weak to express the negation of assertions, or any boolean combinations of such. This is acceptable if an assertion is viewed as constraining some consequence relation. But we claim a more profitable interpretation of $\mathrel{|\!\sim}$ is as a conditional *connective* in some language, rather than a relation symbol in the meta-language.

We can extend the language of conditional assertions in the obvious way, permitting boolean combinations of assertions as well as propositional formulae (with no occurrence of $\mathrel{|\!\sim}$) by viewing $\mathrel{|\!\sim}$ as a connective. We will permit these extended assertions to include sentences such as $\neg(\alpha \mathrel{|\!\sim} \beta)$ or $(\alpha \mathrel{|\!\sim} \beta) \vee (\alpha \mathrel{|\!\sim} \gamma)$, but forbid nesting of the connective $\mathrel{|\!\sim}$ (e.g., $\alpha \mathrel{|\!\sim} (\beta \mathrel{|\!\sim} \gamma)$); so no $\mathrel{|\!\sim}$ can appear within the scope of another.

The well-formed formulae of this enriched language will be called *extended conditional assertions*, the set of which is denoted $\mathbf{L}_{EC}$. It's not hard to see that P and R can be extended with little difficulty to account for these new formulae. To capture reasonable inferences using this language of extended assertions, we must enhance the systems R and P to reason propositionally with assertions. P* and R* will denote the systems obtained by augmenting P and R with the axiom and rule schemata of CPL together with the axiom $(\neg A \mathrel{|\!\sim} A) \supset A$. CPL will allow standard propositional reasoning with conditional assertions as well as nonconditional formulae. The additional axiom allows interaction between propositions and assertions, and corresponds to the modal axiom **T** ($\Box A \supset A$). We also translate the inference rules of P and R into the corresponding Hilbert-style axioms (and note that the inference rules remain valid due to **MP**).[24]

Definition 4.21 P* is the smallest set $S \subseteq \mathbf{L}_{EC}$ containing CPL and the following axioms, and closed under the following rules of inference:

ID $\alpha \mathrel{|\!\sim} \alpha$

And $(\alpha \mathrel{|\!\sim} \beta \wedge \alpha \mathrel{|\!\sim} \gamma) \supset \alpha \mathrel{|\!\sim} \beta \wedge \gamma$

Or $(\alpha \mathrel{|\!\sim} \gamma \wedge \beta \mathrel{|\!\sim} \gamma) \supset \alpha \vee \beta \mathrel{|\!\sim} \gamma$

CM $(\alpha \mathrel{|\!\sim} \beta \wedge \alpha \mathrel{|\!\sim} \gamma) \supset \alpha \wedge \beta \mathrel{|\!\sim} \gamma$

T $(\neg\alpha \mathrel{|\!\sim} \alpha) \supset \alpha$

LLE From $\models_{CPL} \alpha \equiv \beta$ infer $(\alpha \mathrel{|\!\sim} \gamma) \supset (\beta \mathrel{|\!\sim} \gamma)$

RCM From $\models_{CPL} \alpha \supset \beta$ infer $(\gamma \mathrel{|\!\sim} \alpha) \supset (\gamma \mathrel{|\!\sim} \beta)$

MP From $\alpha \supset \beta$ and α infer β

US From α infer α', where α' is a substitution instance of α

Definition 4.22 R* is the smallest set $S \subseteq \mathbf{L}_{EC}$ that contains P* and the following axiom, and is closed under the inference rules **LLE**, **RCM**, **MP**, and **US**:

[24] Thanks to Karl Schlechta for pointing out the need to translate **LLE** and **RCM** further than was done in an earlier paper (Boutilier 1990).

RM $(\alpha \mathrel{|\!\sim} \gamma \wedge \neg(\alpha \wedge \beta \mathrel{|\!\sim} \gamma)) \supset \alpha \mathrel{|\!\sim} \neg\beta$

We will use $\vdash_{P*}$ and $\vdash_{R*}$ to denote derivability in these systems, which is defined in the usual manner. The notions of satisfiability and validity in P-models will be adjusted as follows:

Definition 4.23 Let $M = \langle S, \varphi, \prec \rangle$ be a P-model, and let $s \in S$. The truth of an extended conditional assertion $\alpha \in \mathbf{L}_{EC}$ at s ($M \models_s \alpha$ means α is true at s) is defined inductively as follows:

1. $M \models_s \alpha$ iff $s \in \varphi(\alpha)$ for atomic sentence α.
2. $M \models_s \neg\alpha$ iff $M \not\models_s \alpha$.
3. $M \models_s \alpha \supset \beta$ iff $M \models_s \beta$ or $M \not\models_s \alpha$.
4. $M \models_s \alpha \mathrel{|\!\sim} \beta$ iff $\alpha \mathrel{|\!\sim}_M \beta$.

Sentence α is *valid on* M (written $M \models \alpha$) iff $M \models_s \alpha$ for each $s \in S$. α is *P*-valid* ($\models_{P*} \alpha$) iff $M \models \alpha$ for each P-model M. α is *R*-valid* ($\models_{R*} \alpha$) iff $M \models \alpha$ for each R-model M.

It is not hard to see P* and R* correspond to the classes of P-models and R-models, respectively, using this extended notion of validity, and that these logics extend P and R in a natural way. In fact, P* and R* are not much more interesting than P and R, except they will allow us to show a correspondence between these notions of nonmonotonic consequence and CLNs.

Theorem 4.16 $\models_{P*} \alpha$ *iff* $\vdash_{P*} \alpha$.

Theorem 4.17 $\models_{R*} \alpha$ *iff* $\vdash_{R*} \alpha$.

That each of the inference rules of P is a valid derived rule or theorem of CT4 can be readily shown. Furthermore, the additional rule of Rational Monotonicity from R is also valid in CT4D. Therefore all assertions A derivable from a set of assertions KB in P are derivable from KB in CT4 (provided we translate $\alpha \mathrel{|\!\sim} \beta$ in the first instance to $\alpha \Rightarrow \beta$ in the second). Also, A is derivable from KB in R only if it is derivable in CT4D. Thus CT4 and CT4D subsume P and R, respectively. We'd like to show this inclusion is not proper and that, in fact, each corresponding pair of logics is equivalent. This cannot be the case of course, as the languages are different, $\mathbf{L}_C$ allowing nested conditionals. What we can show is that, phrased as the derivability of assertions only, the restricted logics CT4 and CT4D are equivalent to their corresponding consequence operators.

Theorem 4.18 *Let KB be a set of conditional assertions and A an assertion. A is preferentially entailed by KB iff* $KB \vdash_{CT4} A$.

Theorem 4.19 *Let KB be a set of conditional assertions and A an assertion. A is rationally entailed by KB iff* $KB \vdash_{CT4D} A$.

An important corollary of these results and Lehmann's result, Theorem 4.15, is that CT4 and CT4D are indistinguishable with respect to the task of reasoning with *simple* conditional sentences (where $\Rightarrow$ is the main connective and no nesting occurs).

Definition 4.24 A *simple conditional* is a sentence of the form $\alpha \Rightarrow \beta$, such that $\alpha, \beta \in \mathbf{L}_{CPL}$.

Corollary 4.20 *Let* $KB \subseteq \mathbf{L}_C$ *be a set of simple conditionals and A a simple conditional. Then* $KB \vdash_{CT4} A$ *iff* $KB \vdash_{CT4D} A$.

Rather than prove these results directly, we can show even stronger results via an excursion through the language of extended conditional assertions. Now if KB is a set of extended assertions then again it should be clear that derivability from KB in P* and R* is subsumed by derivability in CT4 and CT4D, respectively. If we restrict attention to the fragments of CT4 and CT4D containing no nested conditionals, we can show the converse holds and that P* and R* are exactly CT4– and CT4D–. Thus the motivation underlying CLNs is corroborated by the equivalence of these existing logics to logics developed within our framework.

Definition 4.25 An *extended conditional* is any sentence $\alpha \in \mathbf{L}_C$ such that no occurrence of $\Rightarrow$ lies within the scope of another (this includes propositional formulae). The language of extended conditionals, $\mathbf{L}_C^-$, is the (largest) sublanguage of $\mathbf{L}_C$ containing no nested conditional sentences.

Definition 4.26 Let system S be a CLN. By S– we denote the system S restricted to extended conditionals. That is, S– = S $\cap$ $\mathbf{L}_C^-$.

Theorem 4.21 *Let* $A \in \mathbf{L}_C^-$. *Then* $\vdash_{CT4-} A$ *iff* $\vdash_{P*} A$.

Theorem 4.22 *Let* $A \in \mathbf{L}_C^-$. *Then* $\vdash_{CT4D-} A$ *iff* $\vdash_{R*} A$.

These theorems are proven using the similarity of the structure of CT4 (respectively CT4D) models and P (respectively R) models, showing that satisfiable conditional assertions have satisfiable counterparts in $\mathbf{L}_C^-$. From these, Theorems 4.18 and 4.19 are easily derivable corollaries. Thus, viewing the nonmonotonic consequence relations of Kraus, Lehmann and Magidor (1990) and Lehmann (1989) as more standard conditional logics, and interpreting the relation symbol $\mid\!\sim$ as a conditional connective, allows the conclusion that they are equivalent to "flat" fragments of our CLNs CT4 and CT4D, as well as to fragments of the standard modal logics S4 and S4.3.

Extending P* and R* additionally to allow nested conditionals suggests that we could push the equivalence even further, but this is not the case. The treatment of conditionals in preferential models is *global* whereas our treatment is *local.* More precisely, if an assertion is satisfied at some world in a P-model, it is satisfied at all worlds. Our semantics is local in that only worlds accessible to w (or more normal than w) can influence the evaluation of a conditional in a CT4-model. So if, say, some model contained an A-world, but w was more normal than the most normal A-worlds, w would satisfy $\Box\neg A$ and, hence, would validate any conditional $A \Rightarrow B$, given the CT4 truth conditions. In contrast, the preferential semantics would make $A \Rightarrow B$ true, *even at* w, only if B were true at these most normal A-worlds. For this reason, our CLNs are equivalent to the usual modal logics. We can express any modality using the connective $\Rightarrow$, while the global quality of $\mid\!\sim$ forbids this. In effect, $\mid\!\sim$ can be used to express the distinct modalities in S4 and S4.3, but cannot express arbitrary modal functions.[25]

There are a number of reasons to consider the global definition of Kraus, Lehmann and Magidor (1990) more reasonable and we explore these in the next chapter, where we will modify the truth conditions for $\Rightarrow$ somewhat to reflect this.

Other notions of nonmonotonic consequence have been proposed by Kraus, Lehmann and Magidor (1990), as well. An interesting avenue to pursue is the further examination of these

[25] We can show that each of the ten distinct modalities of S4.3 (Dummett and Lemmon 1959) is expressible using no nesting of $\Rightarrow$ or $\mid\!\sim$ (hence, within $\mathbf{L}_{EC}$); and we expect that the fourteen modalities of S4 are also so expressible. However, S4.3 (and hence S4) has infinitely many modal functions (Makinson 1966), so obviously a lack of nesting decreases expressive power.

consequence relations to determine the extent to which they can be characterized as CLNs or modal logics. We remark here that the strongest system that they study, denoted M, appears to correspond to the conditional logic CT45 and the modal system S5. The system M is determined by preferential models where the ordering $\prec$ is empty. As noted there, preferential entailment reduces to logical entailment relative to the underlying set of possible worlds. Of course, this is exactly strict entailment in an S5 setting and, as we've noted above, normal implication and strict implication within S5 coincide. The global nature of $\mid\!\sim$ in this circumstance is not restrictive at all, for all modal functions in S5 are reducible to simple modal functions of one modality (cf. (Hughes and Cresswell 1968)). Each of these six distinct modalities in S5 is definable using $\mid\!\sim$.

4.3.2 ε-semantics

Another conditional theory of default reasoning based on quite different intuitions is the theory of ε-semantics pursued by Pearl and his colleagues (Pearl 1988; Goldszmidt and Pearl 1989; Goldszmidt, Morris and Pearl 1990; Pearl 1990). This work takes as a starting point the system of probabilistic inference developed by Adams (1975). Adams's logic generalizes straightforward deductive inference by characterizing inference that preserves "high probability" rather than truth; that is, if a set of premises has some arbitrarily high probability, then the consequences should also have high probability. For this reason, he identifies the probability of a conditional with its conditional probability, instead of with the probability that the conditional is true. For instance, consider the paradox of material implication where one infers $A \supset B$ from $\neg A$. While this inference preserves high probability, if the probability of the conditional "If A then B" is identified with $P(B|A)$ instead of $P(A \supset B)$, this inference is no longer sound.[26]

Pearl (1988) proposes a semantics similar to Adams's for conditional defaults. The idea is roughly that a default statement, say "birds fly," makes an assertion that the conditional probability $P(\texttt{fly}|\texttt{bird})$ is sufficiently high so that one can conclude that an arbitrary bird flies. In general, we do not specify some acceptance criterion, so "sufficiently high" is left intentionally vague. However, if we want to infer new default rules from a given set, then whatever acceptance criteria we adopt (implicitly) in certain situations should be satisfied by the conclusions. Thus, arbitrarily high conditional probabilities must be preserved. These notions are generalized further in (Goldszmidt and Pearl 1989) and we present a number of definitions taken from there. We will assume an underlying propositional language generated by a finite set of variables. Its conditional extension includes sentences $\alpha \rightarrow \beta$ and $\alpha \Rrightarrow \beta$, but these connectives can only appear as the main connective in a sentence (so $\alpha, \beta \in \mathbf{L}_{CPL}$). We refer to these sentences as *rules*. The *material counterpart* of either rule is $\alpha \supset \beta$.

A sentence $A \rightarrow B$ is intended as a default statement of the aforementioned type, and we take it to mean "the conditional probability of B given A is *high*," or "if A then typically B." $A \Rrightarrow B$ is meant to be a strict sentence and asserts that B is certain given A, or if A is true then B must be true. A default theory $T = D \cup S$ consists of a set D of defeasible sentences (having $\rightarrow$ as the main connective) and a set S of strict sentences (having $\Rrightarrow$ as the main connective). We are also interested in how truth assignments (or worlds) treat the material counterparts of such sentences.

Definition 4.27 (Goldszmidt and Pearl 1989) A valuation (possible world) w *verifies* the rule $\alpha \rightarrow \beta$ iff $w \models \alpha \wedge \beta$, *falsifies* the rule iff $w \models \alpha \wedge \neg\beta$, and *satisfies* the rule iff $w \models \alpha \supset \beta$, and similarly for $\alpha \Rrightarrow \beta$.[27]

[26] We can make $P(A \supset B)$ arbitrarily high by making $P(\neg A)$ sufficiently large, but this is not the case for $P(B|A)$.

[27] This use of "satisfies" is ours, and means the rule is not falsified.

The models of interest are *probability assignments* that assign probabilities to each possible world. From such an assignment P we can determine the probability of any sentence; and, of immediate interest, the conditional probability of B given A is given by

$$P(B|A) = \frac{\sum\{P(w) : w \models A \wedge B\}}{\sum\{P(w) : w \models A\}} .$$

As assignment is said to be *proper* (for T) if

$$\sum\{P(w) : w \models A\}$$

is non-zero for every antecedent A in T, and means the conditional probability is defined for each conditional in T. Intuitively, a model for T is any probability assignment that makes each conditional reasonably probable. So if $A \rightarrow B \in D$ then $P(B|A)$ should be high, while $P(B|A)$ should be 1 for any $A \Rrightarrow B \in S$.[28] The definition of consistency reflects the noncommittal nature of these considerations.

Definition 4.28 (Goldszmidt and Pearl 1989) Let $T = D \cup S$ be a default theory. T is *ε-consistent* if, for each $\varepsilon > 0$, there is a proper assignment P such that $P(\alpha) \geq 1 - \varepsilon$ for all $\alpha \in D$, and $P(\alpha) = 1$ for all $\alpha \in S$.

We will say such an assignment P *satisfies* T *to degree* ε. If there is no such proper assignment for T, for some ε, T is *ε-inconsistent*. It can be that T is inconsistent because no proper assignment satisfies some antecedent A, rather than it being the case that the conditional probabilities cannot be made sufficiently high. For instance, if D contains $A \rightarrow B$ and S contains $A \Rrightarrow \neg A$ then $\neg A$ must have probability 1 in any distribution P, making P improper. A stronger notion of consistency rules out such a circumstance. This is analogous to prohibiting the inference of $A \nRightarrow B$ from $\Box\neg A$ in CT4.

Definition 4.29 (Goldszmidt and Pearl 1989) Let T be ε-consistent and α some (strict or defeasible) rule with antecedent A. $T \cup \{\alpha\}$ is *substantively inconsistent* if $T \cup \{A \rightarrow \top\}$ is ε-consistent but $T \cup \{\alpha\}$ is *ε-inconsistent*.

Goldszmidt and Pearl also extend Adams's notion of confirmability, which is used to characterize ε-consistency. α is tolerated by a set of conditionals if it can be verified by some world that satisfies all conditionals in the set.

Definition 4.30 (Goldszmidt and Pearl 1989) Let $\alpha \in T$ be some rule. *α is tolerated* by $T - \{\alpha\}$ iff there is some valuation v such that v verifies α and v satisfies all members of $T - \{\alpha\}$.

Definition 4.31 (Goldszmidt and Pearl 1989) $T = D \cup S$ is *confirmable* iff some rule $\alpha \in D$ is tolerated by $T - \{\alpha\}$ (when $D \neq \emptyset$) or each rule $\alpha \in S$ is tolerated by $T - \{\alpha\}$ (when $D = \emptyset$).

This notion of confirmability allows a generalization of Adams's result on probabilistic consistency.

Theorem 4.23 (Goldszmidt and Pearl 1990) *T is ε-consistent iff every nonempty subset of T is confirmable.*

[28] We identify the probability of statements in T with their conditional probabilities.

Thus a simple decision procedure for testing ε-consistency (and ε-entailment to follow) can be constructed (which we discuss in the next chapter). The notion of entailment in ε-semantics is, of course, based on the idea of preserving arbitrarily high probabilities of the premises.

Definition 4.32 (Goldszmidt and Pearl 1989) Let T be an ε-consistent default theory and α a conditional rule. T *ε-entails* α iff there exists a proper assignment for $T \cup \{\alpha\}$ and, for all $\varepsilon > 0$, there exists some $\delta > 0$ such that for any proper assignment P that satisfies T to degree δ, $P(\alpha) \geq 1 - \varepsilon$.

This is related to ε-consistency as follows:

Theorem 4.24 (Goldszmidt and Pearl 1990) *Let T be an ε-consistent default theory. Then T ε-entails α iff $T \cup \{\neg\alpha\}$ is substantively inconsistent.*

It has been noted by Kraus, Lehmann and Magidor (1990) that for default theories with no strict conditionals ε-entailment corresponds precisely to preferential entailment.[29] From this result the following corollary is automatic.

Definition 4.33 Let $T \subseteq \mathbf{L}_C$ be a finite set of simple conditionals. T^ε denotes the corresponding set of probabilistic rules

$$T^\varepsilon = \{A \rightarrow B : A \Rightarrow B \in T\}.$$

Corollary 4.25 *Let $T \subseteq \mathbf{L}_C$ be a finite set of simple conditionals, and T^ε be ε-consistent. Then $T \vdash_{CT4} A \Rightarrow B$ iff T^ε ε-entails $A \rightarrow B$.*

This does not hold in the general case where T^ε is inconsistent, for CT4 allows conditionals to hold vacuously; for example, $A \Rightarrow B$ and $A \Rightarrow \neg B$ are mutually CT4-consistent and entail $\Box\neg A$. This corresponds to allowing improper assignments for T^ε, something disallowed by ε-entailment. However, we can discount this difference by prohibiting vacuously satisfied conditionals as follows:

Definition 4.34 Let $T \subseteq \mathbf{L}_C$ be a finite set of simple conditionals. The *proper counterpart* of T is

$$T^P = \{\Diamond A : A \Rightarrow B \in T\} \cup T.$$

Now we can use Pearl's notion of confirmability to determine a proof procedure for simple conditional theories.

Theorem 4.26 *T^P is CT4-consistent iff every non-empty subset of T is confirmable.*

As a very simple corollary we have the following.

Corollary 4.27 *Let $T \subseteq \mathbf{L}_C$ be a finite set of simple conditionals. T^P is CT4-inconsistent iff T^ε is not ε-consistent.*

We've seen the substantive inconsistency of a conditional is due to its consequent not achieving the desired probability, rather than its antecedent being unsatisfiable in any proper distribution. In CT4, this is reflected in the possibility of the antecedent.

[29] At least for finite theories, since ε-entailment is not compact (Adams 1975).

Theorem 4.28 *Let $T \subseteq \mathbf{L}_C$ be a finite set of simple conditionals. $T^\epsilon \cup \{A \rightarrow B\}$ is substantively inconsistent iff $T^P \cup \{\Diamond A\}$ is CT4-consistent, while $T^P \cup \{\Diamond A, A \Rightarrow B\}$ is CT4-inconsistent.*

This leads to another key result, the equivalence of ε-semantics to (a fragment of) CT4.

Theorem 4.29 *Let $T \subseteq \mathbf{L}_C$ be a finite set of simple conditionals such that $T^P \cup \{\Diamond A\}$ is CT4-consistent. $T^P \vdash_{CT4} A \Rightarrow B$ iff T^ϵ ε-entails $A \rightarrow B$.*

Thus substantive inconsistency amounts to inconsistency in CT4 of a proper set of conditionals, and Pearl's test for inconsistency can be used to test the consistency of a set of conditionals in CT4, or the substantive consistency of such a set. As well, ε-entailment is seen to be equivalent to a fragment of CT4, and, hence, to the modal logic S4. The obvious extensions including strict sentences follow easily if we translate $A \Rightarrow B$ into $\Box(A \supset B)$.

Definition 4.35 Let $T \subseteq \mathbf{L}_C$ be a finite set of simple conditionals. The set of *strict sentences* in T is the subset of T

$$S(T) = \{\alpha \Rightarrow \beta \in T : \vdash_{CPL} \alpha \equiv \neg(A \supset B) \text{ and } \vdash_{CPL} \beta \equiv A \supset B\}.$$

For each such sentence $\alpha \Rightarrow \beta$ we write $\Box(A \supset B) \in T$. The set of *default sentences* in T is $D(T) = T - S(T)$. The *strict proper counterpart* of T is

$$T^P = \{\Diamond A : A \Rightarrow B \in D(T) \text{ or } \Box(A \supset B) \in S(T)\} \cup T.$$

We usually take the proper counterpart of T to mean this strict version. T^ϵ denotes the corresponding set of probabilistic conditionals

$$T^\epsilon = \{A \rightarrow B : A \Rightarrow B \in D(T)\} \cup \{A \Rightarrow B : \Box(A \supset B) \in T\}.$$

Corollary 4.30 *Let $T \subseteq \mathbf{L}_C$ be a finite set of simple conditionals including strict sentences. T^P is CT4-inconsistent iff T^ϵ is not ε-consistent.*

Corollary 4.31 *Let $T \subseteq \mathbf{L}_C$ be a finite set of simple conditionals including strict sentences. $T^P \vdash_{CT4} A \Rightarrow B$ iff T^ϵ ε-entails $A \rightarrow B$.*

The intuitions that underlie ε-semantics and CT4 are genuinely distinct and it seems somewhat surprising they should sanction precisely the same inferences, and that ε-semantics corresponds to the modal system S4. However, an examination of Adams's (1975) construction equating ε-consistency with confirmability of all subsets reveals a very close relationship, even at a model-theoretic level. Roughly, Adams's construction proceeds as follows: if every nonempty subset of T is confirmable, there exists some sequence of subsets of T, say $T_n \subset \cdots \subset T_2 \subset T_1 = T$, such that for each T_i

(a) There exists a world w_i verifying each $\alpha \in T_i - T_{i-1}$ and falsifying no sentence in T_i (so w_n verifies all of T_n).

(b) T_{i-1} is the set of sentences in T_i not verified by w_i.

Adams shows this is sufficient for ε-consistency by associating, for fixed ε, (unnormalized) probabilities of $(1-\varepsilon)\varepsilon^{i-1}$ with each world w_i, thus making the conditional probability of each member of T at least $1-\varepsilon$.[30] More importantly, we see that the (proper) conditional counterpart of T will be CT4-satisfiable just if it can be partitioned similarly. If this is the case then we can construct a CT4-model for T where the set of worlds is just the set of w_i, and w_i is more normal than w_j just when $i < j$. In other words, if we rank the worlds w_i according to the probability assigned by Adams's construction, we determine a CT4-model for the corresponding set of proper conditionals. In a sense more normal worlds can be arbitrarily more probable than their less normal counterparts.

We note here that by Corollary 4.20 any of the definitions and results relating CT4 to ε-semantics hold for CT4D also.[31] As with the relationship to preferential and rational consequence relations, this is due to the equivalence of rules to *simple* conditionals. To distinguish the logics, more expressiveness is required. Hence, the logics CT4 and CT4D suggest natural extensions to ε-semantics that include nesting and boolean combinations of default rules. It is not clear that the essential nature of the semantic interpretation of rules in terms of high probabilities can be preserved in this extension, however. This would suggest that CLNs (and preferential relations) provide a somewhat more natural and robust interpretation of defaults, in this respect.

4.4 Miscellany

The framework presented for conditional logics of normality seems very general and intuitively appealing. However, its generality and applicability is reinforced by the fact that logics within the literature, while independently motivated, turn out to be equivalent to the "unnested" fragments of logics developed in this framework. Theorems 4.2 and 4.3 show the completeness of our conditional logic CT4 and its equivalence to the modal logic S4, while Corollary 4.5 demonstrates that any extension of CT4 is also equivalent to a modal system. Thus we can use modal logics or conditional logics interchangeably here, taking either as primitive. Theorems 4.21 and 4.22, that demonstrate that Lehmann's (1989) logics P and R are fragments of our conditional logics; and Theorem 4.29 shows the same result for Pearl's (1988) ε-semantics. Moreover, Theorem 4.26 determines a simple proof procedure for simple conditional sentences based on Pearl's notion of confirmability.

While we have seen that our CLNs are very similar to the logics of (Delgrande 1987; Kraus, Lehmann and Magidor 1990; Goldszmidt and Pearl 1989), there are a number of differences between our account of conditional logics and existing accounts in the literature. Viewing CLNs as extensions of CT4 provides several conceptual and practical advantages from the standpoint of default reasoning research. This perspective suggests a wide variety of conditional logics, which may determine useful interpretations of normality. The correspondence with standard modal systems provides a widely-studied, well-developed and well-understood semantics for such logics. Furthermore, this relationship permits the appropriation of a host of ready-made results for these logics, results regarding axiomatizability, axiomatic bases, decision procedures and their complexity, and the like. For example, Lehmann (1989) showed that deciding whether $K \models_P \alpha \mid\sim \beta$ is a problem in co-NP when K is a finite set of assertions. Using the correspondence between R* and CT4D, and the fact that the problem of deciding S4.3-satisfiability is NP-complete (Ono and Nakamura 1980), we can state the following stronger result.

Corollary 4.32 *For a finite set of extended assertions $K \cup \{\alpha\}$, deciding whether $K \models_{R*} \alpha$ is in*

[30] The exception is $P(w_n) = \varepsilon^{n-1}$.

[31] Semantically, this can be seen by the nature of Adams's construction. The verifying worlds are not just partially ordered, but totally ordered by probability. Thus, the corresponding CT4-model is also a CT4D-model.

co-NP.

That this holds for P in the case of extended assertions (i.e. P*) does not follow immediately, for S4-satisfiability is PSPACE-complete (Ladner 1977). It might still be the case however, as the language of extended assertions does not allow the expression of nested S4-modalities (that is, arbitrary S4 modal functions). As well, the validity problem for CT4D– is in co-NP and that of CT4D is co-NP-hard.

Regarding conditional logics as CLNs not only provides a uniform basis for comparison of such logics, but also extends the type of reasoning that can be performed using conditional logics, as they typically appear in the literature. More specifically (as discussed in Section 4.2 in the context of background and evidential knowledge) conditional logics, including those of (Delgrande 1987; Delgrande 1988; Lehmann 1989; Kraus, Lehmann and Magidor 1990; Nute 1984a), do not allow nested occurrences of the conditional connective in the language or do not provide an adequate semantic account of such sentences. Probabilistic accounts of conditionals (e.g., (Adams 1975; Pearl 1988)) suffer from the same weakness. CLNs, on the other hand, do allow such sentences, which might be of some value. A number of examples of default knowledge that appear to require nested defaults are given in (Asher and Morreau 1991). For instance, a sentence "Typically people who do not normally drive do not normally fly either" might be expressed as

$$(P \Rightarrow \neg D) \Rightarrow (P \Rightarrow \neg F).$$

Other sentences are interesting for logical reasons. For example, the following sentences are theorems of CT4 and its extensions:

$$(A \wedge (A \Rightarrow B)) \Rightarrow B,$$

$$(A \Rightarrow C) \Rightarrow ((A \wedge B) \Rightarrow C).$$

The first sentence appears to embody a rough version of the probabilistic principle of direct inference whereby the degree of belief associated with a sentence B, given that A holds, is equal to the conditional probability $P(B|A)$. Here we do not deal with degrees of belief or numerical probabilities, but rather with acceptance or rejection of facts, assuming normality. So when A and $A \Rightarrow B$ hold, we are willing to conclude B (in normal circumstances).

The latter sentence is important when dealing with a "principle of irrelevance" (see Chapter 5), which states that unless otherwise informed, assume that attributes are irrelevant or independent of one another. This principle allows one to conclude, for instance, that yellow birds normally fly, given that birds normally fly. This inference is problematic for most logics of normality (and probabilistic logics (Pearl 1988)) and requires the meta-inference of irrelevance. This theorem of CT4 can be seen as justifying this principle as being true in the normal state of affairs, and therefore "irrelevance" (or "independence" in probabilistic terms) is just another default inference. The next chapter will focus on the problem of handling the problem of irrelevance in a logical framework.

Several avenues for future study of CLNs remain open. One concerns weaker notions of normal implication. These may be investigated by studying logics weaker than CT4, or by allowing weaker definitions of the connective $\Rightarrow$. For instance, in CT4D, $A \Rightarrow B$ is equivalent to

$$\Box\neg A \vee \Diamond(A \wedge \Box(A \supset B))$$

which is weaker than its definition for CLNs in general.[32] This weaker notion of normal implication might be interesting in subsystems of CT4D. In CT4, for example, this definition allows $A \Rightarrow B$

[32] See Proposition 4.8. In Chapter 6, we will motivate this definition in terms of belief revision.

and $A \Rightarrow \neg B$ to be consistent with A, and has possibly useful interpretations. It might lead to some notion of *paraconsistent* default reasoning in which one can believe (by default) both a sentence and its negation without committing to belief in every proposition. Again, the appealing aspect of this model of paraconsistency is the standard modal semantics used to characterize it. In some ways, this logic might be more natural than traditional paraconsistent logics and the associated nonstandard semantics for inconsistent beliefs (Belnap 1977; Levesque 1984b; Lakemeyer 1987). It might also turn out to correspond quite naturally to the idea of a "belief cell" (cf. (McArthur 1988)), for the inconsistent default conclusions only arise when certain sets of worlds are considered incomparable, restricting the interaction of their "consequences."

Another connection worthy of investigation is that of CLNs with probabilistic logics, such as that of (Bacchus 1990). While the relationship to the (nonstandard) probabilistic logic of ε-semantics is very clear it remains to be seen how exactly notions such as conditionalization can be approximated within our framework. Clearly, some parallels exist, for instance, with our qualitative version of direct inference. Bacchus's logic also appears close to ours in the ability to conditionalize on (objective) probabilistic information. We have given a cursory treatment of the relationship to his logic, but have yet to investigate the connection in great detail.

Chapter 5

The Problem of Irrelevance

In the previous chapter we presented a family of logics for representing and reasoning with normative statements, which may be interpreted as default rules or prototypical facts. But we said little about characterizing the process of default reasoning using these logics. Given a set of facts, including normative facts such as "birds normally fly" and "penguins normally don't," and factual statements like "Tweety is a bird," what appropriate default inferences should be forthcoming?

Of course, the obvious answer is that the logical consequences of the knowledge base (in the desired logic, say CT4D) are the desired conclusions. Such a proposal is easily dismissed, for the very properties that make CT4D an attractive logic of normality also encumber the system with a caution unsuitable for default reasoning. For example, given the set of premises

$$KB = \{\texttt{bird}, \texttt{bird} \Rightarrow \texttt{fly}\},$$

a desirable default conclusion is `fly`. However, `fly` cannot be a logical consequence of KB as $\neg$`fly` is consistent with this information. This is logically appealing since we require that the conditional exemplify this exception-allowing quality. Thus the burden of "proof" must be placed on some default reasoning component.

In Section 4.1.3 we alluded to a mechanism for default reasoning using CT4D (or CT4)[1] based on the intuition that we should be interested not only in the *logical* consequences of KB, but in the *normative* consequences. In other words, we should ask what is normally the case when KB is true rather than what must be the case. This exploits the fact that Weak Modus Ponens is a theorem of CT4D. Thus, given a KB including both "background" and "evidential" statements, we ask

$$\vdash_{CT4D} KB \Rightarrow \alpha$$

to ascertain the status of a potential default conclusion α. Naturally, default inference includes logical inference since $\vdash_{CT4D} \Box(KB \supset \alpha)$ whenever KB entails α, and

$$\vdash_{CT4D} \Box(KB \supset \alpha) \supset (KB \Rightarrow \alpha).$$

This proposal deals adequately with a large class of examples and captures the intuition that we are willing to accept the most normal states of affairs that satisfy KB. For instance, given the

[1] We will use CT4D as the logic of choice in our exposition of concepts in this chapter for clarity. When distinctive features of CT4D are integral to some idea this will be made clear.

set of premises above, we can derive

$$\texttt{bird} \wedge (\texttt{bird} \Rightarrow \texttt{fly}) \Rightarrow \texttt{fly}.$$

More complicated examples are also handled quite nicely.

Example 5.1 Let KB be the set of premises

$$\{\texttt{bird} \Rightarrow \texttt{fly},\ \Box(\texttt{penguin} \supset \texttt{bird}),\ \Box(\texttt{emu} \supset \texttt{bird}),$$

$$\texttt{penguin} \Rightarrow \neg\texttt{fly},\ \texttt{emu} \Rightarrow \neg\texttt{fly}\}.$$

Adding bird to KB we can derive fly by default as

$$\vdash_{CT4D} KB \wedge \texttt{bird} \Rightarrow \texttt{fly}.$$

Adding penguin to KB, we can derive $\neg$fly, as well as when it is disjoined with emu, for

$$\vdash_{CT4D} KB \wedge \texttt{penguin} \Rightarrow \texttt{bird} \wedge \neg\texttt{fly} \quad \text{and}$$

$$\vdash_{CT4D} KB \wedge (\texttt{penguin} \vee \texttt{emu}) \Rightarrow \texttt{bird} \wedge \neg\texttt{fly}$$

are both valid derivations.

■

Unfortunately, it turns out that a very natural class of examples is not amenable to such an analysis, namely, examples that include *irrelevant information*. If we add the fact green to any of the premise sets in this example, none of the natural default conclusions (whether fly or $\neg$fly) can be reached. Once again, the cautious nature of the conditional logic, attractive from a logical point of view, prevents desirable default inferences, even when reasoning about the most normal states of affairs. With KB as above, for instance,

$$KB \wedge \texttt{bird} \wedge \texttt{green} \Rightarrow \texttt{fly}$$

is not CT4D-valid. This is due to the nonvalidity of the strengthening of the antecedent inference rule for CT4D. So even though bird $\Rightarrow$ fly holds, it need not be the case for the green birds. Logically,

$$\neg(\texttt{bird} \wedge \texttt{green} \Rightarrow \texttt{fly})$$

must remain consistent with KB, otherwise we would be able to infer not only that green birds fly, but also penguins, emus, dead birds, and so on. Indeed, the validity of strengthening would cause KB to be inconsistent (assuming the possibility of the antecedents).

In the case at hand, however, fly is deemed a reasonable conclusion because green is judged to be irrelevant to the fact that birds normally fly. Nothing in KB indicates green birds are exceptional with respect to flying ability, so we assume they are not. The reason adding penguin to KB altered the default conclusion is the explicit indication that penguins are exceptional.

The caution exhibited by conditional reasoning causes the "inverse qualification problem." In standard systems (such as default logic, etc.) a rule like "birds fly" cannot be stated as is. It must be *qualified* with exceptional conditions: "birds that are not penguins, not emus, not dead, and so on, fly." Conditionals are much more natural representations of default rules in this respect.

But to be reasoned with effectively, we must also describe all conditions that are *irrelevant*, or unexceptional with respect to a conditional. If we know $\texttt{bird} \Rightarrow \texttt{fly}$, we have to explicitly assert $\texttt{green} \wedge \texttt{bird} \Rightarrow \texttt{fly}$, $\neg\texttt{green} \wedge \texttt{bird} \Rightarrow \texttt{fly}$, and so on, if we wish to reason about such contexts.[2]

This *problem of irrelevance* plagues all approaches to default reasoning based on conditional logics. A number of solutions have been proposed to circumvent these difficulties. All consist of schemes whereby the conditional logics in question are augmented with extra-logical machinery with which appropriate default conclusions are drawn. Each of these solutions shares certain underlying intuitions. Very roughly, the schemes of Delgrande (1988), Pearl (1990) and Lehmann (1989) can be viewed as adding sentences of the form $\alpha \wedge \beta \Rightarrow \gamma$ to KB whenever $\alpha \Rightarrow \gamma$ is in KB and the result is consistent. A more semantical view is the following: these systems make states of affairs as normal as possible subject to the constraints of KB. Naturally, these systems are nonmonotonic as the added assumptions are defeasible. Adding constraints in the form of conditionals to KB can render certain assumptions inconsistent. Learning $\neg(\alpha \wedge \beta \Rightarrow \gamma)$ can render the previous assumption inconsistent, for example.

While each of these systems is extra-logical, we will extend the logics CT4 and CT4D to include the expressive power required to enforce these common assumptions logically, as sentences within the logical language. In order to capture these assumptions we require a notion of *conditional only knowing*, similar to the only knowing of (Levesque 1990). Briefly, if we know $\texttt{bird} \Rightarrow \texttt{fly}$ then to conclude $\texttt{bird} \wedge \texttt{green} \Rightarrow \texttt{fly}$ we must assume that at the most normal antecedent (bird) worlds *only* the consequent fly is known. In that way, all worlds consistent with bird and fly are deemed to be among the most normal bird-worlds, and, hence, any irrelevant properties can be discounted. We will define a new conditional connective $>$ and read $A > B$ as "At the most normal A-worlds, at most B is known," roughly the dual of $A \Rightarrow B$. As in the previous chapter, however, the conditional connective will not be primitive. Rather it is defined in terms of two unary modal operators, the standard modal connective $\Box$ that constrains truth at accessible worlds, and the less standard modality $\overleftarrow{\Box}$, that refers to truth at *inaccessible worlds*.

In Section 5.2 we present four bimodal logics of inaccessibility. Two of these logics extend CT4 and CT4D with inaccessibility, and these are further extended to ensure models are "full". Next, in Section 5.3, we discuss only knowing and inaccessibility, defining a notion of conditional only knowing and the connective $>$. In Section 5.4 the connective $>$ is used to axiomatize Pearl's (1990) system of default reasoning known as 1-entailment. This is achieved by defining a natural preference semantics in terms of the logic CO*, proving its equivalence to 1-entailment, and showing how the structure of preferred models can be captured concisely in the extended modal language. Finally, we investigate the qualities of relevance and irrelevance, distinguishing *statistical relevance* from *practical relevance*, and show how either notion may be modeled qualitatively within our logics.

The key results of this chapter are: Theorems 5.4 and 5.7, which show the completeness of our bimodal logics CO and CO*; Corollaries 5.24 and 5.25, which determine axiomatizations of 1-entailment and rational closure within our logics; and Theorem 5.27, which determines a nontrivial characterization of practical relevance.

5.1 Current Solutions to Irrelevance

As we've seen, conditional approaches to default reasoning must account for the cautious nature of conditional connectives. Before examining such models, let us see why traditional default reasoning

[2]Indeed, it is not hard to see that the list of irrelevant conditions is necessarily longer than the list of exceptional conditions, for every exceptional property P to $A \Rightarrow B$ has a corresponding irrelevant condition $\neg P$.

systems do not run into this problem. Imagine a circumscriptive default rule "birds fly" represented as

$$\texttt{bird} \wedge \neg\texttt{ab} \supset \texttt{fly}.$$

Because material implication is used, whenever `bird` is provable and `ab` is not provable (hence, by circumscribing it $\neg$`ab` is derived) `fly` will be derivable. Even when facts such as `green` or `raining` are known, the derivation remains valid, unless something like `green` $\supset$ `ab` is asserted, stating that green birds are known to be exceptional. Default logic is similar, for the default rule

$$\frac{\texttt{bird} : \texttt{fly}}{\texttt{fly}}$$

is applicable whenever `bird` is provable and $\neg$`fly` is not. The ease with which irrelevance is dealt comes at the expense of a compelling semantic account of default rules, and the ability to derive or reason about defaults.

5.1.1 Assumptions of Relevance and Normality

Delgrande's (1987; 1988) conditional logic N is very similar to CT4D and, unsurprisingly, fails to characterize logically all reasonable default inferences. To deal with this inadequacy, Delgrande (1988) specifies some extra-logical criteria to complement his logic and describe appropriate default conclusions. Conditional modus ponens is not valid in N, so from A and $A \Rightarrow B$ one cannot conclude B. Furthermore, one cannot conclude $A \wedge C \Rightarrow B$ from $A \Rightarrow B$, even when there is no reason to believe C is an exceptional property, this due to a lack of strengthening. Delgrande's default scheme attacks these problems separately by making two general assumptions.

The *assumption of normality* states that the world being modeled is as normal as can consistently be believed, given the constraints imposed by KB. So in the first case where $A \Rightarrow B$ and A are in KB, it is not logically necessary that B be believed since the actual world may be exceptional in this respect; but since B normally holds when KB does, this assumption ensures B is believed.

The *assumption of relevance* asserts that the only sentences known to affect the truth of a conditional sentence are those known to have an effect. In the second case, given A, C and $A \Rightarrow B$, it need not be the case that $A \wedge C \Rightarrow B$ holds, so (even with the assumption of normality) we cannot conclude B. However, relevance states that C should not affect the status of $A \Rightarrow B$, so $A \wedge C \Rightarrow B$ is inferred, and by normality so is B.

Delgrande defines two somewhat complementary approaches to default reasoning that embody these assumptions and shows that they determine the same notion of default consequence. We describe each in turn. A *default theory* T is a pair $\langle D, C\rangle$ where D is a set of simple conditional sentences or negations of such, and C is a set of propositional sentences (though Delgrande deals with first-order theories). Intuitively, D is the background or intensional knowledge (both default and necessary) and C forms the evidence or extensional knowledge relevant to the specific reasoning situation.

The first approach takes the assumption of relevance to be primary. Roughly, one extends conditionals in the manner described above, adding these strengthened sentences to the theory when this is consistent. This can lead to conflict however when, say, $Q \Rightarrow P$ and $R \Rightarrow \neg P$ are in the theory, stating Quakers are pacifists and Republicans are not. One can decide that Quaker Republicans (the demographically elusive class of "Nixons") are or are not pacifists, but not both. Since either choice would be arbitrary in the absence of further information, Delgrande's definition allows the addition of $\alpha \wedge \beta \Rightarrow \gamma$ based on $\alpha \Rightarrow \gamma$ only if no other "relevant" conditional (say $\beta \Rightarrow \neg\gamma$) contradicts this.

Definition 5.1 (Delgrande 1988) Sentence $\alpha \Rightarrow \gamma$ is *supported in* D if there is some β such that

(a) $\vdash_{CPL} \alpha \supset \beta$

(b) $D \vdash_N \beta \Rightarrow \gamma$

(c) If there is some β' such that $\vdash_{CPL} \alpha \supset \beta'$ and $D \vdash_N \neg(\beta' \Rightarrow \gamma)$, then $\vdash_{CPL} \beta \supset \beta'$.

So in the case of $Q \Rightarrow P$ and $R \Rightarrow \neg P$, neither of $R \wedge Q \Rightarrow P$ or $R \wedge Q \Rightarrow \neg P$ is supported. With the standard birds and penguins example, $\texttt{bird} \wedge \texttt{penguin} \Rightarrow \neg\texttt{fly}$ is derivable so

$$\texttt{yellow} \wedge \texttt{bird} \wedge \texttt{penguin} \Rightarrow \texttt{fly}$$

is not supported by $\texttt{bird} \Rightarrow \texttt{fly}$. However,

$$\texttt{yellow} \wedge \texttt{bird} \wedge \texttt{penguin} \Rightarrow \neg\texttt{fly}$$

is supported by $\texttt{bird} \wedge \texttt{penguin} \Rightarrow \neg\texttt{fly}$ since the "contradictory" default $\texttt{bird} \Rightarrow \texttt{fly}$ is less specific (or weaker) than the supporting default.

A *maximal default extension* of D, denoted $E(D)$, is a maximal, consistent set of sentences all of which are supported in D, and is defined to be unique. A *default conclusion* based on $T = \langle D, C\rangle$ is any $\alpha \in \mathbf{L}_{CPL}$ such that

$$E(D) \vdash_N C \Rightarrow \alpha.$$

In this manner the assumption of relevance is captured by using the extension of D, and normality is achieved by asking what normally follows from C in this context.

The second approach views normality as the primary assumption. Roughly, whenever $\alpha \Rightarrow \beta$ is believed $\alpha \supset \beta$ must be true at the most normal worlds (as is the case with CT4D). To make the world as normal as possible, as many of these sentences as possible should be added to C. Again, in the case of conflict, more specific conditionals take precedence, as conflict indicates exceptionality. So if we have $\texttt{bird} \Rightarrow \texttt{fly}$ and $\texttt{bird} \wedge \texttt{penguin} \Rightarrow \neg\texttt{fly}$ in D, with $\texttt{bird} \wedge \texttt{penguin}$ in C, both of the corresponding material implicants cannot be added to C consistently. Given context D, C must represent an exceptional state of affairs, so the exceptional, or more specific, sentence $\texttt{bird} \wedge \texttt{penguin} \supset \neg\texttt{fly}$ is added to reflect this.

Definition 5.2 (Delgrande 1988) Sentence $\alpha \supset \gamma$ is *contingently supported in* $T = \langle D, C\rangle$ iff

(a) $D \vdash_N \alpha \Rightarrow \gamma$

(b) $C \cup \{\alpha \supset \gamma\}$ is consistent

(c) If there is some α' such that $\vdash_{CPL} C \supset \alpha'$ and $D \vdash_N \neg(\alpha' \Rightarrow \gamma)$, then $\vdash_{CPL} \alpha \supset \alpha'$.

A *maximal contingent extension* of $T = \langle D, C\rangle$, denoted $E(C)$, is a maximal consistent set of sentences including C such that all sentences in $E(C) - C$ are contingently supported. A *default conclusion* based on T is any $\alpha \in \mathbf{L}_{CPL}$ such that

$$E(C) \vdash_{CPL} \alpha$$

for all contingent extensions $E(C)$. The assumption of normality is captured by assuming $\alpha \supset \beta$ holds for suitable $\alpha \Rightarrow \beta$. Relevance follows trivially for the same reason it does in traditional systems: once the material implicant is asserted, the default is "applicable" no matter what else is known.

The probabilistic reasoning system of Bacchus (1990) makes similar assumptions to deal with the weakness of logical derivability. Given some statistics about the proportion of birds that fly in the form $P(\texttt{fly}|\texttt{bird})$, it is often the case that we want to associate some conditional probability with fly given the class of, say, yellow birds, even though adequate statistics are unavailable. Like the assumption of relevance, Bacchus allows the nonmonotonic assumption that the probability associated with the narrower reference class is identical to that for the known class. Also similar to relevance, conflicts are resolved in favor of more specific reference classes. The assumption of normality in Bacchus's system can be thought of as embodied in the processes of *randomization* and *direct inference*. Given an individual bird about which nothing else is known, the subjective probability that it can fly can be assumed to be $P(\texttt{fly}|\texttt{bird})$. Hence, if birds normally fly (the conditional probability is sufficiently high) then a "random" bird is believed to fly with that degree of belief.[3]

5.1.2 System Z

The other main probabilistic approach we examined in Chapter 4 was Pearl's (1988) logic of ε-semantics. Given the equivalence to the simple fragment of CT4, it should not be surprising to learn that similar problems arise in this system. The assumption of normality is a lesser problem for ε-semantics as evidence C can be represented only implicitly as the antecedent of a simple conditional query $C \rightarrow \alpha$. In particular, evidence cannot form a distinguished part of a theory (which consists solely of simple conditionals). The problem of irrelevance remains.

To handle this difficulty Pearl (1990) proposes System Z, a natural method of ranking rules and arbitrary formulae. This ranking reflects the degree to which sentences are considered abnormal or exceptional. Imagine a rule r tolerated by theory T. Since it violates no constraints imposed by T, r is considered normal and is given low rank. Suppose, however, that r is not tolerable. If r is to be satisfied at some possible world, this world must falsify at least one rule in T. But the falsified rule must be verified at some world that is more normal, so the verifying world for r must be somewhat exceptional. Therefore r is considered exceptional and is assigned a higher rank.

Tolerance (see Definition 4.30) can be used to define a natural ordering on default rules by partitioning T as follows:

Definition 5.3 (Pearl 1990) $T_i = \{r : r \text{ is tolerated by } T - T_0 - T_1 - \cdots T_{i-1}\}, \qquad \text{for } i \geq 0.$

Assuming T is ε-consistent, this results in an ordered partition $T = T_0 \cup T_1 \cup \cdots T_n$. Now to each rule $r \in T$ we assign a rank (the *Z-ranking*), $Z(r) = i$ whenever $r \in T_i$. Roughly, but not precisely (see (Boutilier 1991a)), the idea is that lower ranked rules are more general, or have lower priority. Given this ranking, we can rank possible worlds according to the highest ranked rule they falsify:

$$Z(w) = \min\{n : w \text{ satisfies } r, \text{ for all } r \in T \text{ such that } Z(r) \geq n\}.$$

Again, lower ranked worlds are to be considered more normal.

Now, $\alpha \in \mathbf{L}_{CPL}$ can be ranked according to the lowest ranked world that satisfies it; that is

$$Z(\alpha) = \min\{Z(w) : w \models \alpha\}.$$

[3] Of course, this allows much finer distinctions than normality, which says that one either accepts or rejects the fact that a random bird flies. Note also that this a very loose characterization of the assumptions of Bacchus (1990). More exactly, the *expected values* of certain conditional probabilities (over various possible worlds) are used in these processes.

Given that lower ranked worlds are considered more normal, we can say that a default rule $\alpha \rightarrow \beta$ should hold iff the rank of $\alpha \wedge \beta$ is lower than that of $\alpha \wedge \neg\beta$. This leads to the following definition:

Definition 5.4 (Pearl 1990) Formula β is *1-entailed* by α with respect to T (written $\alpha \vdash_1 \beta$) iff $Z(\alpha \wedge \beta) < Z(\alpha \wedge \neg\beta)$ (where Z is defined with respect to T).

System Z solves the problem of irrelevance by assigning the lowest possible rank to valuations or worlds consistent with the theory T. 1-entailment is defined such that lower ranked worlds are considered more normal. Hence, if we know the rule `bird` $\rightarrow$ `fly`, worlds satisfying `green` and verifying this rule will be given the same rank as the most normal `bird`-worlds (supposing no other rules). Thus, the most normal `green` $\wedge$ `bird`-worlds satisfy `fly` and

$$\texttt{green} \wedge \texttt{bird} \vdash_1 \texttt{fly}.$$

We will examine the structure of Z-ranking in detail in Section 5.4.

5.1.3 Rational Closure

Given a set T of conditional assertions in the sense of (Kraus, Lehmann and Magidor 1990), the Lehmann and Magidor (1990) theory of rational relations maintains that any reasonable set of conditional assertions derivable (by default) from T should form a rational consequence relation. Unfortunately, Theorem 4.15 shows that the intersection of all rational relations extending T is $Cn_P(T)$, the set of preferential consequences of T. As discussed in Section 4.3.1, this is due to the fact that a set of simple conditionals cannot distinguish ranked models from one another. To alleviate this quandary the theory of *rational closure* is proposed (Lehmann 1989; Lehmann and Magidor 1990) whereby certain rational extensions of T are deemed preferable, and from those default conclusions arise. Rational closure also handles irrelevance.

Roughly, the rational closure of T attempts to extend T to include *supported* sentences, much like Delgrande's supported sentences, not sanctioned by preferential entailment. $\alpha \wedge \beta \mid\sim \gamma$ is supported in T if there is some $\alpha \mid\sim \gamma \in T$.[4] Unfortunately, Lehmann and Magidor show that the *perfect extensions* of T do not always exist. Rational closure is an approximation to such perfect extensions that is well-defined for any *admissible* theory T.[5] We provide a simple definition here and note that a number of alternate characterizations and representation results for rational closure, reflecting a number of different intuitions, are provided by Lehmann and Magidor (1990), and the reader is referred there for details.

Definition 5.5 (Lehmann 1989) Let T be a set of conditional assertions. The *degree* of formulae α is defined inductively as follows:

(a) $degree(\alpha) = 0$ iff there is no β such that $\alpha \vee \beta \mid\sim \neg\alpha \in Cn_P(T)$.

(b) $degree(\alpha) = i$ iff $degree(\alpha)$ is not less than i and $\alpha \vee \beta \mid\sim \neg\alpha \in Cn_P(T)$ only if $degree(\beta) < i$.

(c) $degree(\alpha) = \infty$ iff α is assigned no degree by the previous clauses.

[4] More precisely, $\alpha \mid\sim \gamma$ need only be in $Cn_P(T)$ (Lehmann and Magidor 1990) since any rational extension of T must include all preferential consequences of T. For simplicity, we assume T to be closed under preferential consequence in the remainder of this section when T is a set of assertions.

[5] In particular, if T is a *well-founded* preferential relation (which includes any finitely generated T), the closure is defined.

If a sentence $\alpha \vee \beta \mid\sim \neg\alpha$ is a preferential consequence of T, we say α is *more exceptional* than β. Whenever this assertion holds, the minimal $\alpha \vee \beta$-worlds must satisfy $\neg\alpha$. This implies any minimal α-world is less normal than any minimal β-world, hence α is more exceptional. Thus *degree*(α) represents the degree of exceptionality of α, given premises T. As with 1-entailment, we implicitly assign the lowest possible rank to states of affairs using the following definition of rational closure.

Definition 5.6 (Lehmann 1989) Let T be a set of conditional assertions. The *rational closure* of T, denoted $Cn_R(T)$, is defined as

$$\alpha \mid\sim \beta \in Cn_R(T) \text{ iff } degree(\alpha) < degree(\alpha \wedge \neg\beta) \text{ or } degree(\alpha) = \infty.$$

Lehmann and Magidor (1990) also characterize rational closure in terms of a preference ordering on the rational relations extending T, as well as in terms of rational model construction from preferential models of T.

5.1.4 Common Intuitions

In the three main conditional approaches to default reasoning we've examined, the emphasis appears to be the extension of conditionals to include irrelevant conditions in the antecedents, and then reasoning from the augmented knowledge base assuming normality. In Delgrande's model this approach is taken explicitly via syntactic considerations of the theory at hand. In Pearl's System Z irrelevance is handled semantically, extensions of conditionals implicitly determined by constraints on the preferred model of the theory. Rational closure is defined and represented in a number of ways and seems to reflect both semantic and syntactic considerations.

Given the striking similarity of the definitions of 1-entailment and rational closure, the following results of Goldszmidt and Pearl (1990) should not be surprising. For simple conditional sentences, we use $A \rightarrow B$ to represent rules in the case of ε-semantics and assertions in the case of nonmonotonic consequence relations.

Theorem 5.1 (Goldszmidt and Pearl 1990) *Let T be a finite set of simple conditionals. $Z(\alpha) = n$ iff degree(α) = n.*

Theorem 5.2 (Goldszmidt and Pearl 1990) *Let T be a finite set of simple conditionals. $\alpha \vdash_1 \beta$ with respect to T iff $\alpha \rightarrow \beta \in Cn_R(T)$.*

One noticeable aspect of both System Z and rational closure is their extra-logical nature. Both approaches can be viewed as imposing a preference relation (in the sense of Shoham (1988)) on the set of models of a particular theory and drawing conclusions true of the preferred structure. This preferred structure is the model of T in which all worlds are as normal as possible. A likeness exists to circumscription, where models of a theory that minimize the extent of designated predicates are preferred. Unlike circumscription, however, these systems do not come equipped with a sound and complete axiomatization.[6]

[6] At least, not a natural one. It could be argued that the set of supported sentences form an axiom system, similar to Delgrande's maximal default extensions $E(D)$. But this does not have the flavor or simplicity of the second-order circumscription axiom, nor will it generally be finite for logically infinite languages.

In the next three sections, we will develop a logic in which we may axiomatize, in a natural way, the assumptions made by rational closure and 1-entailment. In this way, we present a small set of axioms sufficient for reasoning with irrelevant information, just as the circumscriptive schema characterizes predicate-minimal models. However, just as circumscription requires the increased expressive power of second-order logic to capture the nature of minimality concisely, so too will we require a stronger language in which to model our notion of minimality.

Consider the modal logics we have been using. The modal connective $\Box$ allows us to express what is true or false at accessible worlds. A sentence $\Box A$ says A is true at all accessible worlds, while $\neg\Box A$ says A is false at some accessible world. Given our interpretation of accessibility, we can express truth at more normal worlds. We can also determine what holds at such worlds in a more sophisticated manner by use of the conditional connective. When we say $A \Rightarrow C$ we assert the truth of C at all most normal A-worlds.[7] Reverting to selection function notation, let us say that $f(w, \|A\|)$ is the set of most normal A-worlds for a particular CT4D-structure, relative to w, but understanding $f(w, \|A\|)$ to be defined in terms of R. To say $A \Rightarrow C$ is true is to say $f(w, \|A\|) \subseteq \|C\|$. Since A must be true at all selected (most normal) A-worlds, we can constrain this further as $f(w, \|A\|) \subseteq \|A \wedge C\|$.

The statement $A \Rightarrow C$ says nothing about the truth of $A \wedge B \Rightarrow C$. For instance, $f(w, \|A\|)$ could consist of only worlds satisfying $\neg B$, and the next-to-most normal set of A-worlds could all satisfy B and $\neg C$. Thus $A \wedge B \Rightarrow \neg C$ is conceivable. The approaches to irrelevance we've examined void this possibility whenever possible by saying worlds are as normal as possible. So if some model could consistently place these "extraordinary A-worlds" in $f(w, \|A\|)$, then such a model is preferred. In this preferred model $A \wedge B \Rightarrow C$ holds just because $A \Rightarrow C$ does. Since this is the case for arbitrary propositions α (not just B), in general $A \wedge \alpha \Rightarrow C$ can also be inferred. In summary, whenever $A \Rightarrow C$ holds, we know $f(w, \|A\|) \subseteq \|A \wedge C\|$, and we would like to insist that the converse hold as well, that $\|A \wedge C\| \subseteq f(w, \|A\|)$. When this is true we can derive $A \wedge \alpha \Rightarrow C$ for any α (provided $A \wedge C \wedge \alpha$ is true at some world in the structure).

Expressing the condition $f(w, \|A\|) \subseteq \|A \wedge C\|$ naturally in a modal logic is not feasible, for it requires an inordinate number of sentences of the form described (see Delgrande's (1988) assumption of relevance and Lehmann's (1989) definition of support). While $\|A \wedge C\| \subseteq f(w, \|A\|)$ can be summarized by ensuring certain facts are true at accessible (or more normal) worlds, that is by asserting $A \Rightarrow C$, this converse seems to require the ability to express truth at *inaccessible worlds* or *less normal* worlds. In particular, if we could just say that $A \wedge C$ must be false at all worlds less normal than $f(w, \|A\|)$ then all $A \wedge C$ worlds must be at least as normal as $f(w, \|A\|)$. But $f(w, \|A\|)$ is the set of *most* normal A-worlds, so $\|A \wedge C\| \subseteq f(w, \|A\|)$, or, when $A \Rightarrow C$ holds, $\|A \wedge C\| = f(w, \|A\|)$. This implies the desired strengthening of the conditional $A \wedge \alpha \Rightarrow C$ and irrelevance is obtained. We now turn our attention to logics where such truth at inaccessible worlds can be expressed.

5.2 Modal Logics of Inaccessibility

In this section we will present modal logics of inaccessibility with which conditions of irrelevance can be expressed concisely. In addition to the standard modal operator $\Box$, we add to the language the unary modal connective $\overleftarrow{\Box}$. The sentence $\overleftarrow{\Box}A$ will be true just when A is true at all inaccessible worlds, so just as $\Box A$ is read "A holds at all more normal worlds," $\overleftarrow{\Box}A$ is read "A holds at all

[7] This is a simplification since the Limit Assumption may be violated. However, it will serve the purpose of illustrating the main concepts.

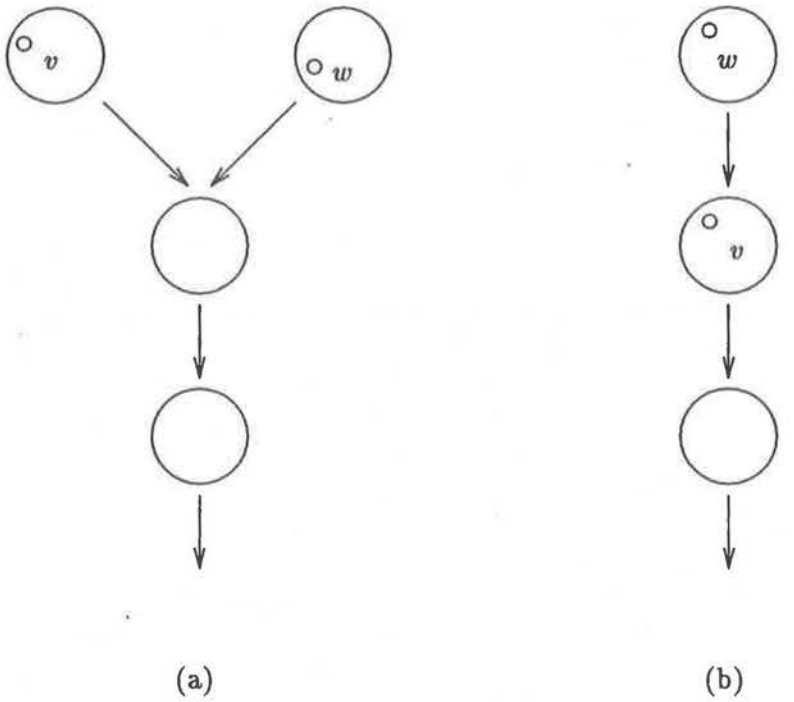

Figure 5.1: In the connected structure (a) w and v are mutually inaccessible. In the totally connected structure (b) this is impossible.

less normal worlds." The semantics for the bimodal logics will be based on the same Kripke structures used for (mono-) modal logics, with additional truth conditions for $\overleftarrow{\Box}$ defined on these models. Unlike the connectives in many multimodal systems, this new operator adds no "ontological baggage" to our semantic conception of normality.

We will start by extending CT4D with this expressive power, referring to the resulting logic as CT4DO, or CO for short.[8] In the standard modal case, CT4D is characterized by the class of connected structures. But CT4D is also characterized by the class of *totally connected* structures (Hughes and Cresswell 1984), where R is totally connected iff vRw or wRv for all v, w. To see this, imagine satisfaction of a formula α at any world w in a connected structure. The truth of α is determined solely by worlds accessible to w (the *submodel generated by* w). This submodel is of course totally connected, for connected structures can "branch backwards" only. Thus, from the point of view of a particular w, all alternatives are totally connected.[9] While connectedness and total connectedness are indistinguishable within our modal language $\mathbf{L}_M$, it is not hard to see the two kinds of models are vastly different when inaccessibility is considered. For instance, in a totally connected structure, if w cannot see v, v must see w. This need not be the case for merely connected structures, so long as no world sees both w and v. They may be situated on separate branches of the model and be incomparable (see Figure 5.1). Since our intuitions about CT4D tell us any two worlds should be comparable in the normality ordering, we take total connectedness to be a basic condition for CO.

[8]The "O" suffix stands for "only knowing", which can expressed in these logics. See Section 5.3 and Chapter 6.
[9]They form a total preorder, or a set of clusters that are totally ordered.

Our bimodal language $\mathbf{L}_B$ will be formed from a denumerable set $\mathbf{P}$ of propositional variables, together with the connectives $\neg$, $\supset$, $\Box$ and $\overleftarrow{\Box}$. The connectives $\wedge$, $\vee$ and $\equiv$ are defined in terms of these in the usual way (see Chapter 4).

Definition 5.7 A *CO-model* is a triple $M = \langle W, R, \varphi \rangle$, where W is a set (of possible worlds), R is a reflexive, transitive totally-connected binary relation on W (the accessibility relation), and φ maps $\mathbf{P}$ into 2^W ($\varphi(A)$ is the set of worlds where A is true).

We repeat the definition of satisfaction with the added clause for $\overleftarrow{\Box}$.

Definition 5.8 Let $M = \langle W, R, \varphi \rangle$ be a CO-model, with $w \in W$. The truth of a formula α at w in M (where $M \models_w \alpha$ means α is true at w) is defined inductively as:

1. $M \models_w \alpha$ iff $w \in \varphi(\alpha)$ for atomic sentence α.
2. $M \models_w \neg\alpha$ iff $M \not\models_w \alpha$.
3. $M \models_w \alpha \supset \beta$ iff $M \models_w \beta$ or $M \not\models_w \alpha$.
4. $M \models_w \Box\alpha$ iff for each v such that wRv, $M \models_v \alpha$.
5. $M \models_w \overleftarrow{\Box}\alpha$ iff for each v such that not wRv, $M \models_v \alpha$.

If $M \models_w \alpha$ we say that M *satisfies* α at w.

We now define several new connectives as follows:

(a) $\Diamond\alpha \equiv_{\text{df}} \neg\Box\neg\alpha$

(b) $\overleftarrow{\Diamond}\alpha \equiv_{\text{df}} \neg\overleftarrow{\Box}\neg\alpha$

(c) $\overleftrightarrow{\Box}\alpha \equiv_{\text{df}} \Box\alpha \wedge \overleftarrow{\Box}\alpha$

(d) $\overleftrightarrow{\Diamond}\alpha \equiv_{\text{df}} \Diamond\alpha \vee \overleftarrow{\Diamond}\alpha$

(Note that $\overleftrightarrow{\Diamond}\alpha$ can also be defined as $\neg\overleftrightarrow{\Box}\neg\alpha$.) It is easy to verify that these connectives have the following truth conditions:

(a) $M \models_w \Diamond\alpha$ iff for some v such that wRv, $M \models_v \alpha$.

(b) $M \models_w \overleftarrow{\Diamond}\alpha$ iff for some v such that not wRv, $M \models_v \alpha$.

(c) $M \models_w \overleftrightarrow{\Box}\alpha$ iff for all $v \in W$, $M \models_v \alpha$.

(d) $M \models_w \overleftrightarrow{\Diamond}\alpha$ iff for some $v \in W$, $M \models_v \alpha$.

These connectives have the obvious readings: $\Box A$ means "A is true at all more normal worlds"; $\Diamond A$ means "A is true at some more normal world"; $\overleftarrow{\Box} A$ means "A is true at all less normal worlds"; $\overleftarrow{\Diamond} A$ means "A is true at some less normal world"; $\overleftrightarrow{\Box} A$ means "A is true at all worlds, whether more or less normal"; finally, $\overleftrightarrow{\Diamond} A$ means "A is true at some world, whether more or less normal."[10] Validity is defined in a straightforward manner.

[10] "More normal" is used in this context in a nonstrict sense, meaning "at least as normal."

Definition 5.9 For any CO-model $M = \langle W, R, \varphi \rangle$, α *is valid on* M (written $M \models \alpha$) iff $M \models_w \alpha$ for each $w \in W$. A sentence α is *CO-valid* (written $\models_{CO} \alpha$) just when $M \models \alpha$ for every CO-model M.

Inaccessibility has been studied in AI by Levesque (1990), who presents a bimodal logic for only knowing that makes use (at least implicitly) of inaccessible worlds. The work of Humberstone (1983) deals explicitly with logics of inaccessibility and we take this as a starting point.

Humberstone presents an extension of the basic normal modal logic K in which inaccessibility can be expressed. The logic KO consists of K^2, the standard bimodal extension of K (see (Segerberg 1970)), but where the accessibility relations are constrained to be the complements of one another via the *Humberstone schemata* **H***.[11] That is, the following denumerable set of axiom schemata must be satisfied:

H* $D(\Box\alpha \wedge \overline{\Box}\beta) \supset B(\alpha \vee \beta)$,
where D is any sequence of the connectives $\Diamond$ and $\overleftarrow{\Diamond}$ having length ≥ 0, and B is any such sequence of $\Box$ and $\overline{\Box}$.

Fortunately, given the additional constraints imposed by total connectedness, we need not consider each instance of **H*** as axiomatic, but only one instance **H**.

Definition 5.10 The conditional logic CO is the smallest $S \subseteq \mathbf{L}_B$ such that S contains classical propositional logic and the following axiom schemata, and is closed under the following rules of inference:

K $\Box(A \supset B) \supset (\Box A \supset \Box B)$

K′ $\overline{\Box}(A \supset B) \supset (\overline{\Box}A \supset \overline{\Box}B)$

T $\Box A \supset A$

4 $\Box A \supset \Box\Box A$

S $A \supset \overline{\Box}\Diamond A$

H $\overleftarrow{\Diamond}(\Box A \wedge \overline{\Box}B) \supset \overline{\Box}(A \vee B)$

Nes From A infer $\overline{\Box}A$.

MP From $A \supset B$ and A infer B

Axiom **4** ensures transitivity of the accessibility relation R and **S** ensures total (or strong) connectedness. Reflexivity, usually associated with the axiom **T**, is derivable from total connectedness.

Definition 5.11 A sentence α is *provable* in CO (written $\vdash_{CO} \alpha$) iff $\alpha \in$ CO. α is *derivable* from a set $\Gamma \subseteq \mathbf{L}$ (written $\Gamma \vdash_{CO} \alpha$) if there is some finite subset $\{\alpha_1, \cdots, \alpha_n\}$ of Γ such that $\vdash_{CO} (\alpha_1 \wedge \cdots \wedge \alpha_r w \alpha$.

Lemma 5.3 *Any instance of a Humberstone schema* **H*** *is derivable in CO.*

This allows us to show the completeness of the logic CO.

[11] KO is referred to as $K^2 + (*)$ in (Humberstone 1983).

Theorem 5.4 *The system* CO *is characterized by the class of CO-models; that is,* $\vdash_{CO} \alpha$ *iff* $\models_{CO} \alpha$.

It is interesting to note that the derivation of **H*** in Lemma 5.3 uses only axiom **H** and the properties of K^2, except that **T** is used to derive instances of **H*** where either B or D is the empty sequence. This suggests that the logic KO does not need to have as axiomatic each instance of **H*** but only **H** and three instances dealing with these empty sequences.

Corollary 5.5 *The logic KO (Humberstone's (1983)* K^2 *+ (*)) has the following axiomatic basis.*

K $\Box(A \supset B) \supset (\Box A \supset \Box B)$

K' $\overline{\Box}(A \supset B) \supset (\overline{\Box} A \supset \overline{\Box} B)$

H $\overline{\Diamond}(\Box A \wedge \overline{\Box} B) \supset \overline{\Box}(A \vee B)$

H1 $(\Box A \wedge \overline{\Box} B) \supset (A \vee B)$

H2 $\overline{\Diamond}(\Box A \wedge \overline{\Box} B) \supset (A \vee B)$

H3 $(\Box A \wedge \overline{\Box} B) \supset \overline{\Box}(A \vee B)$

Nes *From* A *infer* $\overline{\Box} A$.

MP *From* $A \supset B$ *and* A *infer* B.

Thus we have the following result, not appearing in Humberstone's original paper:

Corollary 5.6 *KO is finitely axiomatizable.*

Often it is the case when ranking states of affairs according to normality we want to consider all logically possible worlds. No matter how implausible, each should be somehow ranked and should occur in our models. This consideration is important when dealing with the concept of only knowing and will be crucial in our account of irrelevance in the next section. For this reason we consider a class of CO-models in which all propositional truth valuations are represented by some possible world. Recall **P** is the set of atomic variables in our language.

Definition 5.12 Let $M = \langle W, R, \varphi \rangle$ be a Kripke model. For all $w \in W$, w^* is defined as the map from **P** into $\{0,1\}$ such that $w^*(A) = 1$ iff $w \in \varphi(A)$; in other words, w^* is the valuation associated with w.

Definition 5.13 A *CO*-model* is any $M = \langle W, R, \varphi \rangle$, such that M is a CO-model and

$$\{w^* : w \in W\} \supseteq \{f : f \text{ maps } \mathbf{P} \text{ into } \{0,1\}\}.$$

We can enforce this constraint axiomatically by appropriating the rather nonstandard schema of Levesque (1990), and show that CO* is sound and complete with respect to the class of CO*-models.

Definition 5.14 CO* is the smallest extension of CO closed under all rules of CO and containing the following axioms:

NB $\overleftarrow{\Box}\alpha \supset \neg\Box\alpha$ for all falsifiable propositional α.[12]

Theorem 5.7 *The system* CO* *is characterized by the class of CO*-models; that is,* $\vdash_{CO*} \alpha$ *iff* $\models_{CO*} \alpha$.

Clearly, any theorem of CO is also a theorem of CO*. However, CO* is a proper extension of CO since, for any satisfiable propositional sentence α, CO* has as a theorem $\overleftrightarrow{\Diamond}\alpha$ while $\neg\overleftrightarrow{\Diamond}\alpha$ is CO-satisfiable. Little else distinguishes the logics.

Finally, we define the bimodal extension of CT4 (or S4) that incorporates inaccessibility, though we will not have occasion to use this logic until the next chapter. The logic CT4O is based on the class of CT4-models, now also referred to as CT4O-models. In particular, a CT4O-model is a reflexive, transitive Kripke model with the additional truth conditions for $\overleftarrow{\Box}$ understood.

Definition 5.15 The conditional logic CT4O is the smallest $S \subseteq \mathbf{L}_B$ such that S contains classical propositional logic and the following axiom schemata, and is closed under the following rules of inference:

K $\Box(A \supset B) \supset (\Box A \supset \Box B)$

K′ $\overleftarrow{\Box}(A \supset B) \supset (\overleftarrow{\Box}A \supset \overleftarrow{\Box}B)$

T $\Box A \supset A$

4 $\Box A \supset \Box\Box A$

H $\overleftrightarrow{\Diamond}(\Box A \wedge \overleftarrow{\Box}B) \supset \overleftrightarrow{\Box}(A \vee B)$

Nes From A infer $\overleftrightarrow{\Box}A$.

MP From $A \supset B$ and A infer B.

Lemma 5.8 *Any instance of a Humberstone schema* **H*** *is derivable in CT4O.*

Theorem 5.9 *The system* CT4O *is characterized by the class of CT4O-models; that is,* $\vdash_{CT4O} \alpha$ *iff* $\models_{CT4O} \alpha$.

Definition 5.16 A *CT4O*-model* is any $M = \langle W, R, \varphi\rangle$, such that M is a CT4O-model and

$$\{w^* : w \in W\} \supseteq \{f : f \text{ maps } \mathbf{P} \text{ into } \{0,1\}\}.$$

Definition 5.17 CT4O* is the smallest extension of CT4O closed under all rules of CT4O and containing the following axioms:

NB $\overleftarrow{\Box}\alpha \supset \neg\Box\alpha$ for all falsifiable propositional α.

Theorem 5.10 *The system* CT4O* *is characterized by the class of CT4O*-models; that is,* $\vdash_{CT4O*}$ α *iff* $\models_{CT4O*} \alpha$.

[12] Alternatively, we could use $\overleftrightarrow{\Diamond}\alpha$ for all satisfiable α.

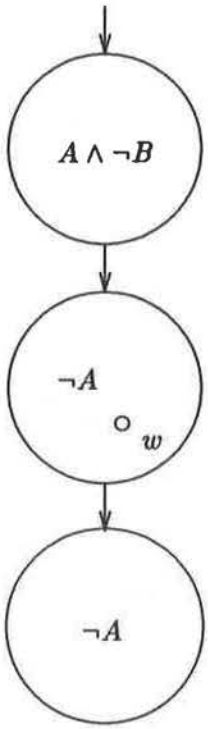

Figure 5.2: The difference between the local and global definitions of $\Rightarrow$. On the earlier (local) definition w satisfies $A \not\Rightarrow B$ since it can see no A-worlds. On the new (global) definition the truth of $A \Rightarrow B$ is determined by the entire structure. Since the minimal A-worlds satisfy $\neg B$, $A \Rightarrow B$ is false at w (and at all other worlds) even though w cannot see the minimal A-worlds. This reflects the use of $\overset{\leftrightarrow}{\Box}$ and $\overset{\leftrightarrow}{\Diamond}$ in the new definition instead of $\Box$ and $\Diamond$.

5.3 Conditional Only Knowing

Before returning to the problem of irrelevance, we will redefine the connective $\not\Rightarrow$ so that it more closely resembles the nonmonotonic consequence relations $\mid\!\sim$ of (Kraus, Lehmann and Magidor 1990) and their global nature (see Chapter 4). We use $\Rightarrow$ to denote the new connective as defined in CO.

Definition 5.18 The connective $\Rightarrow$ is defined in $\mathbf{L}_B$ as

$$A \Rightarrow B \equiv_{\mathrm{df}} \overset{\leftrightarrow}{\Box}\neg A \vee \overset{\leftrightarrow}{\Diamond}(A \wedge \Box(A \supset B))$$

Proposition 5.11 *Let M be a CO-model. Then $M \models A \Rightarrow B$ iff $M \models_w A \Rightarrow B$ for some w .*

Figure 5.2 shows a CO-model at which a world satisfies a conditional according to the definition given in Chapter 4, but not according to the new global definition. On this definition of $\Rightarrow$, if $A \Rightarrow B$ holds at any world in a model then it holds at all worlds. Previously, $A \not\Rightarrow B$ could hold "vacuously" if there were no *accessible* worlds satisfying A (i.e. $\Box\neg A$). While this is in accord with an epistemic reading of the relation R, it does not conform to our normative interpretation. Accessible worlds

are not meant to represent states of affairs considered possible by some world, but those states that are more normal. It is entirely unreasonable to expect only more normal worlds to determine which normative statements we take to be true. Worlds that are exceptional should also play a role in such deliberations. On this definition conditionals are satisfied vacuously only when no world, *accessible or inaccessible*, satisfies the antecedent. This is the perspective reflected in Kraus, Lehmann and Magidor's (1990) definition of $\mid\!\sim$. One advantage of our modal definition of $\Rightarrow$ is that the connective $\overleftarrow{\Box}$ allows us to define the truth of $\Rightarrow$ at individual worlds, whereas the truth conditions of $\mid\!\sim$ can only be defined with respect to entire structures. We note, however, that no differences of great consequence exist between the properties of $\Rightarrow$ and $\Rightarrow$. We take $\Rightarrow$ (and hence CO) to be the connective (and logic) of choice for representing normative implication.

To deal with the problem of irrelevance, we want to ensure that the set of most normal A-worlds, $f(w, \|A\|)$, is as large as possible given the constraints of a theory, for any sentence A. When $A \Rightarrow C$ holds it must be that no $\neg C$-worlds are in $f(w, \|A\|)$. So $\|\neg C\| \cap f(w, \|A\|) = \emptyset$. If we consider the set of worlds $f(w, \|A\|)$ to represent an agent's state of knowledge then we can say an agent *knows* C, that any (epistemically) possible world must satisfy C.[13] In a sense, $A \Rightarrow C$ expresses a *conditional* form of knowledge, for we may read this as "If A were the case, normally an agent would know C." Of course, at the most normal A-worlds A must hold, so $A \Rightarrow C$ says even more: "If A were the case, normally an agent would know $A \wedge C$" (see Figure 5.3).

Our considerations in the previous section suggest that we should also insist an agent normally know *nothing more* than $A \wedge C$ when A holds. For an agent to know more than $A \wedge C$ it must exclude certain possible worlds from its state of knowledge, worlds that make this additional knowledge false. If an agent knows no more than $A \wedge C$ then all $A \wedge C$-worlds must be considered epistemically possible. Suppose we define a new connective $A > C$ that holds just when $\|A \wedge C\| \subseteq f(w, \|A\|)$. When $A > C$ is true it is legitimate to say "If A were the case, normally an agent would *know at most* $A \wedge C$" (see Figure 5.4). It is important to keep in mind, however, that this type of "knowledge" is conditional on accepting the most normal states of affairs satisfying A as epistemically possible.

This is analogous to Levesque's (1990) concepts of *knowing at least* and *knowing at most*, except conditionalized on a particular antecedent A. In general, given $A \Rightarrow C$ we'd like to assume an agent normally *only knows* $A \wedge C$ when A is true; that is, the agent knows at least $A \wedge C$ (i.e. $A \Rightarrow C$) and knows at most $A \wedge C$ (i.e. $A > C$). If this is the case, then $f(w, \|A\|) = \|A \wedge C\|$ and assumptions of irrelevance are forthcoming.

We now must define the connective $>$ in terms of the primitive operators $\Box$ and $\overleftarrow{\Box}$. The sentence $A > C$ says at most $A \wedge C$ is known at the set of most normal A-worlds. For this to be true in the principal case we must insist at least that $\overleftarrow{\Diamond} A$ be true, for otherwise there are no A-worlds, normal or not, and all sentences are trivially satisfied by the (empty) set $f(w, \|A\|)$; thus, everything is known. In the case where C is a falsehood, we will let $A > C$ hold only when $\neg\overleftarrow{\Diamond} A$ holds — everything is known at least (since $A \Rightarrow C$) so everything will be known at most.

If there is some A-world w such that $A \wedge C$ holds at some inaccessible (less normal) world v, then $A > C$ cannot hold. If this is the case w is more normal than v, so there is some $A \wedge C$-world (v) that is not among the most normal A-worlds. So a requirement for $A > C$ is

$$\overleftarrow{\Box}(A \supset \overleftarrow{\Box}(A \supset \neg C)).$$

[13] For details regarding logics of knowledge see, e.g., (Hintikka 1962; McArthur 1988; Levesque 1986b). Further discussion of only knowing is given in Chapter 6.

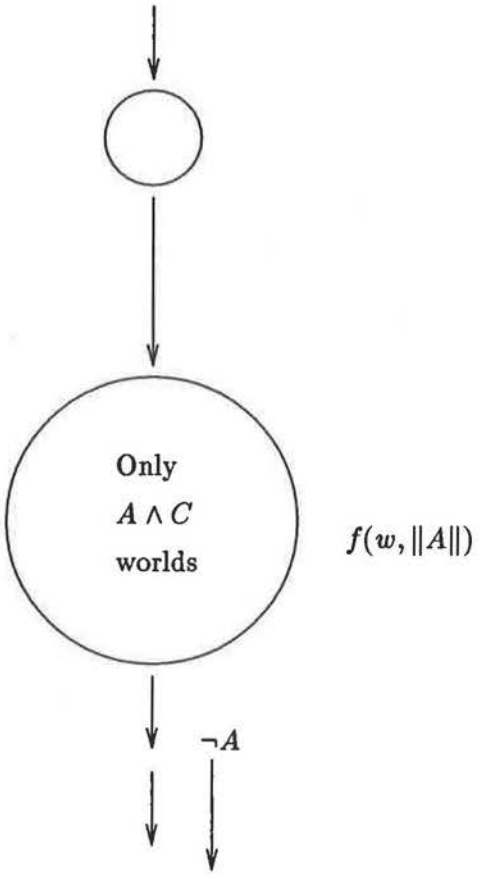

Figure 5.3: When $A \Rightarrow C$ all worlds in $f(w, \|A\|)$ satisfy $A \wedge C$.

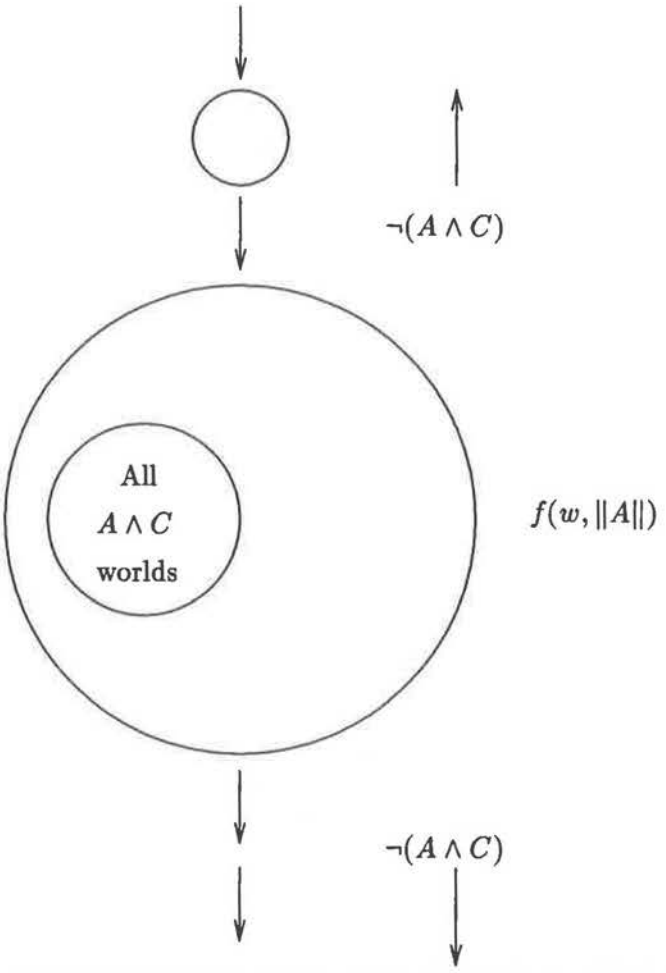

Figure 5.4: When $A > C$ all $A \wedge C$-worlds are in $f(w, \|A\|)$.

Furthermore, since all $A \wedge C$-worlds are to be included in $f(w, \|A\|)$, every A-world should see some $A \wedge C$-world (in fact, all $A \wedge C$-worlds). So we insist that

$$\vec{\Box}(A \supset \Diamond(A \wedge C)).$$

Putting these together we arrive at the following definition.

Definition 5.19 The connective $>$ is defined in $\mathbf{L}_B$ as

$$A > B \equiv_{\mathrm{df}} \vec{\Box}(A \supset (\vec{\Box}(A \supset \neg B) \wedge \Diamond(A \wedge B))) \wedge \vec{\Diamond} A.$$

Proposition 5.12 *Let $M = \langle W, R, \varphi \rangle$ be a CO-model. Then for any w, $M \models_w A > B$ iff $M \models A > B$. If $M \models A > B$ then there exists a cluster $\mathcal{C}$ in W such that $\|A \wedge B\| \subseteq \mathcal{C}$ and no A-world is strictly more normal than $\mathcal{C}$. Furthermore, $\|A \wedge B\|$ must be nonempty.*

Thus $A > B$ ensures all $A \wedge B$-worlds are among the most normal A-worlds.

Proposition 5.13 *Let $M = \langle W, R, \varphi \rangle$ be a CO-model. If $M \models A > B$ and $M \models A \Rightarrow B$ then there exists a cluster $\mathcal{C}$ in W such that $\|A \wedge B\| \cup \mathcal{N} = \mathcal{C}$, where each world in $\mathcal{N}$ satisfies $\neg A$, and no A-world is strictly more normal than $\mathcal{C}$.*

Given $A \Rightarrow B$ and $A > B$ as premises one can infer that all $A \wedge B$-worlds are most normal A-worlds, and conversely that all most normal A-worlds satisfy $A \wedge B$. This means if any proposition α consistent with $A \wedge B$ is possible (represented in the structure), then $A \wedge \alpha \Rightarrow B$ is derivable.

Proposition 5.14 $A \Rightarrow B, A > B \vdash_{CO} \vec{\Diamond}(A \wedge B \wedge \alpha) \supset (A \wedge \alpha \Rightarrow B)$.

Thus, determining the irrelevance of condition α to $A \Rightarrow B$ depends on the existence of some state of affairs where all three propositions are satisfied. For example, given $\mathtt{bird} \Rightarrow \mathtt{fly}$ one cannot conclude $\mathtt{bird} \wedge \mathtt{green} \Rightarrow \mathtt{fly}$ unless the existence of a $\mathtt{bird} \wedge \mathtt{green} \wedge \mathtt{fly}$-world is assured. However, such assurance is guaranteed in CO* as all valuations are represented.

Proposition 5.15 *Let $A \wedge B \wedge \alpha$ be propositionally satisfiable. Then*

$$A \Rightarrow B, A > B \vdash_{CO*} A \wedge \alpha \Rightarrow B$$

While the logic CO* seems able to express the concept of irrelevance, it is not clear how a default reasoner should proceed given such a logic and a set of facts KB. A modest proposal is simply to assert $A > B$ for each $A \Rightarrow B$ in KB, so long as the result is consistent. This works on a wide variety of examples.

Example 5.2 Let $KB = \{\mathtt{bird} \Rightarrow \mathtt{fly}\}$. If we assert $\mathtt{bird} > \mathtt{fly}$, we can derive conditionals such as

$$\mathtt{bird} \wedge \mathtt{green} \Rightarrow \mathtt{fly}.$$

If $\mathtt{penguin} \Rightarrow \neg\mathtt{fly}$ and $\Box(\mathtt{penguin} \supset \mathtt{bird})$ are added to KB, $\mathtt{bird} > \mathtt{fly}$ is no longer consistent. However,

$$\mathtt{bird} \wedge \neg\mathtt{penguin} \Rightarrow \mathtt{fly} \quad \text{and} \quad \mathtt{bird} \wedge \mathtt{penguin} \Rightarrow \neg\mathtt{fly}$$

are both derivable and it is consistent to assert

$$\texttt{bird} \wedge \neg\texttt{penguin} > \texttt{fly} \quad \text{and} \quad \texttt{bird} \wedge \texttt{penguin} > \neg\texttt{fly}.$$

Adding these to KB, we obtain the following theorems:

$$(KB \wedge \texttt{bird}) \Rightarrow \texttt{fly}$$

$$(KB \wedge \texttt{bird} \wedge \texttt{green}) \Rightarrow \texttt{fly}$$

$$(KB \wedge \texttt{penguin}) \Rightarrow \neg\texttt{fly}$$

$$(KB \wedge \texttt{penguin} \wedge \texttt{green}) \Rightarrow \neg\texttt{fly}$$

■

Such an approach, however, has limitations.

Example 5.3 Consider a KB of two independent conditionals $A \Rightarrow B$ and $C \Rightarrow D$, where A, B, C and D are distinct variables. In this case, it is inconsistent to assert both $A > B$ and $C > D$. If each of the four sentences hold then all $A \wedge B$-worlds are most normal A-worlds, including $C \wedge D$-worlds and $C \wedge \neg D$-worlds. From $C > D$ we can infer that all $C \wedge D$-worlds are equally normal and must be just as normal as the aforementioned $C \wedge \neg D$-worlds. This contradicts $C \Rightarrow D$.

■

In this case it is not clear what "extendible" conditionals of interest are derivable from such a KB. Thus, the use of the connective $>$ for dealing with irrelevance requires further investigation. Another simple proposal, that adequately handles this KB, can be described as follows: since the material counterparts of these sentences, $A \supset B$ and $C \supset D$, must be normally true (that is $\top \Rightarrow (A \supset B \wedge C \supset D)$), it should be the case that $\top > (A \supset B \wedge C \supset D)$ holds as well. While this scheme seems quite syntactic, it has a more detailed semantic counterpart. Extending this idea, we will show that the connective $>$ is capable of representing a certain form of default reasoning, namely 1-entailment, or rational closure.

5.4 An Axiomatization of 1-entailment and Rational Closure

5.4.1 A Simple Preference Relation

Both rational closure and 1-entailment intuitively determine a preference for models in which worlds are ranked as low as possible, or where worlds are considered as normal as possible. We will now formalize this notion of preference explicitly by defining a preference relation $\leq$ on the class of CO-models. If M_1 and M_2 are two CO-models then following Shoham (1988) we say M_1 is *as preferable as* M_2 when $M_1 \leq M_2$. If $M_1 \leq M_2$ but not $M_2 \leq M_1$ then we say M_1 is *preferred to* M_2, writing $M_1 < M_2$. In standard fashion, the conclusions derivable nonmonotonically from a theory T are just those sentences true in all preferred models of T.

Recall, a CO-model consists of a totally ordered set of clusters of possible worlds. For illustrative purposes, imagine this totally ordered set forms a discrete, well-founded order, so there is a minimum element (cluster), labeled 0, which is the set of most normal worlds, and a sequence of increasingly less normal clusters that we label 1, 2, 3 and so on. In general, CO-models will not

be so well-behaved, but this assumption will clarify the discussion.[14] We will also assume in our discussion of various models that W and φ are fixed, so comparing models is a matter of shuffling around a fixed set of states of affairs in a normality ordering.

The question to answer is the following: when should one CO-model be preferred to another? Intuitively, in a preferred model, each world should be as normal as possible. Obviously, the degree of normality (or rank) of any world w can be viewed as the label of the cluster to which it belongs, and wRv just when the rank of w is greater than or equal to that of v. Hence, the most preferred CO-model assigns a rank of 0 to each world, making each equally (and most) normal. Of course, such a structure is an S5-model and is maximally ignorant, satisfying no interesting conditionals $A \Rightarrow B$ (see Section 4.2). Therefore, if we have a theory with some conditional $A \Rightarrow B$, no $A \wedge \neg B$-world can have rank 0, for they would then be among the most normal A-worlds.

It should be clear that $M_1 \leq M_2$ ought to hold just when each world in M_1 has rank at least as low as its rank in M_2. We can say world w is more normal in M_1 than in M_2 if it has been removed from its cluster and put in a lower cluster, that is, given a lower rank. In terms of accessibility, this means there exists some v mutually accessible to w (in the cluster of w) in M_2 such that v sees w in M_1 but w does not see v. Of course, if w forms a singleton cluster in M_2, this is meaningless, but for technical reasons we say w is more normal in M_1 in this case.

Now, if all worlds are to have rank at least as low in M_1 as in M_2, then every world ought to see at least as many worlds in M_1 as in M_2. To see this suppose w sees fewer worlds in M_1 than in M_2 (assuming w is not more normal in M_1). Then some v accessible to w in M_2 is no longer accessible in M_1. This means v has become less normal in M_1, and M_1 should not be preferred. So if M_1 is preferable to M_2 we say each world that is not more normal in M_1 than in M_2 should see at least as many worlds (in the sense of set inclusion) in M_1 as M_2.

More formally, assume $M_1 = \langle W, R_1, \varphi \rangle$ and $M_2 = \langle W, R_2, \varphi \rangle$ are CO-models. We are interested in the preferred models of simple conditional theories in order to characterize 1-entailment.

Definition 5.20 $w \in W$ is *more normal in* M_1 than in M_2 (written $N(w, M_1, M_2)$) iff there is some $v \in W$ such that vR_1w, wR_1v, and not vR_2w, or there is no v such that wR_1v and vR_1w $(w \neq v)$.

Definition 5.21 M_1 is *as preferable as* M_2 (written $M_1 \leq M_2$) iff for all $w \in W$, $N(w, M_1, M_2)$ is false only if $\{v : wR_2v\} \subseteq \{v : wR_1v\}$.

M_1 is *preferred to* M_2 (written $M_1 < M_2$) iff $M_1 \leq M_2$ and not $M_2 \leq M_1$.

Definition 5.22 Let $T \subseteq \mathbf{L}$ be a set of simple conditionals. M is a *preferred model* of T iff $M \models T$ and for all M' such that $M' \models T$, $M' \not\leq M$.

Definition 5.23 α is a *default conclusion* based on T (written $T \models_{\leq} \alpha$) iff $M \models \alpha$ for each minimal model M of T.

This definition compares only models that agree on possible worlds, hence W and φ must agree. If we are considering only CO*-models, this makes little difference (as long as we "rename" worlds appropriately). We will see below that duplicate worlds (having the same induced valuation) have no effect on default conclusions. In the case of CO-models, we will obtain sets of minimal

[14] In fact, the models constructed for simple conditional theories in the next subsection will have precisely this structure, and for finite theories these will be finite sets (of clusters).

models for each set of worlds W. If we wish to allow different sets of worlds to be comparable, we may "relativize" the preference criteria (below) to those worlds the models have in common. However, in the subsequent developments CO*-models will be of primary interest. As well, T should consist only of conditional sentences since a model will in general not satisfy any non-tautologous propositions (they will not be true at all worlds in the model, unlike conditionals). If we wish to consider propositional facts as well, our preference relation can be extended easily to "models" of T consisting of pairs $\langle M, w\rangle$ where w is some world in structure M. Similarly, default conclusion α should be a conditional sentence. Extending the relation to pairs $\langle M, w\rangle$ and a (finite) theory T including propositional facts, we can then conclude any α such that

$$T \models_{\leq} T \Rightarrow \alpha.$$

Such α will include all conditional default conclusions sanctioned by Definition 5.23, as well as propositional sentences that normally follow from T.

In what follows we assume T to be a simple conditional theory, ignoring propositional evidence. The types of inference allowed by this preference criterion will become apparent as we compare it to 1-entailment and rational closure.

5.4.2 Equivalence to 1-entailment

We now assume a language generated by a finite set of propositional variables, and consider only CO*-models based on this language, assumptions made by System Z (where all valuations are ranked).[15] Given a set of simple conditionals T, Z-ranking provides a unique "preferred model" for T from which 1-entailment is derived. For such a fixed T we can define a corresponding CO*-model.

Definition 5.24 $Z_T = \langle W, R, \varphi\rangle$ is the CO*-model where wRv iff $Z(w^*) \geq Z(v^*)$.[16]

Corollary 5.16 *Z_T is a CO*-model.*

Corollary 5.17 $Z_T \models_{CO*} T$.

Corollary 5.18 $Z_T \models_{CO*} \alpha \Rightarrow \beta$ *iff* $\alpha \vdash_1 \beta$ *whenever* α *is satisfiable.*

We can also show that Z_T is the (unique) $\leq$-minimal model of T. Let M be an arbitrary CO*-model.

Lemma 5.19 *If* $M \models_{CO*} T$, *then* $Z_T \leq M$.

Lemma 5.20 *If* $M \models_{CO*} T$ *and* $M \leq Z_T$, *then* $M = Z_T$.

Theorem 5.21 $T \models_{\leq} \alpha \Rightarrow \beta$ *iff* $\alpha \vdash_1 \beta$ *with respect to* T.

This means that the minimal Z-ranking of worlds corresponds to a theory-dependent instance of the more general preferential ranking of CO*-models. Thus, the preference relation $\leq$ seems

[15] While all such CO*-models need not be finite, the nature of the theories and preferred models we consider ensure nothing is lost if we assume all models are finite, or even if we assume all models have unique worlds corresponding to the finite set of propositional valuations.

[16] Recall that w^* is the valuation associated with w.

to respect the intuition that worlds should be given the lowest possible rank. Furthermore, the explicit nature of Z-ranking allows us to describe the exact nature of the (unique) preferred model Z_T. In particular, if T is ε-consistent and is partitioned as $T_0, \cdots, T_n$, then Z_T consists of $n+2$ clusters of mutually accessible (or equally normal) worlds; cluster 0 consists of all worlds of rank 0, cluster 1 consists of all worlds of rank 1, and so on, with the most exceptional worlds being those of rank $n+1$.

Example 5.4 Let T be the theory consisting of the default rules $B \Rightarrow F$, $P \Rightarrow B$ and $P \Rightarrow \neg F$ (the usual "bird and penguin" example). The rule $B \Rightarrow F$ has rank 0, since the valuation $\{B, F, \neg P\}$ verifies the rule without violating the other two. The other rules cannot be verified in the presence of $B \Rightarrow F$ (since either B or $\neg F$ must be made true). Hence, they are assigned rank 1. Worlds that violate no rules are assigned rank 0. All such worlds satisfy both $B \supset F$ and $\neg P$. So the "most normal" worlds are exactly those at which birds fly (if they exist), but no penguins exist. The rank 1 worlds are the next most normal, and they consist of the worlds where birds are penguins and they do not fly. Finally, the least normal worlds are the rank 2 worlds, those that have penguins that either can fly or are not birds.

■

Since the preferred model of T is unique, we can capture the exact structure of Z_T using sentences in the logic CO containing the connective $>$ as worlds in each cluster can be characterized by the rules they violate. For instance, the worlds in cluster 0 are exactly those that satisfy all rules (or falsify no rules) in T. Suppose $R_0^\wedge$ stands for the conjunction of the material counterparts $\alpha \supset \beta$ of all rules in T. If we assert $\top > R_0^\wedge$, then any model of T that satisfies this sentence will have a cluster of most normal worlds consisting of exactly all worlds of rank 0. To see this, recall that $\top > R_0^\wedge$ means that at the most normal $\top$-worlds (i.e. the most normal worlds) at most $R_0^\wedge$ is known. Thus any world of rank 0 (i.e. satisfying $R_0^\wedge$) must be most normal. But no other worlds can be considered as normal, for any world of rank greater than 0 must falsify *some* rule, violating T.

Similarly, if $R_1^\wedge$ stands for the conjunction of the material counterparts of rules in $T - T_0$, asserting $\neg R_0^\wedge > R_1^\wedge$ assures that the most normal worlds falsifying some rule will consist exactly of those worlds satisfying rules of rank 1 or greater; in other words, the cluster of next-to-most normal worlds will be the set of worlds of rank 1.

Let T be a finite set of conditionals, partitioned as $T_0, T_1, \cdots, T_n$.

Definition 5.25 Let $R_{-1}^\wedge \equiv_{\text{df}} \bot$. For $0 \leq i \leq n+1$, define $R_i^\wedge$ as

$$R_i^\wedge \equiv_{\text{df}} \bigwedge \{\alpha \supset \beta : \alpha \Rightarrow \beta \in T - T_0 - \cdots T_{i-1}\}.$$

We assume $\bigwedge \emptyset \equiv_{\text{df}} \top$; therefore, $R_{n+1}^\wedge \equiv \top$.

Definition 5.26 For theory T as above, the *closure* of T is defined as

$$Cl(T) = T \cup \{\neg R_i^\wedge > R_{i+1}^\wedge \ : \ -1 \leq i \leq n\}.$$

The closure of T constrains models of T to have exactly the structure we desire.

Lemma 5.22 *$Cl(T)$ is consistent iff T is, and is "categorical" in the sense that there is a unique CO*-model that satisfies it, namely Z_T.*[17]

Theorem 5.23 *$Cl(T) \vdash_{CO*} \alpha \Rightarrow \beta$ iff $T \models_{\leq} \alpha \Rightarrow \beta$.*

Corollary 5.24 *$Cl(T) \vdash_{CO*} \alpha \Rightarrow \beta$ iff $\alpha \vdash_1 \beta$ with respect to T.*

Corollary 5.25 *$Cl(T) \vdash_{CO*} \alpha \Rightarrow \beta$ iff $\alpha \mathrel{|\!\sim} \beta \in Cn_R(T)$.*

Just as the (second-order) circumscriptive axiom applied to a theory T closes that theory to correspond to (predicate-) minimal models, so too does this closure correspond to our notion of minimality. Theorem 5.23 shows that $Cl(T)$ can be regarded as an axiomatization of the notion of preference described above, and of the implicit preference ordering determined by System Z. Hence, the types of conclusions sanctioned by 1-entailment (see (Pearl 1990) for details) are also determined by this form of closure. This implies, given the results of (Goldszmidt and Pearl 1990), that $Cl(T)$ determines the same consequence relation as that of rational closure (Lehmann 1989). Notice, however, that the size and number of additional axioms required to form the closure in our case is not large (roughly bounded by twice the number of original defaults), in contrast with the inordinately large number of sentences required by, say, the scheme of Delgrande (1988).

We provide an example, due to Pearl (1990), illustrating the types of conclusions sanctioned (and not) by System Z, rational closure and $Cl(T)$.

Example 5.5 Let T contain the following conditionals:

$$P \Rightarrow B,\ B \Rightarrow F,\ P \Rightarrow \neg F,\ P \Rightarrow A,\ B \Rightarrow W,\ F \Rightarrow M,$$

where we read P, B, F, A, W, and M as "penguin," "bird," "fly," "antarctic-dweller," "has-wings," and "mobile," respectively. Deductive consequence in CT4D (and CO*) allows the derivation of the following from T:

$$B \wedge P \Rightarrow \neg F,\ F \Rightarrow \neg P,\ B \Rightarrow \neg P,\ P \wedge A \Rightarrow B.$$

Using the closure of T, we can derive from $Cl(T)$ in CO* the further consequences:

$$\neg B \Rightarrow \neg P,\ \neg F \Rightarrow \neg B,\ B \Rightarrow M,\ \neg M \Rightarrow \neg B,\ P \wedge \neg W \Rightarrow B.$$

Underivable facts include (appropriately) $F \Rightarrow B$ and $\neg F \Rightarrow P$. Unfortunately, other underivable conclusions seem intuitively desirable, in particular

$$P \Rightarrow W, \quad \text{and} \quad P \wedge \neg A \Rightarrow B.$$

We briefly discuss these difficulties in the closing section of this chapter.

■

[17] Again, it is categorical if we ignore "duplicate" worlds (associated with the same valuation), which contribute nothing to the truth or falsity of formulae in Z_T.

5.5 Statistical and Practical Relevance

We have discussed the problem of irrelevance and various solutions that lead to a reasonable notion of default inference, yet little has been said about what it means for one proposition to be relevant to another. Intuitively, when we nonmonotonically assume B is irrelevant to C given A, we intend that $A \wedge B \Rightarrow C$ holds when $A \Rightarrow C$ does; in other words, B does not affect our willingness to accept C when A is known. But few formal studies of the concept of irrelevance have been undertaken.

An obvious set of criteria for deciding the relevance of some data to other information in a quantitative setting is based on *conditional independence*. Suppose we have background evidence (or *context*) E and want to know if proposition A is relevant to B in context E. We should judge A to be relevant iff it can affect our degree of belief in B given E. A is relevant to B iff $P(B|E) \neq P(B|E \wedge A)$, or iff B and A are *conditionally dependent* given E. A is irrelevant to B iff A and B are *conditionally independent*. Such a notion appears to rely heavily on quantitative information of the type we have assumed is typically unavailable. However, the concept of independence can be represented qualitatively as a set of relations about independence. Pearl (1988) presents a set of five sound axioms for reasoning with conditional independence relationships (which are also conjectured to be complete) that can be viewed as a logic of independence.[18]

While independence is a vital concept, there are reasons to think irrelevance should not be identified with conditional independence. Suppose we take the following definition of relevance as determined by independence.[19]

Definition 5.27 (Gärdenfors 1978b) (a) p is *relevant* to r on evidence e iff $P(r|p \wedge e) \neq P(r|e)$.

(b) p is *irrelevant* to r on evidence e iff $P(r|p \wedge e) = P(r|e)$ or $\vdash \neg(p \wedge e)$.

Thus if p is logically impossible given e, we say p is irrelevant by fiat.[20]

As noted by Keynes (cf. Gärdenfors (1978b)), there is a sense in which aspects of relevance are missing from this definition. Certain information p may be judged to be irrelevant to r on this definition even though various components of p are separately viewed as relevant. For instance, we can have $P(r|p \wedge q \wedge e) = P(r|e)$ while also having $P(r|p \wedge e) \neq P(r|e)$ and $P(r|q \wedge e) \neq P(r|e)$. The combination of p and q negates the individual effects these pieces of evidence have on the acceptance (or degree of belief) of r.[21] However, one may still be willing to view $p \wedge q$ as relevant to r, so Keynes proposed a stronger definition of relevance.

Definition 5.28 (Gärdenfors 1978b) (a) p is *irrelevant* to r on evidence e iff there is no sentence q, that is derivable from $p \wedge e$ but not from e alone, such that $P(r|q \wedge e) \neq P(r|e)$.

(b) p is *relevant* to r on evidence e otherwise.

Given this new definition, the feature of independence allowing the combination of relevant evidence to become irrelevant is counteracted. If p is relevant to r, then so is any information

[18] The axioms are sound and (likely) complete in the sense that any set of relationships satisfying the axioms corresponds precisely to the set of independencies determined by some probability model P.

[19] The definition is taken from Gärdenfors (1978b) following Carnap (1950).

[20] The definition is formulated asymmetrically to handle the case of logically impossible information. Otherwise relevance and irrelevance are completely determined by a shift (or not) in conditional probability.

[21] For example, adapting the party example from the counterfactual literature, we can state that Pete will likely have a good time at the party (with some probability); but he will have a rotten time if either Mary or Jane come (the probability goes down in each case); but he will have a good time if they both come (with the original probability), since they usually go off together and leave him alone.

that includes p. Unfortunately, Carnap (1950) has shown Keynes's definition to be vacuous in the following sense: if proposition r is contingent on evidence e, then p is irrelevant to r on e iff $\vdash e \supset p$. This is certainly not a reasonable definition of relevance, for we must allow some sentences p, contingent on e, to be irrelevant to r (which is also contingent on e). Otherwise, there is no need to define relevance at all — everything is relevant to everything else!

Gärdenfors (1978b) shows that Carnap's trivialization result does not rely on the formulation of Definition 5.28 in terms of conditional probabilities. Rather it is the essential nature of Keynes's proposal, that information is relevant if any part of it is relevant, that is untenable. Gärdenfors proposes five basic qualitative postulates for relevance and irrelevance, and a sixth that captures the spirit of Definition 5.28. Let $p\mathcal{R}_e r$ mean p is relevant to r on evidence e and $p\mathcal{I}_e r$ mean p is irrelevant to r on e. The Gärdenfors (1978b) postulates are:

(I0) If $\vdash e \supset (p \equiv q)$ then $p\mathcal{R}_e r$ iff $q\mathcal{R}_e r$.

(I1) $p\mathcal{R}_e r$ iff not $p\mathcal{I}_e r$.

(I2) $p\mathcal{R}_e r$ iff $\neg p\mathcal{R}_e r$.

(I3) $\top\mathcal{I}_e r$.

(I4) If r is contingent on e, there is some q such that $q\mathcal{R}_e r$.[22]

(I5) If $p\mathcal{R}_e r$ and $\not\vdash \neg(p \wedge q \wedge e)$ then $(p \wedge q)\mathcal{R}_e r$.

(I0)–(I4) are considered basic postulates any notion of relevance must satisfy. So, for instance, relevance depends only on semantic content (I0) and is the complement of irrelevance (I1). (I2) says that if p is relevant to r then so is $\neg p$. This is motivated by statistical considerations, for if the conditional probability of r given p is different from $P(r)$, then so is the conditional probability given $\neg p$. We return to this point later. (I5) reflects Keynes's considerations for a stronger concept of relevance. In conjunction with the other postulates, (I5) leads to the following trivialization result.

Theorem 5.26 (Gärdenfors 1978) *For fixed evidence e, let $\mathcal{R}_e$ and $\mathcal{I}_e$ satisfy (I0) through (I5). Then every sentence contingent on e is relevant to every other sentence contingent on e.*

Thus, Keynes's definition cannot be satisfied meaningfully, assuming the other five postulates, (I0) through (I4), are accepted. For this reason, Gärdenfors puts forth and examines two weaker conditions that separately reflect part of the motivation for Definition 5.28 and do not lead to trivialization. However, another approach, which we will now investigate, is to keep (I5) and weaken the other postulates. The concept of relevance thus determined is certainly distinct from that envisioned by Gärdenfors and Keynes, but (we will argue) is no less meaningful.

We want to provide a qualitative account of relevance that we may relate to the normative conditional sentences of CO*. Just as quantitative relevance is determined by some probability model P that determines all conditional probabilities, we will define qualitative relevance in terms of CO*-models (or CO* theories) that determine the truth or falsity of conditional sentences. If sentence $A \Rightarrow B$ is true we may think of A as some background evidence from which we are willing to infer B (analogous to assigning some high conditional probability to $P(B|A)$). Without actual probabilities, we cannot determine the effect of C has on this degree of belief assigned to B (given

[22] In fact, Gärdenfors considers both this weaker postulate and the stronger $r\mathcal{R}_e r$, but requires only the weaker version for the following results.

A), but we can tell if C causes a *change in the acceptance* of B. If $A \wedge C \Rightarrow B$ is true, then C is irrelevant since B is still accepted given A, even if C is learned. Otherwise, when $\neg(A \wedge C \Rightarrow B)$ holds, C is relevant to B given A.

One consequence of these preliminary considerations is the abandonment of postulate (I2). Indeed, whenever $e \Rightarrow q$ holds both of p and $\neg p$ cannot be relevant to q.

Example 5.6 Suppose $e \Rightarrow q$ and p is relevant to q on evidence e, that is, $\neg(p \wedge e \Rightarrow q)$. In CO*, we can derive $e \Rightarrow \neg p$ (by **RM**) and hence $\neg p \wedge e \Rightarrow q$ (by **CM**). So $\neg p$ must be irrelevant to q on evidence e.

■

While (I2) seems well-motivated statistically, the shift to a qualitative notion of relevance justifies the violation of (I2). Consider the example of Gärdenfors (1978b). If I am about to cross a long wooden bridge in a heavy truck and someone informs me that an earthquake is going to occur within a minute, I certainly will consider this new information relevant. Suppose e forms my evidential knowledge, s means "I will cross safely" and q indicates an impending earthquake. Assuming I was willing to cross before learning of the earthquake, we have $P(s|e) > P(s|q \wedge e)$ and, by Definition 5.28, $q\mathcal{R}_e s$. Given an appropriate representation in CO*, we also consider q relevant to s based on the facts $e \Rightarrow s$ and $q \wedge e \Rightarrow \neg s$.

Now consider a similar situation where someone decides to tell me there will be no earthquake as I'm about to cross the bridge. I'm liable to dismiss my informant as a lunatic and discount the new information $\neg q$ as being irrelevant. On Definition 5.28, it must be the case that $\neg q \mathcal{R}_e s$, since $P(s|e) \neq P(s|\neg q \wedge e)$; that is, learning of no earthquake is relevant to my deliberations. This is a simple consequence of the fact that

$$P(s|e) = P(s|q \wedge e) \cdot P(q \wedge e) + P(s|\neg q \wedge e) \cdot P(\neg q \wedge e),$$

and $P(s|e) > P(s|q \wedge e)$. Intuitively, though, $\neg q$ is irrelevant to s because $P(s|e)$ is very close to $P(s|\neg q \wedge e)$, assuming $P(q \wedge e)$ is very slight. Since the probability of an earthquake is judged to be negligible anyway, the shift in the chance of a safe crossing on learning $\neg q$ is smaller than need practically be considered. Thus, we'd like to say $\neg q$ is irrelevant.

The problem with notions of relevance satisfying (I2), which we dub *statistical relevance*, or s-relevance, is the indistinguishability of the marginal relevance of $\neg q$ to fact s and the significant relevance of q to s. Note that it is not the magnitude of the change of degree of belief that is important, but rather the change in the degree of acceptance of a belief. Defining relevance in terms of normative conditionals, we lose the quantitative distinctions, but gain qualitatively in terms of acceptance. If $e \Rightarrow s$ and $q \wedge e \Rightarrow \neg s$, then we can infer $\neg q \wedge e \Rightarrow s$. While $\neg q$ might change the quantitative degree of belief in s given e it does not change the acceptance of s, whereas q does. Learning of an earthquake is relevant because it causes the belief that a crossing will be safe to be given up. Learning of no earthquake is irrelevant because the fact that a crossing will be safe is accepted both before and after being so informed. This sort of relevance is called *practical relevance*, or p-relevance, as it is based on coarse-grained, nonstatistical principles. We contrast p-relevance with s-relevance by noting that s-relevance satisfies (I2) at the expense of (I5) while p-relevance can be defined to satisfy (I5) but not (I2).

Definition 5.29 Let M be a CO*-model, and $p, q, r, e \in \mathbf{L}_{CPL}$. The *p-relevance relation determined by* M is defined as follows:

(a) If $\vdash \neg(p \wedge e)$ then $p\mathcal{I}_e r$.

(b) If $M \models e \Rightarrow r$ then
$p\mathcal{I}_e r$ iff there is no q such that $e \vdash p \supset q$ and $M \models \neg(q \wedge e \Rightarrow r)$.

(c) If $M \models e \Rightarrow \neg r$ then
$p\mathcal{I}_e r$ iff there is no q such that $e \vdash p \supset q$ and $M \models \neg(q \wedge e \Rightarrow \neg r)$.

(d) If $M \models \neg(e \Rightarrow r) \wedge \neg(e \Rightarrow \neg r)$ then
$p\mathcal{I}_e r$ iff there is no q such that $e \vdash p \supset q$ and $M \models q \wedge e \Rightarrow r$ or $M \models q \wedge e \Rightarrow \neg r$.

(e) $p\mathcal{R}_e r$ iff not $p\mathcal{I}_e r$.

We say p is p-irrelevant to r on evidence e, writing $p\mathcal{I}_e r$, just when no weaker sentence q changes the acceptance or nonacceptance of r (given e). So if r is accepted given e ($e \Rightarrow r$) then it is accepted after learning p (or any weaker q). If r and $\neg r$ are unaccepted (both $\neg(e \Rightarrow r)$ and $\neg(e \Rightarrow \neg r)$) then both remain unaccepted when learning p (or any weaker q).

We can show that this notion of relevance respects postulate (I5) without falling prey to trivialization.

Theorem 5.27 *Let M be a CO*-model. The p-relevance relation determined by M satisfies (I0), (I1), (I3), (I4) and (I5).*

Proposition 5.28 *P-relevance is nontrivial, in general. That is, there are CO*-models M and sentences p and r, both contingent on e and pairwise contingent, such that $p\mathcal{I}_e r$ where $\mathcal{I}_e$ is determined by M.*

Thus, it seems Keynes's intuitions that lead to Definition 5.28, although incompatible with s-relevance, can be fully realized using the weaker notion of p-relevance. The only part of s-relevance given up is Gärdenfors's postulate (I2). But we have seen how this discrepancy can be justified when one views relevance based on changing degrees of acceptance rather than changing degrees of belief. Thus, nothing intrinsic in (I5) leads to triviality.

The connective $>$, defined in Section 5.3, can now be related to a formal definition of irrelevance. In particular, the following holds.

Proposition 5.29 *Let M be a CO*-model and $\mathcal{I}_e$ and $\mathcal{R}_e$ the p-relevance relation determined by M. If $M \models e > r$ and $M \models e \Rightarrow r$ then $p\mathcal{I}_e r$ for all p such that $\not\vdash \neg(p \wedge e \wedge r)$.*

The claim that $A > B$ extends the conditional $A \Rightarrow B$ with irrelevant properties can be justified by appeal to the relation $\mathcal{I}_e$.

5.6 Miscellany

In this chapter we described several bimodal logics in which truth at inaccessible worlds is expressible. With the increased expressive power, we defined a new connective $>$ expressing *conditional knowing at most*, in some sense the dual of $\Rightarrow$, *conditional knowing at least*. Combining the two we have a conditional version of Levesque's (1990) *only knowing* connective. Theorems 5.4 and 5.7 show the completeness of our bimodal logics CO and CO*.

We formalized to some extent the problem of irrelevance and showed how two approaches, 1-entailment and rational closure, can be axiomatized in CO* using $>$. Corollaries 5.24 and 5.25 demonstrate the adequacy of our axiomatization of 1-entailment and rational closure within our

logics. We view this as analogous to the second-order circumscriptive axiomatization of predicate-minimal theories. Because our modal language allows information other than simple conditional sentences to be expressed, we expect that similar criteria can be developed for extended theories that include boolean combinations of conditionals (e.g., negated or disjoined defaults) and propositional facts, as well as nested conditionals. We have yet to investigate this idea fully, but the modal framework should allow the extension of rational closure to more expressive theories.

Finally, we discussed the concept of relevance in more detail, discussing properties of two kinds of relevance relations. While certain intuitions regarding relevance have been shown to be incompatible with statistical relevance, Theorem 5.27 and Proposition 5.28 demonstrate that practical relevance is in accord with these intuitions.

Much research remains to be carried out in relating irrelevance to CO*. The use of the connective $>$ and the additional expressive power is somewhat unsatisfying. While the closure of a theory T corresponds to 1-entailment, the formulation of $Cl(T)$ uses axioms that "mimic" the structure of the unique preferred model Z_T.[23] Ultimately, we would like to see a purely logical characterization of default inference in the spirit of Levesque (1990). There we can derive default conclusions simply by asking what beliefs are entailed by supposing an agent *only* knows a knowledge base; that is, we derive α such that $\models_{OL} \mathsf{O}(KB) \supset B\alpha$. While Levesque's characterization is semantically very clear and quite elegant, it relies on an autoepistemic interpretation of default rules. Such an interpretation has drawbacks. For instance, new default rules are not derivable in general. In this respect, the use of conditional logic to represent defaults is more desirable.[24]

We consider this approach of conditional only knowing as a step towards unifying these two views and combining the natural representation of default rules as conditionals with the clear semantics of only knowing. Again, the goal is to find a connective and a natural semantics, analogous to Levesque's O operator, such that $\mathsf{O}(KB)$ yields the proper default conclusions using conditional default representations. In this respect, the logical characterization of 1-entailment presented here is incomplete, in the sense that it still relies on certain extra-logical machinery.

While the use of the modal connective $\bar{\Box}$ has been somewhat *ad hoc* in characterizing 1-entailment, it does illustrate the power afforded by inaccessibility. In the next chapter, we will use $\bar{\Box}$ more naturally to capture only knowing a knowledge base in order to model belief revision, and thus demonstrate its wider applicability.

Rational closure and 1-entailment make apparently reasonable assumptions, but forcing worlds to be as normal as possible is a bit too heavy-handed in many circumstances. In general, more subtle preference criteria are needed for default reasoning with conditionals. Consider Example 5.5. Although we have $P \Rightarrow B$ and $B \Rightarrow W$, we cannot conclude $P \Rightarrow W$ via 1-entailment even though nothing in the rule base contradicts this. This is precisely because worlds are made as normal as possible. Since we have $P \Rightarrow \neg F$ and $B \Rightarrow F$, P is more exceptional than B, so the most normal B-worlds (satisfying, e.g., F and W) are not P-worlds. But because P-worlds are more exceptional, we can consider both $P \wedge W$ and $P \wedge \neg W$-worlds to be among the most normal P-worlds. This will not violate the constraint $B \Rightarrow W$ (satisfied due to the most normal B-worlds). Since it is consistent to allow $P \wedge \neg W$ to be as normal as $P \wedge W$, 1-entailment and rational closure insist on it.

As Pearl (1990) explains, $P \wedge \neg W$-worlds are intuitively more exceptional because they violate more rules than $P \wedge W$-worlds. In particular, such worlds cannot satisfy both $P \Rightarrow B$ and $B \Rightarrow W$.

[23] The term "hack" comes to mind!

[24] We discuss autoepistemic logic and conditional default rules further in Chapters 6 and 7, along with Levesque's work.

The *maximum entropy formalism* of Goldszmidt, Morris and Pearl (1990) addresses this problem by (roughly) ranking worlds according to the number and weight (Z-ranking) of the rules they falsify. More precisely, this system assigns the probability distribution of maximum entropy to the set of worlds subject to the constraint that the conditional probabilities represented by default rules are at least $1 - \varepsilon$ (thus we have a family of distributions parameterized by ε). This seems to capture the spirit of counting weighted rule violations (Goldszmidt, Morris and Pearl 1990).

We can explicitly capture "counting weighted rule violations" in CO* by mimicking the structure of such models, for fixed theories, just as we did for 1-entailment. However, there are two objections to such an approach: first, while this might yield useful conclusions, it will not usually correspond to the ranking of maximum entropy.[25] Second, the usefulness and naturalness of such a characterization is limited. Indeed, what new insights this lends to default reasoning is unclear.

A more important avenue for exploration is the relationship between only knowing and maximum entropy. Maximum entropy (all other things being equal) prefers distributions where the probability associated with each possible world is the same (or close to the same). Interpreting more probable worlds as more normal, this is very similar to the bias of only knowing. If one only knows a theory T, then all T-worlds are indistinguishable, or have the same rank. It would be interesting to discover the extent to which we can view only knowing as a symbolic, qualitative interpretation of maximum entropy, and just what such a formulation should look like. This work is merely a starting point for discussing such questions, and we hope logics like CO* provide an adequate framework within which to address such issues.

[25] Moisés Goldszmidt, personal communication. In fact, it is not clear which is more appropriate.

Chapter 6

Conditional Logics for Belief Revision

In the previous two chapters we developed an approach to default reasoning, the task of drawing plausible conclusions based on a static set of beliefs. We've seen that this notion of consequence is nonmonotonic in the sense that as a belief set is augmented with new information certain inferences are deemed less plausible than they once were, and certain conclusions unacceptable. Partly because default reasoning exhibits such nonmonotonicity our beliefs are in a constant state of flux, and it is necessary to adjust our viewpoint to allow the revision of beliefs. In this chapter we will present a formal model for belief revision, inspired by the AGM method of revision and based on a modally-defined conditional logic.

One point that emerges in Chapter 3, where various approaches to revision are surveyed, is that the AGM model, along with other characterizations such as those of Grove (1988), Katsuno and Mendelzon (1990; 1991), Nebel (1989) and Dalal (1988), have a somewhat extra-logical nature.[1] Such models are based on postulates that describe the properties of revision operators or structures that can be said to represent such operators. Clearly, some of these models carry intuitive appeal, but none can be said to be a *logic* of revision in the traditional sense. They do not provide a logical calculus or explicit consequence operation with which one can reason about the process of revision, the results of revising a KB, the constraints imposed by certain facts on the revision of a KB, and so on. The goal of this chapter will be to provide just such a logic for belief revision. We will develop a possible worlds semantic characterization of revision, strongly related to the representational structures of Grove (1988) and Katsuno and Mendelzon (1990), based on the logics CT4O and CO, and define a conditional connective within these modal logics that is adequate for revision (with respect to the AGM postulates and a certain generalization of them).

In Section 6.3 we will show how revision is related to subjunctive conditionals, claiming that our conditional for revision is just such a subjunctive. We will use this relationship to develop a framework for answering subjunctive queries of a knowledge base. In the next section, we will discuss a peculiarity of this relationship discovered by Gärdenfors (1986) known as the triviality result and attempt to "explain it away." Finally, in Section 6.5, we investigate the epistemic nature of revision and show that our logic for belief revision subsumes, in a certain sense, autoepistemic logic.

[1] One notable exception is the logic of Grahne (1991); however, that work describes the related but distinct *update* operator. As well, Lewis's (1973a) counterfactual logic VC can be viewed to some extent as effecting revision. See Section 6.4.

There are a number of important results in this chapter. Theorem 6.1 and Corollary 6.3 show how the concept of only knowing can be defined in the language of CT4O and CO. Theorems 6.7 and 6.8 demonstrate the equivalence of our logic for belief revision to the AGM postulates. Theorems 6.13, 6.16 and 6.19 show how various types of integrity constraints on the revision process can be enforced, and other intensional constraints, in the form of plausibility and entrenchment, are shown to be expressible by Theorems 6.23 and 6.24 and Corollaries 6.25 and 6.26. Proposition 6.29 is used to refute the classic triviality results in belief revision. Finally, Propositions 6.30 and 6.35 show how the notion of belief can be defined in CO*, while Theorem 6.38 demonstrates that CO* is a generalization of autoepistemic logic.

6.1 A Conditional for Revision

Consider the problem of belief revision in the case where some new fact A must be reconciled with a theory or belief set K such that $K \models \neg A$. To accommodate this new fact certain beliefs in K must be given up before A is accepted because inconsistency must be avoided at all costs.[2] The maxim of informational economy dictates that as "few" beliefs as possible be given up, where by "few" we mean that information loss should be kept to a minimum. As discussed in Chapter 3, there are few logical constraints on what counts as an acceptable revision, or what form minimal loss of information should take. For this reason, the AGM approach to revision allows one to consider arbitrary maximal subsets of K consistent with A in the course of these deliberations. The only requirement is that the set of all such subsets, $K \perp A$, should be ordered in a way that reflects the amount of information they contain, and the maximal elements of this ordering (the *maximal* maximal elements if you will) should be the basis for the revised belief set. This is essentially the motivation for partial meet revision.

A key observation of Grove (1988) and Katsuno and Mendelzon (1990) is that such an ordering of subsets can alternatively be viewed as an ordering on possible worlds reflecting a preference on states of affairs an agent would accept as epistemically possible if change in belief required it. We take this observation as a starting point for our Kripkean possible worlds semantics for belief revision.

6.1.1 Preorder Revision

Our semantics will be based on structures consisting of a set of possible worlds W and a binary accessibility relation R over W. Implicit in any such structure for revision will be some theory of interest K that is intended as the object of revision. We return momentarily to the problem of specifying K within the structure. The interpretation of R is as follows: wRv iff v is as *close* to theory K as w. As usual, v is *closer* than w iff wRv but not vRw. Closeness is a pragmatic measure that reflects the degree to which one would accept w as a possible state of affairs given that belief in K might have to be given up. If v is closer to K than w, loosely speaking, v is "more consistent" with our beliefs than w, and is a preferable alternative world to adopt. This view may be based on some notion of *comparative similarity*, for instance.[3]

[2] This, of course, only applies to ideally rational agents since consistency can always be maintained by such hypothetical beasts. But even allowing inconsistent beliefs to be held by less than ideal agents does not asperse this normative goal. To quote Levi (1980, pp.27–28): "To allow X to consider a contradictory corpus feasible does not imply that if he should detect inconsistency in his corpus he should rest content. When X's corpus is inconsistent, it breaks down as a standard of serious possibility. ... It is useless as a resource for inquiry and deliberation."

[3] See (Lewis 1973a; Lewis 1973b; Stalnaker 1984) for a defense of this notion.

The minimal requirements on relation R are quite straightforward. Clearly R should be reflexive, for w is surely as close as itself to any belief state. As well, R should be transitive, for if w is closer than v which is closer than u, then w ought to be closer than u. Other requirements on R are a bit harder to defend, for instance a requirement of forward- or total-connectedness, which we examine below. So we take reflexivity and transitivity to be the only properties, definable solely in terms of R, that it need satisfy.

As it stands, the modal logic S4 seems suitable for the task, as we appear to be dealing with simple preorder Kripke frames. However, there are other restrictions that must be imposed on our modal structures if they are to be considered appropriate for revision of K. For instance, we must insist that no world is closer to K than any world consistent with K. That is, any world *minimal* in R must be a K-world. This condition is hard to express in general, but when K is finitely specifiable as KB (the case in which we are most interested), this corresponds to insisting that $\Diamond\Box KB$ be true on the model M. To see this imagine that some R-minimal world w does not satisfy KB; then obviously $M \not\models_w \Diamond\Box KB$.

This sentence does not give equal status to all K-worlds, for there may be such structures in which *only* K-worlds are minimal, but not *all* K-worlds. This condition is not definable in our monomodal language, but if we consider the bimodal extension of S4 that allows for inaccessible worlds, namely CT4O, we can enforce this condition by asserting $KB \supset \overline{\Box}\neg KB$, when K is finitely representable. To see this, imagine some KB-world w is not R-minimal on M. Then there must be some minimal KB-world v such that vRw fails; but then $M \not\models_v KB \supset \overline{\Box}\neg KB$.

Making some simplifying assumptions, these conditions combine to give us

$$\overline{\Box}(KB \supset (\Box KB \wedge \overline{\Box}\neg KB)).^4 \tag{6.1}$$

We will abbreviate sentence (6.1) as $O(KB)$ and intend it to mean we "only know" KB. The reason for this nomenclature will become apparent later in the chapter.

This condition gives models a structure in which the set of K-worlds forms a mutually accessible cluster of worlds minimal in R, but in general CT4O-models need not be *cohesive*.[5] This implies that certain $\neg K$-worlds need not be related to the cluster of K-worlds at all in the ordering of closeness. Intuitively, any $\neg K$-world should be related, and should be further away than any K-world. To enfore this requirement, we insist that

$$\overline{\Box}\Diamond KB \tag{6.2}$$

be satisfied at some (or equivalently, all) worlds in the structure.

Definition 6.1 A *preorder revision model for theory* K is any structure $M = \langle W, R, \varphi\rangle$ such that R is reflexive, transitive and cohesive on W and $w \in W$ is R-minimal in M iff $M \models_w \alpha$ for all $\alpha \in K$.

Theorem 6.1 *Let K be finitely specified by KB and $M = \langle W, R, \varphi\rangle$ be a CT4O-model such that*

---

[4] We use $\overline{\Box}$ in front of this formula α so that truth of α at all worlds in structure M can be expressed as $M \models_w \alpha$ for any w. We could equivalently ask $M \models \alpha$ when α has no $\overline{\Box}$, but find the former notation more convenient.

[5] That is, it may be that w and v are not related in the transitive closure of the relation $(R \cup R^{-1})$, for some worlds w and v. We use R^{-1} to denote the inverse of R, so $wR^{-1}v$ iff vRw. If a frame is cohesive then for any pair of worlds w and v, v can be reached from w using some number of forward or backward steps along R, as if the directed graph corresponding to R were undirected.

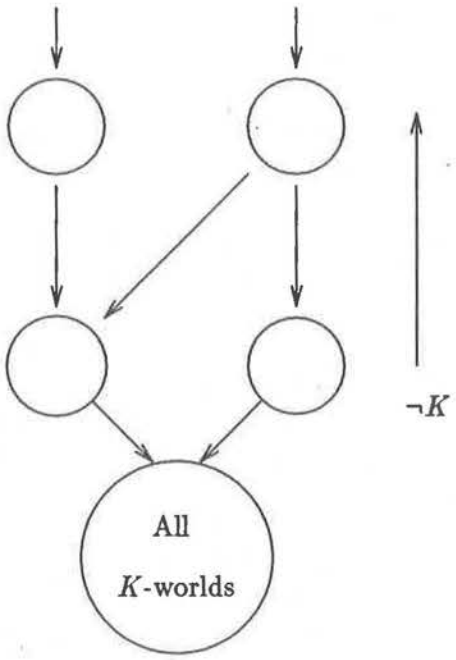

Figure 6.1: A preorder revision model for K.

for some $v, w \in W$, $M \models_w O(KB)$ *and* $M \models_v \Box\Diamond KB$. *Then* M *is a preorder revision model for* K.

So the class of CT4O-models is appropriate for what we term *preorder revision*, revision satisfying the minimal requirements on the closeness relation (see Figure 6.1). The question remains: how do we express the revision of K by A within the modal logic? The intuition is that when revising by A we should consider the worlds closest to K at which A is satisfied to represent the revised state of affairs. It is precisely these worlds that represent the minimal change or loss of information in our belief set. One problem with this characterization of the revised belief set is the assumption that such closest or minimal A-worlds exist.[6] Nothing about CT4O-models presupposes such a constraint, or prevents an infinite "descending chain" of closer and closer A-worlds. Fortunately, we can define an approach to revision without this constraint.

By adopting a different perspective on belief revision in which conditionals are key, we can ignore the Limit Assumption. Often when revising a belief set, we are not interested so much in characterizing the entire new belief state as in certain consequences of the revised theory. In other words, we are interested in facts of the sort "If I revised my beliefs to include A then I would believe B," meaning $B \in K^*_A$. Indeed, we typically cannot (or do not want to) specify the entire belief set that results when revising since it will often be too large (even for a finite KB) or contain many facts that aren't of interest. We'd rather specify certain constraints on revision in this conditional form.

We will use a conditional connective $\xrightarrow{\text{KB}}$ to represent these statements and read $A \xrightarrow{\text{KB}} B$ as "B is a consequence of revising (an implicit theory K or KB) by A." In the case where some

[6] This is known as the *Limit Assumption* and will be discussed in Section 6.2.

set of minimal A-worlds exists, clearly, such an assertion is true iff B is true at every such world. However, if such a set does not exist, for instance, if every A-world has some A-world closer than itself, we can still tell if $A \overset{\text{KB}}{\longrightarrow} B$ is true. Suppose for any A-world there is some closer A-world w (which cannot be minimal in this case) such that B holds at w, and for all worlds closer than w, B holds whenever A does. Then $A \overset{\text{KB}}{\longrightarrow} B$ should be true, for even though there are no minimal A-worlds, in some hypothetical limit B would be true. Another way of phrasing this is to say that for every state of affairs for which *some* closer A-world exists, there exists some closer $A \wedge B$-world such that $A \supset B$ holds at all worlds closer still. Yet another wording: for every $A \wedge \neg B$-world, there exists a closer $A \wedge B$-world w such that no $A \wedge \neg B$ is as close to K as w.

The goal is to express the truth conditions for such a connective within our bimodal language, that is, the requirement that B be true at all "minimal" (at least, at some hypothetical limit) A-worlds. Roughly, our considerations lead us to postulate that for each state of affairs there should be a closer state where A holds and $\Box(A \supset B)$ holds; that is $A \supset B$ holds at all even closer worlds. We express this as

$$\overset{\leftrightarrows}{\Box}\Diamond(A \wedge \Box(A \supset B)).$$

But this isn't quite right, for some worlds may not have any A-worlds that are closer, though the conditional should still hold. For instance, if $K \models \neg A$ then $\neg\Diamond A$ will hold at each K-world. We can ignore any world where $\Box\neg A$ holds as having no influence on what should count as a "minimal" A-world. Thus we are lead to the following definition.

Definition 6.2 The *revision conditional* $A \overset{\text{KB}}{\longrightarrow} B$ is defined in $\mathbf{L}_B$ as

$$A \overset{\text{KB}}{\longrightarrow} B \equiv_{\text{df}} \overset{\leftrightarrows}{\Box}(\Box\neg A \vee \Diamond(A \wedge \Box(A \supset B))).$$

In Figure 6.2 a model is shown verifying $A \overset{\text{KB}}{\longrightarrow} B$. Interestingly, the definition of $A \overset{\text{KB}}{\longrightarrow} B$ is precisely that given for $A \Rightarrow B$ in Chapter 4. The consequences of this equivalence are quite important for the relationship between default reasoning and belief revision. The bulk of Chapter 7 will examine these consequences and relate the two connectives. It is important to note that $\overset{\text{KB}}{\longrightarrow}$ does not describe a family of related connectives indexed by KB. It is a conditional connective in the usual sense. "KB" is used to emphasize the fact that $\overset{\text{KB}}{\longrightarrow}$ is typically used for the revision of some intended knowledge base. The connective is perfectly well-defined and meaningful when KB is left unspecified, as some derived theorems below indicate.

As expressed in the AGM revision postulates, in particular (R5), usually we want to allow revision by any satisfiable sentence to result in a consistent belief set. In other words, we rule out no logically possible worlds in the course of deliberations. This intuition is captured by insisting that preorder revision models have among their set of worlds all propositional valuations. In other words, they should be CT4O*-models. We refer to this class of preorder revision models as full.

Definition 6.3 A *full preorder revision model* for K is any preorder revision model M for K such that M is a CT4O*-model.

Of course, the analogous result to Theorem 6.1 holds for CT4O*-models that satisfy the sentences given above. Before examining consequences of these definitions, we turn our attention to a specialization of them.

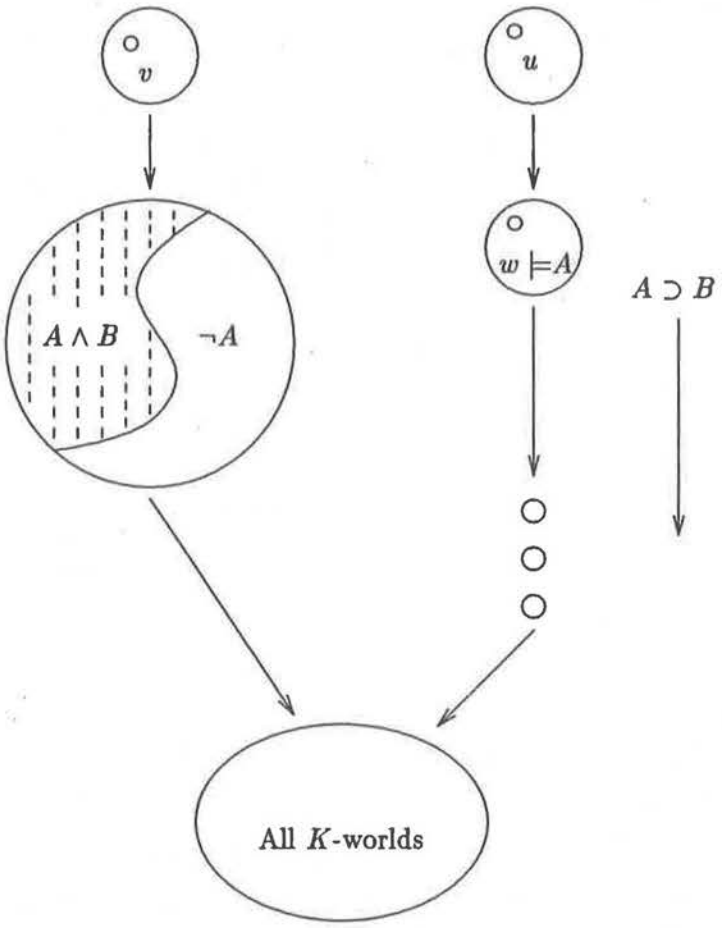

Figure 6.2: A model where revising K by A results in belief B. At each world where A can be "seen" $A \wedge \Box(A \supset B)$ can also be seen. For v there is a set of minimal (closest) A-worlds verifying B (the shaded area). For u there is no such set of minimal A-worlds; but there is a point w at which A and B hold and at which $A \supset B$ holds at all lower points.

6.1.2 Total Order Revision

We noted in Chapter 3 that Grove's (1988) system of spheres model and the model of Katsuno and Mendelzon (1990), that correspond to the class of AGM revision functions, can be viewed as placing a total preorder on worlds instead of merely a preorder. This is in agreement with the intuition that any two worlds should be comparable according to the closeness or similarity measure. If we consider two states of affairs to be such that neither is closer to K than the other, then we insist that both be judged *equally* close or similar. It should be evident that imposing this additional constraint on R yields the logic CO, as shown by the following theorem.

Definition 6.4 A *total order revision model for theory* K is any structure $M = \langle W, R, \varphi \rangle$ such that R is reflexive, transitive and totally connected on W and
$v \in \{w : M \models_w \alpha \text{ for all } \alpha \in K\}$ iff v is R-minimal in M.

Theorem 6.2 *Let K be finitely specified by KB and $M = \langle W, R, \varphi \rangle$ be a CO-model such that for some $v, w \in W$, $M \models_w O(KB)$ and $M \models_v \overset{\exists}{\Box}\Diamond KB$. Then M is a total order revision model for K.*

Of course, the constraint of cohesiveness specified by sentence (6.2) is redundant in the case of CO-models, which are totally connected.

Corollary 6.3 *Let $M = \langle W, R, \varphi \rangle$ be a CO-model such that $M \models_w O(KB)$ for some $w \in W$. Then M is a total order revision model for K.*

We will refer to such models more simply as *revision models* for K, or K-revision models. In case $K = Cn(KB)$, a revision model for KB denotes a revision model for K.

We have motivated the definition of $A \overset{\text{KB}}{\longrightarrow} B$ in terms of CT4O-models and certainly such a definition is applicable in the case of total order revision, it being a special case of preorder revision. However, we can provide an alternative definition based on clearer intuitions that take advantage of the additional structure. Recall that CO-models are total preorders, and consist of a totally ordered set of clusters of possible worlds. Hence a revision model has the cluster $\|K\|$ as a minimum.

In the case where A is inconsistent with K, clearly $A \overset{\text{KB}}{\longrightarrow} B$ is true exactly when the formula

$$\overset{\exists}{\Diamond}(A \wedge \Box(A \supset B))$$

is true at any $w \in \|K\|$. This formula says there is some world satisfying A such that $A \supset B$ is true at all closer worlds. Since the ordering on W is total, this ensures any "minimal" A-world satisfies B. However, if there is an A-world in $\|K\|$, this is unsatisfactory, as the operator $\overset{\exists}{\Diamond}$ refers only to possibility at inaccessible worlds (hence, not to any K-worlds). In such a circumstance $A \overset{\text{KB}}{\longrightarrow} B$ is true when

$$\Diamond(A \wedge \Box(A \supset B))$$

holds at any $w \in \|K\|$. This is so because $\Box(A \supset B)$ means $K \models A \supset B$ so that the closest A-worlds, those in $\|K\|$ satisfy B. Disjoining the two,[7] we arrive at

$$\overset{\exists}{\Diamond}(A \wedge \Box(A \supset B)).$$

We can drop the restriction that $A \overset{\text{KB}}{\longrightarrow} B$ be evaluated at some $w \in \|K\|$, since if it holds at some $w \in W$, it holds at all worlds in W. This sentence does not account for impossible antecedents, in

[7] This disjunction is valid because if the first condition holds and some $K \wedge A$-world exists, the second will be true also.

which case we should expect the conditional to be vacuously true; that is, revising by impossible A entails every sentence B. Finally, the definition is given as

$$A \xrightarrow{\text{KB}} B \equiv_{\text{df}} \overset{\text{u}}{\Box}\neg A \vee \overset{\text{u}}{\Diamond}(A \wedge \Box(A \supset B)). \tag{6.3}$$

If this motivation is accurate the definition of $\xrightarrow{\text{KB}}$ given in Definition 6.2 should be equivalent to this sentence in CO. This is indeed the case.

Proposition 6.4 $M \models_{CO} \overset{\text{u}}{\Box}\neg A \vee \overset{\text{u}}{\Diamond}(A \wedge \Box(A \supset B))$ *iff* $M \models_{CO} \overset{\text{u}}{\Box}(\Box\neg A \vee \Diamond(A \wedge \Box(A \supset B)))$.

When dealing with total order revision, we will use the simpler definition of $\xrightarrow{\text{KB}}$ afforded by sentence (6.3).[8] Once again, we usually want to consider only those models by which every consistent sentence is capable of being revised consistently.

Definition 6.5 A *full total order revision model* (or full revision model) for K is any total order revision model M for K such that M is a CO*-model.

6.1.3 Characterization Results

The connective $\xrightarrow{\text{KB}}$ does not characterize belief revision in the manner of a revision function $*$. Rather than asserting that a belief set K_A^* results when we revise by A, $A \xrightarrow{\text{KB}} B$ makes a weaker assertion, that B will be believed when we revise by A, or (abusing notation) $B \in K_A^*$. This view of revision could be adopted within the AGM framework; for instance, one could define a conditional connective $>$ such that $A > B$ is true iff $B \in K_A^*$ (see Section 6.3). In this sense, our conditional approach is no more general than the AGM approach. In fact, one might claim that using a revision function $*$ strictly subsumes the use of $\xrightarrow{\text{KB}}$. However, this is not the case; we can define a revision operator $*$ in terms of $\xrightarrow{\text{KB}}$ that does map belief sets into revised belief sets. This is fortunate as it allows us to compare the conditional model for revision to the AGM postulates.

Definition 6.6 Let M be a preorder revision model for K. The *revision function determined by* M is denoted $*^{\text{M}}$ and is defined for each $A \in \text{L}_{CPL}$ by

$$K_A^{*^{\text{M}}} = \{B \in \text{L}_{CPL} : M \models A \xrightarrow{\text{KB}} B\}.$$

Some derived theorems and rules of inference for the CT4O will shed some light on the nature of the connective $\xrightarrow{\text{KB}}$ and on the induced revision functions $*^{\text{M}}$.[9]

Proposition 6.5 *The following are derived theorems and inference rules in* CT4O *(assuming as a premise $O(KB)$ wherever KB is mentioned).*

RCM *From* $B \supset C$ *infer* $(A \xrightarrow{\text{KB}} B) \supset (A \xrightarrow{\text{KB}} C)$

And $(A \xrightarrow{\text{KB}} B) \wedge (A \xrightarrow{\text{KB}} C) \supset (A \xrightarrow{\text{KB}} B \wedge C)$

[8] Of course, given the equivalence of the definitions of $\xrightarrow{\text{KB}}$ and $\Rightarrow$ we can use this simpler definition for $\Rightarrow$ in CO. See Chapter 4.

[9] More theorems of the logics and other properties will be examined in the next section.

ID $A \xrightarrow{\text{KB}} A$

KC $(A \xrightarrow{\text{KB}} B) \supset (KB \wedge A \supset B)$

CK $\overset{\leftrightarrow}{\Diamond}(KB \wedge A) \supset (\overset{\leftrightarrow}{\Box}(KB \wedge A \supset B) \supset (A \xrightarrow{\text{KB}} B))$

Cons $\overset{\leftrightarrow}{\Box}\neg A \equiv (A \xrightarrow{\text{KB}} \bot)$

LLE *From* $A \equiv B$ *infer* $(A \xrightarrow{\text{KB}} C) \supset (B \xrightarrow{\text{KB}} C)$

KCI $(A \wedge B \xrightarrow{\text{KB}} C) \supset (A \xrightarrow{\text{KB}} (B \supset C))$

These theorems ensure that $*^{\text{M}}$ behaves in a reasonable manner. In fact, these correspond precisely to the first seven AGM postulates. **RCM** ensures that if revising by A entails belief in B and B implies C then C will also be believed. Together with **And** this ensures $K_A^{*^{\text{M}}}$ is a deductively closed set, that is, a belief set (postulate (R1)). **ID** asserts that A will be believed when K is revised by A, so $A \in K_A^{*^{\text{M}}}$ (R2). **KC** says $B \in K_A^{*^{\text{M}}}$ only of $KB \wedge A$ implies A, where $K = Cn(KB)$. In other words, $B \in K_A^{*^{\text{M}}}$ only if $B \in K_A^{+}$ (R3). **CK** asserts that if $KB \wedge A$ is possible then B is a consequence of $KB \wedge A$ only if it is believed when KB is revised by A. If we assume our revision models are full so that all logically possible worlds are represented, then **CK** means $B \in K_A^{+}$ iff $B \in K_A^{*^{\text{M}}}$ whenever A is consistent with K (R4). We return to the case of arbitrary revision models below. **Cons** ensures that $K_A^{*^{\text{M}}}$ is the inconsistent belief set iff A is not possible. Once again, in the case of full revision models this is postulate (R5). **LLE** asserts that revision depends only on the semantic content of the new information, not its syntactic structure. Hence $K_A^{*^{\text{M}}} = K_B^{*^{\text{M}}}$ if A and B are semantically equivalent (R6). **KCI** ensures that if $C \in K_{A \wedge B}^{*^{\text{M}}}$ then $B \supset C \in K_A^{*^{\text{M}}}$, hence that $C \in (K_A^{*^{\text{M}}})_B^{+}$, ensuring the satisfaction of (R7).

A further theorem of CO is the following.

Proposition 6.6 CKI *is derivable in CO.*

CKI $\neg(A \xrightarrow{\text{KB}} \neg B) \supset ((A \xrightarrow{\text{KB}} (B \supset C)) \supset (A \wedge B \xrightarrow{\text{KB}} C))$

CKI ensures that whenever B is consistent with $K_A^{*^{\text{M}}}$ then $C \in (K_A^{*^{\text{M}}})_B^{+}$ if $C \in K_{A \wedge B}^{*^{\text{M}}}$. This is just postulate (R8).

If we assume that we are dealing solely with full revision models or CO*-models, clearly the revision functions $*^{\text{M}}$ determined by revision models satisfy the AGM postulates.

Theorem 6.7 *Let M be a full revision model and $*^{\text{M}}$ the revision function determined by M. Then $*^{\text{M}}$ satisfies postulates (R1) through (R8).*

In the case of arbitrary CO-models, where some logically possible worlds may be excluded from consideration, postulates may be violated.

Example 6.1 Let M be a (nonfull, preorder or total order) revision model for K such that $M \models \neg\overset{\leftrightarrow}{\Diamond}A$ for some satisfiable proposition A. Then $K_A^{*^{\text{M}}} = Cn(\bot)$, since (by definition of $\xrightarrow{\text{KB}}$) $M \models A \xrightarrow{\text{KB}} B$ for all B. This contradicts (R5).

■

CK and **Cons** correspond only weakly to (R4) and (R5) unless we assume that we are using the stronger logic CT4O*. **CK** states that if $KB \wedge A$ is possible in the structure under consideration, which we take to be an agent's set of logical possibilities, then KB revised by A is to be the same as merely conjoining A to KB. Similarly, **Cons** asserts that revising by A results in an inconsistent belief set only when A is not possible relative to the structure of interest.

Let A be *relatively consistent* with respect to a particular model (or agent who "accepts" this model) if $\overset{\leftrightarrow}{\Diamond}A$ is true in that model (the agent considers A logically possible[10]). We can understand A being relatively consistent to mean that an agent would be willing to accept A if circumstances or new information warranted. If we replace the notion of logical consistency in (R4) and (R5) with relative consistency then **CK** and **Cons** correspond to these weakened postulates, namely

(WR4) If A is relatively consistent with K then $K_A^+ \subseteq K_A^*$.

(WR5) $K_A^* = Cn(\bot)$ iff A is relatively consistent.

Of course, the conception of expansion used in the remaining postulates and any implicit use of consistency must be adjusted to this idea of relative consistency. Consider (R3) for instance: if A is relatively inconsistent, but logically satisfiable, then certainly $K_A^{*^M} \not\subseteq K_A^+$, since K_A^+ will be logically consistent whereas $K_A^{*^M}$ will not. To get this correspondence with (R3), (R7) and (R8), we say that *expansion relative to* M is given by

$$K_A^+ = \{B : M \models_w B \text{ whenever } M \models_w K \text{ and } M \models_w A\}.$$

In the case of logic CT4O* or CO* the notions of logical consistency and relative consistency coincide, since $\overset{\leftrightarrow}{\Diamond}A$ is derivable for any propositionally satisfiable A. In this case, **CK** and **Cons** are equivalent to (R4) and (R5). If we adjust the AGM postulates to reflect consistency relative to a model, then the theorem above will hold for any K-revision (CO) model, not just full revision (CO*) models.

In general, only the full versions of (preorder or total order) revision models will be of interest. Therefore when we refer to revision models of either type we will intend full revision models unless otherwise stated. Most results for full revision models will hold for their more general counterparts if any (implicit or explicit) reference to logical consistency is replaced by relative consistency.

Theorem 6.7 shows revision models to satisfy the eight AGM postulates and to form a subclass of the space of AGM revision operators. If this conditional approach to revision is to be completely general we must show that any AGM revision operator has a corresponding revision model M for each theory K in the domain of $*$. This can be shown by constructing, for any AGM revision function $*$, a revision model such that $*^M$ is identical to $*$ (for the theory K implicit in M). Roughly, we proceed by considering all possible worlds over the propositional language and ordering them as follows: let $v \leq w$ (i.e. wRv — recall that worlds "see" other worlds lower in the order of closeness) iff there is some A such that $w \models A$ and $v \in \|K_A^*\|$. The idea is that if v satisfies theory K_A^* then it should be at least as close to K as any other A-world. This leads us to one of the main results of this thesis, that demonstrates, along with Theorem 6.7, that CO* is an adequate logic for reasoning about belief revision.

Theorem 6.8 *Let $*$ be a revision function satisfying postulates (R1) through (R8). Then for any theory K there exists a full revision model M such that $K_A^* = K_A^{*^M}$ for all A.*

[10]Epistemically possible might be a better term here, but we have reserved that to refer to the facts consistent with an agent's beliefs, rather than those facts an agent might be willing to accept if new information forces such revision. Perhaps physical possibility would capture our intent in certain circumstances.

In other words, the class of revision functions determined by revision models is exactly the class of AGM revision functions. Hence, the modal logic CO* is an appropriate calculus for reasoning about belief revision. In fact, CO*, together with the defined connective $\xrightarrow{\text{KB}}$, appears to be the first sound and complete logical axiomatization of AGM revision in the traditional sense. No other system for AGM revision is specified solely in terms of "standard" logical consequence over a fixed logical language. Furthermore, the weaker logic CO allows a generalization of AGM revision in which certain logical possibilities can be excluded from such deliberations. Perhaps an agent considers these possibilities too remote, or there might exist certain integrity constraints on a database that prohibit such possibilities.[11]

One aspect of this theorem that is perhaps not obvious is the manner in which the inconsistent belief set $Cn(\bot)$ is dealt. We have imposed no special conditions for revising $Cn(\bot)$. Of course, the only set of worlds satisfying each sentence in this set is the empty set, thus representing $Cn(\bot)$ as the minimal cluster in some K-belief model is impossible. Indeed, the constraint that we only know $\bot$ is vacuous, since $O(\bot)$ is a tautology in CO.[12] While we are not primarily concerned with this belief set, the AGM postulates require that a revision function is applicable to *any* belief set, and that a consistent set should result when we revise a belief set by a consistent sentence. So while there is no K-revision model corresponding to the inconsistent belief set, there must still be a model corresponding to its revision.

It is not hard to see that we can still represent revision of $Cn(\bot)$ by imagining some K-revision model as having an empty cluster of worlds lying "below" all other worlds, that is, having an empty minimal cluster. When $Cn(\bot)$ is revised by A, we can quite easily determine the minimal A-worlds in the model; but since there are no worlds in the (imagined) "minimal" cluster, this forces us to accept (as representative of the revised theory) worlds that are not part of the "current" belief set. In other words, there can be no consistent revision of the type described by (R4). This is the key difference when revising the inconsistent belief set. However, this does not imply that no K-revision model corresponds to a revision function with respect to the belief set $Cn(\bot)$. Rather, there is not (necessarily) a model in which the set of worlds satisfying $Cn(\bot)$ is minimal in the model. For a particular AGM revision function $*$, we may take as the minimal cluster of an appropriate K-revision model the set of worlds satisfying the belief set

$$K = \{A : A \in (Cn(\bot))^*_\top\}$$

Thus, the minimal set of worlds is made up of those that are deemed possible by merely "returning" $Cn(\bot)$ to consistency; and a K-revision model for this theory K can also be construed as a revision model for the inconsistent belief set.

We now return our attention to the class of preorder revision models. It is easy to see that any such model determines a revision operator that satisfies postulates (R1)–(R7). However, postulates for preorder revision have not been proposed by AGM, so we cannot compare such revision models to an existing general standard. However, Katsuno and Mendelzon (1990) have proposed postulates for preorder revision that they show to be adequate for propositional languages with finitely many atomic variables in which belief sets are modeled as (finite) propositional sentences. These postulates cannot be directly applied to our situation for they do not deal with deductively closed (infinite) theories. We do not pursue the properties of preorder revision here, though this is certainly an interesting avenue for further investigation. In particular, it should prove instructive

[11] We return to this point in Section 6.3.

[12] This is discussed in more depth in Section 6.5 in the context of autoepistemic logic, where such belief sets are important.

to construct a set of postulates, similar to (and including most of) the eight AGM postulates. Certainly, (R8) will not be satisfied, but a weaker version will likely be required to prove the appropriate representation theorems. Such a representation would allow further comparison of this preorder revision model with that Katsuno and Mendelzon, though on the surface they appear to reflect the same intuitions.

6.2 Properties of the Logics

6.2.1 Some Derived Theorems and Examples

In general, reasoning about the process of revising an objective KB with the modal logics CT4O* and CO* requires one to assert as a background theory the sentence $O(KB)$, together with the sentence $\Box\Diamond KB$ in the case of CT4O*. This background theory enforces the restriction that all states of affairs consistent with our objective knowledge are minimal in our ordering of closeness, that we prefer to "hang on to these beliefs" if possible. This background induces the logical constraints required of belief revision but specifies little else. Consider the following example.

Example 6.2 Consider a background theory $O(KB)$ where

$$KB = \{\texttt{bird}, \texttt{fly}, \texttt{bird} \supset \texttt{fly}, \texttt{penguin} \supset \neg\texttt{fly}\}.$$

If we want to ask about revision of KB by green, we can show

$$O(KB) \vdash_{CO*} \texttt{green} \xrightarrow{\texttt{KB}} \alpha \quad \text{iff} \quad KB \cup \{\texttt{green}\} \vdash \alpha$$

for any proposition α, since green is consistent with KB. However, if we ask

$$O(KB) \vdash_{CO*} \texttt{penguin} \xrightarrow{\texttt{KB}} \alpha$$

very little is derivable as penguin is inconsistent with KB. In general, this will only hold for those α that are logical consequences of penguin.

■

This example illustrates the extreme generality of the AGM postulates and our approach to revision. From the context of KB (in particular, the suggestively labeled atomic propositions), it seems we ought to be able to derive more than penguin when we revise by it, but this is a matter of pragmatics. It should be evident that, say, $\neg$fly should not be derivable without further constraints on the revision process. For example, penguin $\supset \neg$fly might only be in KB because it is vacuously true and has no information content. Therefore revising by penguin should result in keeping all beliefs but this one. It could also be that bird or bird $\supset$ fly should be given up.

These contextual difficulties are resolved by introducing as premises conditional sentences constraining the revision process. In fact, most of our information might be of this form. Asserting penguin $\xrightarrow{\texttt{KB}} \neg$fly ensures that penguin $\supset \neg$fly is not given up.

Example 6.3 Let KB be as before and let S be the set of additional premises containing

$$\texttt{bird} \xrightarrow{\texttt{KB}} \texttt{fly}, \quad \texttt{penguin} \xrightarrow{\texttt{KB}} \neg\texttt{fly}, \quad \texttt{penguin} \xrightarrow{\texttt{KB}} \texttt{bird}.$$

From $S \cup \{O(KB)\}$ we can derive in CO*

$$\texttt{penguin} \xrightarrow{\texttt{KB}} \neg\texttt{fly}, \quad \texttt{penguin} \wedge \texttt{bird} \xrightarrow{\texttt{KB}} \neg\texttt{fly}.$$

■

Indeed, the content of KB only constrains revision when the new information is consistent with KB. As these examples illustrate, the AGM postulates say very little about what information should be given up in the case of revising a knowledge base by a fact inconsistent with it. Only by asserting as premises constraints on the revision process can we determine some consequences of such a revision. In the next section we will present a methodology for reasoning about belief revision using both the content of the knowledge base and premises expressing constraints on revision. These examples each fit within this framework.

Some further theorems derivable in CO are given below. The first four are also valid in CT4O.

Proposition 6.9 *The following theorems are derivable in CO.*

And $(A \overset{\text{KB}}{\longrightarrow} B) \wedge (A \overset{\text{KB}}{\longrightarrow} C) \supset (A \overset{\text{KB}}{\longrightarrow} B \wedge C)$

Or $(A \overset{\text{KB}}{\longrightarrow} C) \wedge (B \overset{\text{KB}}{\longrightarrow} C) \supset (A \vee B \overset{\text{KB}}{\longrightarrow} C)$

RT $(A \overset{\text{KB}}{\longrightarrow} B) \supset ((A \wedge B \overset{\text{KB}}{\longrightarrow} C) \supset (A \overset{\text{KB}}{\longrightarrow} C))$

CM $(A \overset{\text{KB}}{\longrightarrow} B) \wedge (A \overset{\text{KB}}{\longrightarrow} C) \supset (A \wedge B \overset{\text{KB}}{\longrightarrow} C)$

RM $(A \overset{\text{KB}}{\longrightarrow} C) \wedge \neg(A \wedge B \overset{\text{KB}}{\longrightarrow} C) \supset A \overset{\text{KB}}{\longrightarrow} \neg B$

CV $\neg(A \overset{\text{KB}}{\longrightarrow} B) \supset ((A \overset{\text{KB}}{\longrightarrow} C) \supset (A \wedge \neg B \overset{\text{KB}}{\longrightarrow} C))$

As well, revision by A need not result in the belief of B or $\neg B$, hence $\overset{\text{KB}}{\longrightarrow}$ fails to satisfy CEM, the *Conditional Law of Excluded Middle* (Stalnaker 1968); for in general the set

$$\{\neg(A \overset{\text{KB}}{\longrightarrow} B), \neg(A \overset{\text{KB}}{\longrightarrow} \neg B)\}$$

is consistent. However, if revision by A results in belief B, it cannot also result in $\neg B$ (for consistent A), since

$$\overset{\leftrightarrow}{\Diamond} A \supset ((A \overset{\text{KB}}{\longrightarrow} B) \supset \neg(A \overset{\text{KB}}{\longrightarrow} \neg B))$$

is valid.

Example 6.4 Let S be the set of premises

$$\{\texttt{bird} \overset{\text{KB}}{\longrightarrow} \texttt{fly},\ \texttt{penguin} \overset{\text{KB}}{\longrightarrow} \texttt{bird},\ \texttt{emu} \overset{\text{KB}}{\longrightarrow} \texttt{bird},$$

$$\texttt{penguin} \overset{\text{KB}}{\longrightarrow} \neg\texttt{fly},\ \texttt{emu} \overset{\text{KB}}{\longrightarrow} \neg\texttt{fly}\}.$$

Revising some KB by emu or by penguin results in the belief $\texttt{bird} \wedge \neg\texttt{fly}$ because by **And** (for example)

$$S \vdash_{CO*} \texttt{emu} \overset{\text{KB}}{\longrightarrow} \texttt{bird} \wedge \neg\texttt{fly}.$$

Furthermore, if we only want to revise by the disjunction $\texttt{penguin} \vee \texttt{emu}$ this still holds as by **Or**

$$S \vdash_{CO*} \texttt{penguin} \vee \texttt{emu} \overset{\text{KB}}{\longrightarrow} \texttt{bird} \wedge \neg\texttt{fly}.$$

By **RM** one can also conclude

$$\texttt{bird} \overset{\text{KB}}{\longrightarrow} \neg\texttt{penguin} \quad \text{and} \quad \texttt{bird} \overset{\text{KB}}{\longrightarrow} \neg\texttt{emu}.$$

■

Certain facts in the objective KB can also influence the conditionals that hold. For instance, if in this example $\texttt{bird} \in KB$ then for $\{O(KB)\} \cup S$ to be consistent, it must be the case that $\neg\texttt{emu}$, $\neg\texttt{penguin}$ and $\texttt{fly}$ are in (or are entailed by) KB. Revising by $\texttt{bird}$ should leave KB intact, and by the premises S, should entail these facts; hence, they must be contained in KB.

Proposition 6.10 *Let M be a revision model for KB. If $M \models A \overset{\text{KB}}{\longrightarrow} B$ then $KB \models A \supset B$.*

Example 6.5 The "unemployed student" example reveals the default quality of revision. Let S be the set of premises

$$\{\texttt{student} \overset{\text{KB}}{\longrightarrow} \texttt{adult}\ , \texttt{student} \overset{\text{KB}}{\longrightarrow} \neg\texttt{employed}\}.$$

By **CM** we can derive

$$\texttt{student} \wedge \texttt{adult} \overset{\text{KB}}{\longrightarrow} \neg\texttt{employed}.$$

If $\texttt{adult} \overset{\text{KB}}{\longrightarrow} \texttt{employed}$ is added to S then we can also infer

$$\neg(\texttt{student} \wedge \texttt{adult} \overset{\text{KB}}{\longrightarrow} \texttt{employed})$$

and by **RM**

$$\texttt{adult} \overset{\text{KB}}{\longrightarrow} \neg\texttt{student}.$$

So if we learn someone is an adult we should believe she is not a student.

■

Example 6.6 In general, since CEM is not valid, we can have as consistent premises in S both

$$\neg(\texttt{holding(Cup)} \overset{\text{KB}}{\longrightarrow} \texttt{holdingright(Cup)}) \quad \text{and}$$

$$\neg(\texttt{holding(Cup)} \overset{\text{KB}}{\longrightarrow} \texttt{holdingleft(Cup)}).$$

However, considerations of tractability might require a default assumption, say,

$$\texttt{holding(Cup)} \overset{\text{KB}}{\longrightarrow} \texttt{holdingright(Cup)},$$

that the robot typically grabs things with its right hand (see Chapters 2 and 4). In this case, if we have a constraint[13]

$$\texttt{holdingleft(Cup)} \equiv \neg\texttt{holdingright(Cup)}$$

then we can derive

$$\texttt{holding(Cup)} \overset{\text{KB}}{\longrightarrow} \neg\texttt{holdingleft(Cup)}.$$

■

Example 6.5 demonstrates that our approach to revision has a definite default character. Revising by $\texttt{adult}$ alone would cause a belief in $\texttt{employed}$ while revising by this together with $\texttt{student}$

[13]This constraint is not useful if expressed in KB. It could be expressed less naturally in S; however, in the next section we discuss a method for enforcing such intensional constraints.

results in believing **unemployed**. Nevertheless, Example 6.4 and Proposition 6.10 illustrate an important difference between default reasoning and deliberations of revision, that is a commitment to (a form of) modus ponens. If $\texttt{bird} \overset{\text{KB}}{\longrightarrow} \texttt{fly}$ is a true description of revision and **bird** is among our beliefs about the world in KB, then it must be that **fly** is in (or entailed by) KB. Otherwise, $\texttt{bird} \overset{\text{KB}}{\longrightarrow} \texttt{fly}$ could not be true, for revising by **bird** must result in the original KB. The relationship between belief revision and default reasoning will be examined in greater detail in Chapter 7.

6.2.2 The Limit Assumption and Intensional Constraints

The results of the previous section would seem to suggest that there are no important differences in the type of reasoning sanctioned by the AGM postulates and our conditional (or modal) approach to belief revision. While the logic CO* captures AGM revision functions via logical axioms, all (CO*) revision functions satisfy the AGM postulates. However, we will examine Grove's (1988) representation theorem for AGM functions in order to highlight some important advantages of our logical characterization of revision.

In Chapter 3 we briefly discussed Grove's system of spheres model of revision, for which we now provide a formal definition.

Definition 6.7 (Grove 1988) Let W be the set of valuations (or possible worlds)[14] for some fixed language and K some theory. A *system of spheres centered on* K is a collection S of subsets of W such that

1. S is totally ordered under set inclusion, $\subseteq$.
2. $\{w : w \models K\}$ is the $\subseteq$-minimum of S.
3. $W \in S$ (and hence is the maximum of S).
4. For any sentence A, there is a smallest $s \in S$ intersecting $\|A\|$ (the set of worlds satisfying A). This smallest s is denoted $s(A)$.

Thus S consists of a sequence $S_0 \subseteq S_1 \subseteq \cdots \subseteq W$ of subsets of W with $S_0 = \|K\|$ being its minimum. The intuition is that smaller members of S (spheres) contain worlds "closer" to the actual state of affairs K. The sentences resulting from revising K by A are just those true at all worlds in $s(A) \cap \|A\|$. That is, we define K^*_A as

$$K^*_A = \{\alpha : s(A) \cap \|A\| \models \alpha\}.$$

Grove shows that this model of revision exactly characterizes AGM revision function.

Theorem 6.11 (Grove 1988) *A revision operator induced by a system of spheres (as described above) satisfies postulates (R1) through (R8).*

Theorem 6.12 (Grove 1988) *If a revision operator $*$ satisfies postulates (R1) through (R8) then for any theory K there exists a system of spheres centered on K that induces $*$.*

[14] Grove actually uses maximal consistent sets of the underlying language. This distinction is unimportant for our purposes.

We observe that a system of spheres can be viewed as imposing a total preorder on the set of worlds W as follows: let $w \leq v$ iff every sphere that contains v also contains w. In other words, the smallest sphere containing w is enclosed in the smallest sphere containing v. This observation leads to a quick proof of Theorem 6.8, which we sketch here.

Proof (sketch) Let $*$ be any AGM revision function. By Grove's theorem there exists a system of spheres S for any K that determines $*$. Let $\leq$ be the total preorder on W induced by S and define a CO*-model $M = \langle W, R, \varphi \rangle$ as follows:

1. $W = \cup\{S_i : S_i \in S\}$
2. φ is defined in the usual manner
3. wRv iff $v \leq w$.

It should be clear that R is reflexive, transitive and connected and that W contains all propositional valuations (as $W \in S$); hence, M is a CO*-model. Furthermore, M is a full revision model for suitable for K since S is centered on K.

In case A is inconsistent, $K_A^* = Cn(\bot)$ as does K_A^{*M}, since $M \models \overset{\forall}{\Box}\neg A$. So suppose $B \in K_A^*$ for some consistent A. By Grove's result,

$$s(A) \cap \|A\| \models B$$

so there is some sphere $s(A)$ such that for all A-worlds $w \in s(A)$, $w \models B$. By construction (and the fact that $s(A)$ is A-permitting)

$$M \models_w A \wedge \Box(A \supset B), \text{ so}$$

$$M \models \overset{\forall}{\Diamond}(A \wedge \Box(A \supset B)),$$

and thus $M \models A \overset{\text{KB}}{\longrightarrow} B$; i.e. $B \in K_A^{*M}$.

Conversely, assume $B \in K_A^{*M}$. Then $M \models A \overset{\text{KB}}{\longrightarrow} B$; and there exists some w such that

$$M \models_w A \wedge \Box(A \supset B).$$

Since $w \models A$, wRv for all $v \in s(A)$. Hence $v \models A \supset B$ for all $v \in s(A)$. This means

$$s(A) \cap \|A\| \models B$$

and by Grove's result, $B \in K_A^*$.

■

This lends further support to the claim that there is no distinction between our conditional approach to revision and the AGM model. Notice however that a system of spheres is defined so that for any sentence A there is some minimal A-permitting sphere. This points to a crucial divergence in the intuitions underlying spheres models and revision models, which do not necessitate minimal A-worlds in the accessibility relation R. The truth conditions for $B \in K_A^*$ with a spheres model state that B must hold at all minimal A-worlds. We circumvent this restriction by insisting $B \in K_A^*$ iff $A \wedge \Box(A \supset B)$ holds at some world, that there is some A-world such that B is true at all closer worlds.

The requirement that there exists a closest A-world in the ordering for revision has been dubbed the *Limit Assumption*, and has received considerable attention in the philosophical literature on counterfactual conditionals (Lewis 1973a; Lewis 1973b; Stalnaker 1968; Stalnaker 1984). Lewis (1973b) has argued that the Limit Assumption is inappropriate in general when discussing orderings of comparative similarity, or closeness, in the context of counterfactual reasoning. These considerations apply directly to deliberations of revision as well (see the next section). Consider the following example of Lewis:

"If I were over seven feet tall, I would play basketball."

Imagine the speaker of this utterance, say Pete, has a height of six feet. Intuitively, an ordering of comparative similarity (in this context) should rank possible worlds according to Pete's height; the nearer this value is to six feet, the closer to the actual state of affairs a possible world should be ranked. Now to evaluate the conditional (or to revise by its antecedent) is to ask whether Pete would play basketball at the closest world(s) where he is over seven feet tall. But intuitively there is no *closest* set of such worlds, only an infinite sequence in which Pete's height approaches a limit of seven feet. Lewis claims that this circumstance is perfectly acceptable, and that these truth conditions for counterfactuals are malformed. We should ask instead if, as we consider closer and closer worlds, we find a point where Pete plays basketball at all closer worlds (short of the seven foot limit).

Stalnaker argues that such considerations are not relevant in practice, though admitting: "Nothing I can think of in the concept of similarity ...would motivate imposing this restrictive formal structure [the Limit Assumption] in the ordering determined by a similarity relation" (Stalnaker 1984, p.140). For instance, assuming that such a fine-grained level of detail is not relevant to the conditional, we need not pick the closest antecedent-world, but merely the closest one(s) worlds that differ only in *relevant respects*. In this case, a world where Pete is seven-feet one-inch tall might suffice as a standard for evaluating the conditional. The basis for this argument seems to be the use of selection functions (see Chapter 2) for conditional semantics. Such a selection function may ignore irrelevant aspects of similarity and pick "relevantly closest worlds."

It might be the case, nevertheless, that such aspects as Pete's exact height are important (though perhaps not in this context). Stalnaker would then argue that the conditional is worded unsuitably, that Pete should assert

"If I were the shortest height greater than seven feet I would play basketball."

In such a circumstance the selection function would pick out no world and any consequent would follow vacuously. Unfortunately, this seems unnecessarily restrictive, and, we claim, an indictment of selection function semantics. Without the Limit Assumption a selection function fails in this case, since in attempting to determine the closest antecedent-world it comes up empty and makes all conditionals vacuously true. Certainly it seems some conditionals should remain true and some false in this case; for instance,

"If I were over seven feet tall I would be under seven feet tall"

seems absurd even without assuming the existence of a limiting world.

The Limit Assumption is a technical device postulated for the convenience of selection functions. Without such a requirement we can still provide adequate truth conditions for conditionals in terms of comparative similarity by following Lewis's suggestion. Hence, the "expressive" benefits of the Limit Assumption are negligible in comparison, and certainly not sufficient to outweigh the

ontological constraint it imposes. What our approach to revision discloses is that Lewis's account of truth conditions for counterfactuals can be defined or axiomatized in terms of modal operators.

Returning to Grove's system of spheres model for belief revision, the constraint that minimal A-permitting spheres exist is precisely the Limit Assumption and allows Grove to specify truth conditions for K_A^* in terms of a selection function, namely $s(A) \cap \|A\|$. The logic CO* and Theorem 6.7 show that Grove's restriction is unnecessary, that given the appropriate truth conditions (corresponding to $\overrightarrow{\Diamond}(A \wedge \Box(A \supset B))$) the AGM postulates are still satisfiable.

Of course, Grove's representation theorem gives the impression that this bit of philosophical exorbitance is unnecessary for modeling AGM revision functions, for any such function has a corresponding (set of) system of spheres (or CO* revision models) that satisfies the Limit Assumption. Furthermore, one could argue that, since our underlying logic is propositional and we are only interested in finite theories, expressing such concepts of infinite extent is impossible anyway.

The second claim is easier to dispose of, for concepts such as "least height greater than seven feet" and constraints on the ordering of possible worlds that reflect, say, a rational ordering are easily expressible in a first-order language, and we must suppose that any notion of revision should be applicable (in essential ways) to a first-order logic. We can imagine imposing constraints on the closeness ordering of possible worlds in our modal language, suitably extended to encompass first-order concepts. We first require some (necessarily incomplete, but sufficient for our purposes) theory of the real (or rational) numbers, say $\mathcal{R}$, that we assert holds at all worlds with the sentence $\overrightarrow{\Box}\mathcal{R}$. For this example, we could probably get by with a partial theory of $\leq$ together with the sentence

$$\overrightarrow{\Box}\forall x \exists y((x > 7) \supset (y > 7 \wedge y < x)). \tag{6.4}$$

We want Pete's height to be unique, so we assert

$$\overrightarrow{\Box}(\texttt{height(Pete)} = x \wedge \texttt{height(Pete)} = y \supset x = y). \tag{6.5}$$

To express the fact that worlds closer to Pete's actual height of six feet are considered more similar we assert that $\texttt{height(Pete)} = 6$ and

$$\overrightarrow{\Box}\forall y > 6[y < x \supset (\texttt{height(Pete)} = x \equiv \overleftarrow{\Box}\neg\texttt{height(Pete)} = y)]. \tag{6.6}$$

Given this background theory, we want to ask what follows from the revision of an objective theory K by

$$\texttt{height(Pete)} > 7, \tag{6.7}$$

in particular, if "Pete plays basketball" follows from this revision.

This leads us to the first objection, that any AGM revision can be modeled by a structure satisfying the Limit Assumption. A problem immediately crops up for the AGM theory and Grove's spheres models, for the *intensional constraints* we have proposed in this example cannot even be expressed in their language. We cannot assert a sentence such as that constraining the notion of similarity to be applied, sentence (6.6); nor can we even ensure in a natural manner that our revised theories contain certain sentences, such as sentences (6.4) or (6.5), that can be thought of as "integrity constraints" on the belief set. Other than the logical axiomatization it provides, the key advantage of regarding revision in terms of modal logic is the ability to express such intensional constraints on the revision process. Revising by sentence (6.7), since it is inconsistent with KB, need not even satisfy the constraints (6.4) or (6.5), if these are expressed within the (objective)

KB.[15] In CO* we can specify these constraints as premises in the logical language, as well as the restriction that the accessibility relation be "defined" in terms of Pete's height; they are just kept separate from the objective KB.[16] We will examine these integrity constraints further in the next section.

Returning to the Limit Assumption, Grove's result assures us that any AGM function has a model satisfying the assumption. This is the case precisely because the "traditional" language of revision cannot express intensional constraints, and hence cannot distinguish models that satisfy the Limit Assumption from those that do not. Any revision function that says Pete plays basketball when he is over seven feet tall has a model (system of spheres or CO*-model) with a minimal world where Pete is over seven feet tall. As Stalnaker claims, the existence of such a model might not be problematic in practice. But suppose we have a large knowledge base and a language that permits intensional constraints to be specified. In this knowledge base there might exist some theory $\mathcal{R}$ of the real numbers. A user, asking some query about Pete, might now wish to express the (quite natural) constraint that the notion of similarity appropriate in this context should respect Pete's varying height. On interaction with $\mathcal{R}$, this constraint ensures no minimal antecedent-world exists, and on Grove's truth conditions revising by the antecedent results in an inconsistent theory. Our truth conditions suffer from no such drawback, although in the cases where limiting worlds do exist the semantics coincides with Grove's.

6.3 A Framework for Subjunctive Queries

To this point we have investigated the revision conditional connective $\xrightarrow{\text{KB}}$ in the context of changing a knowledge base. However, the question of how to revise a KB is important not just in the presence of changing information, but also when we want to investigate questions of the form "What if A were true?" A subjunctive conditional $A > B$ is one of the form[17] "If A were the case then B would be true." Subjunctives have been widely studied in philosophy and it is generally accepted that (some variant of) the *Ramsey test* is adequate for evaluating the truth of such conditionals:

> First add the antecedent (hypothetically) to your stock of beliefs; second make whatever adjustments are required to maintain consistency (without modifying the hypothetical belief in the antecedent); finally, consider whether or not the consequent is true. (Stalnaker 1968, p.44)

The connection to belief revision is quite clearly spelled out in this formulation of the Ramsey test: to evaluate a subjunctive conditional $A > B$, we revise our beliefs to include A and see if B is believed. On this view, $A \xrightarrow{\text{KB}} B$ is nothing but a subjunctive conditional where the (implicit) KB represents our initial state of knowledge, and will be true exactly when $B \in K^*_A$, in accordance with the Ramsey test.[18]

A number of people have argued that counterfactual and subjunctive conditionals have an important role to play in AI, logic programming and database theory. Bonner (1988) has proposed

[15] Although, these could be explicitly "asserted" in the form $B \in K^*_A$.

[16] This is only because they are not propositional. It is possible to "only know" these sentences in conjunction with objective knowledge.

[17] At least, in "deep structure."

[18] It would seem, however, that subjunctives based on the Ramsey test and belief revision are not the only type of subjunctives. They may also be based on other types of "knowledge base revision," for example, update (see Section 7.1).

a logic for hypothetical reasoning in which logic programs or deductive databases are augmented with *embedded implications* of the form

$$A \leftarrow (B \leftarrow C).$$

Roughly, such an implication is read as "If adding C to the database causes B to be true, then A is true." The embedded rule $(B \leftarrow C)$ acts like a subjunctive premise "If C were true, then B would be," from which one can infer A. Bonner provides an interesting intuitionistic semantics and proof procedure for logic programs augmented with this feature, and the logic is extended to allow embedded universal quantification in (Bonner, McCarty and Vadaparty 1989).

Ginsberg (1986) has identified a number of areas in AI in which counterfactuals might play an important role in the semantic analysis of various tasks. For instance, in a planning domain a robot might ask of a state of belief

"If I removed the cup from my hand could I pick up the spoon?"

If this counterfactual $A > B$ is true, it suggests a natural regression of goals: in order to accomplish B the robot should take steps to ensure the truth of A. Similarly, in diagnosis we may consider the observation of a fault to counterfactually imply a diagnosis or explanation of that fault, as in

"If outputs x and y did not coincide then circuit C would be faulty."

This conditional must be interpreted as a counterfactual with respect to the specification of a device's (proper) behavior because such a description should entail the negation of the antecedent, making such a material conditional vacuously true.

Ginsberg proposes a system for reasoning about counterfactuals that is similar in spirit to Nebel's (1989) system for belief revision. Given some KB, to determine the counterfactual consequences of A we consider those maximal subsets of KB consistent with A, and add A to these. Notice, if A is consistent with KB then only one such subset (KB itself) exists and $A > B$ is true iff $A \supset B$ is entailed by KB, coinciding with our intuitions about consistent revision. Certain subsets can be preferred to others through some ordering $<$ on these subsets, similar to partial meet revision. However, Ginsberg's model of counterfactual reasoning suffers from the same defect as Nebel's revision model, specifically sensitivity to the syntactic structure of KB.

Jackson (1989) considers the problems with Ginsberg's approach and presents a model-theoretic system BERYL that addresses these difficulties. We assume a finitary propositional language and a theory K that must be revised by A. As in the possible models approach to update of (Winslett 1988), those worlds that satisfy A and differ minimally (with respect to set inclusion of atoms) from some world in $\|K\|$ are models of the revised state of affairs. To satisfy the postulate of consistent revision (R4) (which distinguishes BERYL from the possible models approach) Jackson, in a somewhat *ad hoc* fashion, requires that

$$\|K_A^*\| = \|K\| \cap \|A\|$$

whenever A is consistent with K. This would seem to void either the claim that set inclusion of atoms reflects some notion of comparative similarity or that BERYL respects this notion.

6.3.1 Belief Revision and Subjunctive Conditionals

These systems both take seriously the idea that counterfactuals are intimately tied to belief revision. However, this connection had not gone unappreciated by the revision community. Gärdenfors

(1978a, as presented in (1988)) provides an explicit postulate for revision and conditional reasoning based on the Ramsey test. If we assume that conditionals can be part of our belief sets then a concise statement of the Ramsey test is

(RT) $A > B \in K$ iff $B \in K_A^*$.

Gärdenfors also describes a formal semantics for conditionals in terms of *belief revision systems* (as discussed in the next section). By imposing certain constraints on these models in the form of postulates (R1) through (R8), and using (RT) to evaluate conditionals, Gärdenfors comes up with a conditional logic based on a revision "style" semantics that corresponds exactly to Lewis's (1973a) counterfactual logic VC. The logic VC is Lewis's "official logic of counterfactuals," an axiomatization of which is given below.

(1) From A, $A \supset B$ infer B

(2) From $C_1 \wedge \cdots \wedge C_n \supset B$ infer $((A > C_1) \wedge \cdots \wedge (A > C_n)) \supset (A > B)$ for any $n \geq 1$

(3) All truth functional tautologies

(4) $A > A$

(5) $(\neg A > A) \supset (B > A)$

(6) $(A > \neg B) \vee (((A \wedge B) > C) \equiv (A > (B \supset C)))$

(7) $(A > B) \supset (A \supset B)$

(8) $(A \wedge B) \supset (A > B)$

Let KB be as usual a set of beliefs representing our knowledge of the world. We also expect there to be some conditional beliefs among these that constrain the manner in which we are willing to revise our (objective) beliefs. These take the form $\alpha \xrightarrow{\text{KB}} \beta$ (or $\alpha > \beta$), and will be referred to as *subjunctive premises*. By a *subjunctive query* we intend something of the form "If A were true, would B hold?" In other words, is $A > B$ a consequence of our beliefs and subjunctive premises?

Given the connection between VC and belief revision, and assuming the Ramsey test is an appropriate truth test for subjunctives, it would appear that VC is exactly the logical calculus required for formulating subjunctive queries. However, we have misrepresented the Gärdenfors result to a certain degree; in fact, his semantics does not account for the postulate of consistent revision (R4). It is excluded because it results in *triviality* (see the next section) and, together with the other postulates, is far too strong to be of use. Because (R4) is unaccounted for in VC, it is inadequate for the representation of certain subjunctive queries.

Example 6.7 Suppose $KB = \{B\}$, a belief set consisting of a single propositional letter. If we were to ask "If A then B?" intuitively we would expect the answer YES, when A is some distinct atomic proposition. With no constraints (such as $A > \neg B$), the postulate of consistent revision should hold sway and revising by A should result in $KB' = \{A, B\}$. Hence, $A > B$ should be true of KB. Similarly, $\neg(A > C)$ should also be true of KB for any distinct atom C.

■

In VC there is no mechanism for drawing these types of conclusions. At most one could hope to assert B as a premise and derive $A > B$ or $\neg(A > C)$, but neither of

$$B \vdash_{VC} A > B \quad \text{or}$$

$$B \vdash_{VC} \neg(A > C)$$

is true, nor should they be. It *should* be the case that if A is *consistent with our beliefs* that $A > B$ holds, but merely asserting B doesn't carry this force. In a sense, when B is a premise we mean "B is believed," but this does not preclude the possibility of A, or $\neg A$, or C, or anything else being believed. When $KB = \{B\}$ we intend something stronger, namely that "B is *all* that is believed." Because B is the only sentence in KB, we convey the added information that, say, neither A nor $\neg A$ is believed. In Levesque's (1990) terminology, we *only know* KB.

To only know some sentence is to both know (or believe) A and to know nothing more than A. To know A is to restrict one's set of epistemic possibilities to those states of affairs where A is true. If some $\neg A$-world were considered possible an agent could not be said to know A, for the possibility of $\neg A$ has not been ruled out. To know nothing more than A is to include all possible A-worlds among one's set of epistemic possibilities. Adding knowledge to a belief set is just restricting one's set of epistemic possibilities to exclude worlds where these new beliefs fail, so if some A-world were excluded from consideration, intuitively an agent would have some knowledge other than A that ruled out this world. We return to these ideas in Section 6.5.

In our logic CO* we have precisely the mechanism for stating that we only know a knowledge base. In CO*-structures for revision we consider the set of minimal worlds to represent our knowledge of the actual world. Exactly those possible worlds consistent with our beliefs KB are minimal in any KB-revision model; this is precisely what the sentence $O(KB)$ asserts. It says that KB is believed (since only KB-worlds are minimal) and that KB is all that is believed (since only minimal worlds are KB-worlds).

Returning to the subjunctive query $A \xrightarrow{\text{KB}} B$, we take this analysis to mean that

$$B \vdash_{CO*} A \xrightarrow{\text{KB}} B$$

is not the proper formulation of the query. This derivation is not valid (just as it is not in VC). Our analysis suggests that we ought to ask if $A \xrightarrow{\text{KB}} B$ holds if we only know B. In fact this is the case: both

$$O(B) \vdash_{CO*} A \xrightarrow{\text{KB}} B \quad \text{and}$$

$$O(B) \vdash_{CO*} \neg(A \xrightarrow{\text{KB}} C)$$

are legitimate derivations.

This leads to an obvious framework for subjunctive query answering, given a set of beliefs. Our knowledge of the world is divided into two components, a set KB of *objective* or propositional facts or beliefs, and a set S of subjunctive conditionals acting as premises, or constraints on the manner in which we revise our beliefs. To ask a subjunctive query Q of the form $\alpha \xrightarrow{\text{KB}} \beta$ is to ask if β would be true if we believed α, given that our *only* current beliefs about the world are represented by KB, and that our deliberations of revision are constrained by subjunctive premises S. The expected

answers YES, NO and UNK (unknown) to Q are characterized as follows.

$$ASK(Q) = \begin{cases} \texttt{YES} & \text{if } \{O(KB)\} \cup S \models_{CO*} Q \\ \texttt{NO} & \text{if } \{O(KB)\} \cup S \models_{CO*} \neg Q \\ \texttt{UNK} & \text{otherwise} \end{cases}$$

Objective queries about the actual state of affairs (or, more precisely, our beliefs) can be phrased as a Q of the form $\top \xrightarrow{\text{KB}} \beta$ where β is the objective query of interest. It's easy to see that

$$ASK(Q) = \texttt{YES} \quad \text{iff} \quad \vdash_{CO*} KB \supset \beta.$$

The ability to express that *only* a certain set of sentences is believed allows us to give a purely logical characterization of subjunctive queries of a knowledge base. The logic VC seems adequate for reasoning from subjunctive premises and for deriving new conditionals, but it cannot account for the influence of factual information on the truth of conditionals in a completely satisfying manner; for it lacks the expressive power to enforce compliance with postulate (R4). In fact, it is not hard to verify that the axioms for VC are each valid in CO* if we replace nonsubjunctive (factual) information (say α) by statements to the effect that α is believed (in CO*, belief in α is expressed as $\Diamond\Box\alpha$; see Section 6.5). The approaches of Ginsberg and Jackson take VC to be the underlying counterfactual logic. Indeed, their approaches (under certain assumptions) satisfy the Lewis axioms. However, they recognize that the ability to only know a knowledge base is crucial for revision and subjunctive reasoning, an expressive task not achievable in VC. Therein lies the motivation for their extra-logical characterizations, and the underlying idea that KB is representable as a set of sentences or set of possible worlds from which we construct new sets in the course of revision. Winslett's (1988) possible models approach to update has a similar extra-logical quality. CO* can be viewed as a logic in which one can capture just this process.

Naturally, important distinctions between the proposals of Ginsberg and Jackson and ours do exist. For instance, our characterization is not swayed by the syntactic form of KB; and a "consistent" ordering on states of affairs is adhered to for any type of revision, in contrast to BERYL, which requires that its ordering of set inclusion be ignored during consistent revision.

6.3.2 Integrity Constraints

Often only certain states of knowledge, certain belief sets, are permissible. The concept of *integrity constraints*, widely studied in database theory, is a way to capture just such conditions. For a database (or in our case, a belief set) to be considered a valid representation of the world, it must satisfy these integrity constraints. For instance, we might not consider feasible any belief set in which certain commonsense laws of physics are violated; or a database in which there exists some student with an unknown student number might be prohibited.

This apparently straightforward concept actually has several distinct interpretations. Reiter (1990) surveys those and proposes the definition we favor, which essentially asserts that an integrity constraint C should be entailed by KB. The distinguishing characteristic of Reiter's definition is that integrity constraints can be phrased using a modal knowledge operator that refers to "what is known by the database." This connective is given a semantics, and satisfaction of an integrity constraint is defined using an implicit appeal to the concept of only knowing. We can phrase the definition within our language as follows (though Reiter's underlying logic is first-order).

Definition 6.8 (Reiter 1990) Let $C \in \mathbf{L}_B$ and $KB \subseteq \mathbf{L}_{CPL}$. KB satisfies integrity constraint C iff $O(KB) \vdash_{CO*} C$.

Typically, C will contain operators referring to the knowledge contained in KB (see (Reiter 1990) for examples). For our purposes, since we do not intend to give a full account of what a knowledge base knows,[19] we will assume constraints are propositional and that KB satisfies C just when KB entails C.

As emphasized in (Fagin, Ullman and Vardi 1983) and (Winslett 1990), integrity constraints are particularly important when updating a database. Any new database (or belief set) should satisfy these constraints, therefore any reasonable model of update or revision must explicitly account for integrity constraints (Winslett 1990). Consider the following example.

Example 6.8 Let the constraint C, that a particular department has only one chair, be expressed as

$$\texttt{chair(x,d)} \wedge \texttt{chair(y,d)} \supset \texttt{x=y} \tag{6.8}$$

Suppose we want to update KB with

$$\texttt{chair(Ken,DCS)} \vee \texttt{chair(Maria,DCS)}.$$

If (6.8) is consistent with KB, the constraint C can be placed in KB and will ensure that exactly one of Ken or Maria is the chair of Computer Science.[20] Suppose, on the contrary, that

$$KB = \{\texttt{chair(Derek,DCS)}\}$$

so the new fact is inconsistent. The constraint can no longer be enforced in the updated KB', for nothing about (6.8) says it must be true in the revised state of affairs. Simply viewing the constraint as an objective fact is inappropriate.

■

This example illuminates the need for integrity constraints to be expressed intensionally. They refer not only to the actual world, but to *all (preferred) ways in which we may view the world.* We can ensure the satisfaction of a constraint by adding it to KB.[21] This will not, however, guarantee that the constraint remains consistent when KB is revised. Since revision often entails giving up certain facts, this "constraint" expressed as a fact in KB can be lost as easily as any other. We want those revisions to be preferred that keep the constraint C in the revised KB.

As mentioned in the previous section, we can ensure a revised belief set or database satisfies a constraint C by asserting $\Box C$ as a premise in our background theory (on the same level as $O(KB)$). This has the effect of ensuring *any* possible worlds ever considered satisfy C. This tack has two disadvantages. First, it takes us from the realm of CO*, and requires the logic CO. Thus we lose the uniformity of our models and require a proper subset of the axiom schemata corresponding to CO* (and perhaps lose a more uniform proof procedure). Second, it might be too strong an assertion in many applications. Such a statement will force any revision of KB by a fact inconsistent with C to result in the inconsistent belief set $Cn(\bot)$. In certain (maybe most) circumstances, we can imagine a constraint C ought to be satisfied if at all possible; but if it cannot be we should not be forced into inconsistency.

---

[19] A rough characterization would be fairly straightforward, however. We can replace Reiter's knowledge modality K with the sequence $\widetilde{\Diamond}\Box$, or even with $KB\,\supset$, as we discuss in the next section in terms of autoepistemic logic. In the context of only knowing, KB knows (propositional) α just when KB entails α.

[20] We adopt the Unique Names Hypothesis (Reiter 1978a), so, e.g., $\texttt{Ken} \neq \texttt{Maria}$ is true of KB. Propositionally this can be written as $\texttt{chairMaria} \supset \neg\texttt{chairKen}$.

[21] This is true only of simple propositional constraints. These are not the only constraints we will want to enforce in general (Reiter 1990).

Instead of abolishing $\neg C$-worlds outright, we'd like to say all C-worlds are "preferred" to $\neg C$-worlds, or they are closer to the actual state of affairs, theory K, than any world violating the constraint. Such a condition is expressible as

$$\vec{\Box}(C \supset \Box C).$$

To see this, imagine some $\neg C$-world v is closer than some C-world w, in a model M of K. Then wRv and $M \not\models_w C \supset \Box C$.

Of course, we are often concerned with a set of constraints $C = \{C_1, \cdots, C_n\}$. A set of sentences of the form

$$\vec{\Box}(C_1 \supset \Box C_1), \cdots, \vec{\Box}(C_n \supset \Box C_n)$$

will be CO*-inconsistent if any C_i, C_j are pairwise contingent, since it is impossible to satisfy the condition, for some $w \models C_i \wedge \neg C_j$ and $v \models \neg C_i \wedge C_j$, that v is closer than w and w is closer than v. For a set C of multiple constraints, we use C to denote their conjunction

$$\vec{\Box}(C \supset \Box C) \quad \text{where} \quad C = \bigwedge_{i \leq n} C_i.$$

We now define a revision model with integrity constraints. We assume C is satisfiable.

Definition 6.9 *M is a revision model for K with weak integrity constraints $C_1, \cdots, C_n$* iff *M is a revision model for K and $M \models \vec{\Box}(C \supset \Box C)$ where C is the conjunction of the constraints. We will often say that M is a revision model with constraints C where C is understood to be either the set or conjunction of the individual constraints. The set $\{\vec{\Box}(C \supset \Box C)\}$ will be denoted WIC.*

Theorem 6.13 *Let M be a revision model for K with weak integrity constraints C. Then $K_A^{*M} \models C$ for all A consistent with C.*

Thus we validate the definition of integrity constraint. If a sentence A is consistent with C it must be that revising by A results in a belief set that satisfies C. Of course, this requires that the original belief set must also satisfy the integrity constraints.

Corollary 6.14 *Let M be a revision model for K with weak integrity constraints C. Then $K \models C$.*

Corollary 6.15 *If $KB \not\models C$ then $\{O(KB)\} \cup WIC$ is CO*-inconsistent.*

These corollaries state that if KB does not entail constraints C, it can have no revision model satisfying WIC. That is, if the initial KB violates C we cannot postulate a model in which KB satisfies C no matter how it is revised. Our concern is not with the (traditional) question of when KB satisfies C, but instead with the question of when KB will satisfy C after arbitrary revisions.[22]

In order to get off the ground, we assume the original KB in our examples satisfy C (i.e. $KB \models C$). The simplest way to do this is to "throw in" the constraints, assuming they are part of

[22]The problem we address here might also differ from the classical problem of integrity constraints in the sense that we are dealing with revisions of databases rather than updates of databases (see Section 3.3 and 7.1). This suggests that, just as there are (at least) two distinct types of theory change, there are also two kinds of integrity constraint.

KB. Otherwise $O(KB)$ will require that $\neg C$-worlds be among those considered to model the actual world. Often, integrity constraints and objective knowledge are kept separately. In our examples, we will always consider KB to include as an objective fact any constraint C,[23] although we may not explicitly write C as a member of KB.

Example 6.9 Let $KB = \{\texttt{chair(Derek,DCS)}\}$ and

$$C = \{\texttt{chair(x,d)} \wedge \texttt{chair(y,d)} \supset \texttt{x=y}\}.$$

Suppose we want to revise KB with

$$\texttt{chair(Ken,DCS)} \vee \texttt{chair(Maria,DCS)}.$$

Then from $\{O(KB)\} \cup WIC$ we can derive in CO*

$$\texttt{chair(Ken,DCS)} \vee \texttt{chair(Maria,DCS)} \xrightarrow{\texttt{KB}}$$

$$\texttt{chair(Ken,DCS)} \equiv \neg\texttt{chair(Maria,DCS)}.$$

■

This definition of integrity constraint has the unappealing quality of being unable to ensure that as many constraints as possible be satisfied. For instance, if some update A violates some C_i of C, then revision by A is not guaranteed to satisfy other constraints. Only worlds that satisfy *all* constraints in C are preferred. For this reason we included the qualifier "weak" in the definition. In general, we want other members of C to be satisfied even though a certain C_i may be violated.

Example 6.10 Let $KB = \{\texttt{chair(Derek,DCS)}\}$ and $C = \{C_1, C_2\}$, where

$$C_1 = \texttt{chair(x,d)} \wedge \texttt{chair(y,d)} \supset \texttt{x=y},$$

$$C_2 = \texttt{chair(x,DCS)} \equiv \texttt{teachnocourse(x)}.$$

C_2 is intended to constrain the database such that only the chair of Computer Science teaches no courses. From $\{O(KB)\} \cup WIC$ we cannot derive in CO*

$$\texttt{chair(Ken,DCS)} \wedge \texttt{chair(Maria,DCS)} \xrightarrow{\texttt{KB}} \texttt{teachnocourse(Ken)},$$

nor can we derive

$$\texttt{chair(Ken,DCS)} \wedge \texttt{chair(Maria,DCS)} \xrightarrow{\texttt{KB}} \texttt{teachnocourse(Maria)}.$$

■

So once C_1 has been violated, we cannot assume C_2 will be satisfied.

In order to ensure a more reasonable behavior on the part of integrity constraints, we'd like to insist that as many of them as possible be satisfied. This can be accomplished by asserting that worlds in which a certain subset of constraints $C = \{C_1, \cdots, C_n\}$ is violated are preferred to any world in which a larger subset of these is violated, where by preferred we intend worlds are lower

[23] Although, from previous examples we have seen this is not *sufficient* for C to be considered an integrity constraint.

in our ranking of closeness. Let $S_1, S_2, \ldots, S_{2^n}$ be an enumeration of the subsets of C. We will use S_i to denote the sentence

$$S_i = \bigwedge\{C_j : C_j \in S_i\}.$$

$\overline{S_i}$ will denote its "complement"

$$\overline{S_i} = \bigwedge\{\neg C_j : C_j \notin S_i\}.$$

We can think of $S_i \wedge \overline{S_i}$ as stating that some state of affairs satisfies exactly the constraints in S_i. Suppose $w \models S_i \wedge \overline{S_i}$. To say w is preferred to any world that violates strictly more constraints is to assert that, for every $S_j \subset S_i$,

$$\Box((S_i \wedge \overline{S_i}) \supset \Box\neg\overline{S_j}).$$

To see this, suppose this sentence is violated. Then for some such w there is a world v that is at least as close as w at which $\overline{S_j}$ holds. This means some superset, $C - S_j$, of the constraints violated at w, $C - S_i$, is violated at v, yet v is equally or more preferred. We can state this more concisely, without quantifying over all subsets of S_i as follows:

$$\Box((S_i \wedge \overline{S_i}) \supset \Box(\overline{S_i} \supset S_i)).$$

The *(strong) integrity constraints* determined by C are specified by the set

$$IC = \{\Box((S_i \wedge \overline{S_i}) \supset \Box(\overline{S_i} \supset S_i)) : i \leq 2^n\}.$$

Definition 6.10 M is a *revision model for K with (strong) integrity constraints C* iff M is a revision model for K and $M \models IC$.

Once again, we assume C is satisfiable. The set IC determines a partial order on the subsets of C that respects set inclusion. When revising by a sentence A, in general there will be several maximal subsets $S_i \subseteq C$ consistent with A. The revision of K by A will result in either the belief in one such subset of C or the disjunction of several such maximal S_i, if the revision model fails to "complete" the partial order determined by sentence IC.

Theorem 6.16 *Let M be a revision model for K with strong integrity constraints C. Let $\mathcal{S}$ be the collection of maximal subsets $S_i \subseteq C$ such that S_i is consistent with A. For some $S \subseteq \mathcal{S}$, it is the case that $K_A^{*M} \models \bigvee S$.*

Corollary 6.17 *Let M be a revision model for K with strong integrity constraints C. Then $K \models C$.*

Corollary 6.18 *If $KB \not\models C$ then $\{O(KB)\} \cup IC$ is CO*-inconsistent.*

Example 6.11 Let $KB = \{\texttt{chair(Derek,DCS)}\}$ and $C = \{C_1, C_2\}$, where

$$C_1 = \texttt{chair(x,d)} \wedge \texttt{chair(y,d)} \supset \texttt{x=y},$$

$$C_2 = \texttt{chair(x,DCS)} \equiv \texttt{teachnocourse(x)}.$$

From $\{O(KB)\} \cup IC$ we can derive in CO* both

$$\texttt{chair(Ken,DCS)} \wedge \texttt{chair(Maria,DCS)} \xrightarrow{KB} \texttt{teachnocourse(Ken)}$$

and

$$\texttt{chair(Ken,DCS)} \wedge \texttt{chair(Maria,DCS)} \xrightarrow{\text{KB}} \texttt{teachnocourse(Maria)}.$$

■

The specification of integrity constraints as the set IC seems much more reasonable, but gives equal weight to each constraint in C. Fagin, Ullman and Vardi (1983) have argued that sentences in a database can have different priorities and that updates should respect these priorities by "hanging on to" sentences of higher priority whenever possible during revision. For instance, as we have argued throughout, sentences with more information content should be retained in a belief set if it is possible to give up those with less. In their theory of updates, sentences in a database are *tagged* with nonnegative integer priorities, and sentences are removed from the database during the update process according to this ranking.

Consider the previous examples where constraints assert that a department has one chair and that the chair of Computer Science is the only person without a course to teach. It can be the case that certain information cannot satisfy both of the constraints, but could satisfy either one singly — for example, when we learn that Maria is the chair and Ken has no course load. It might also be that we prefer to violate the constraint that a non-chair faculty member teaches no course in deference to the fact that Computer Science has only one chair; it seems more plausible that Ken has cut a special deal than that he is another chair.

Suppose that the set $C = \{C_1, \cdots, C_n\}$ is now an *ordered* set of integrity constraints with C_i having higher priority than C_j whenever $i < j$. So we prefer C_i when a conflict arises with C_j. Expressing the fact that C_1 is preferred at all costs is merely asserting that C_1 is a weak integrity constraint, so

$$\overset{\text{H}}{\Box}(C_1 \supset \Box C_1)$$

should be satisfied. Since C_2 has the next highest priority, it should be satisfied when possible unless that entails violating C_1. This means any world that satisfies C_2 and C_1 is preferred to any world that does not. In other words,

$$\overset{\text{H}}{\Box}((C_1 \wedge C_2) \supset \Box(C_1 \wedge C_2)).$$

In conjunction with the first sentence, this implies that any preferred world that violates C_2 will satisfy C_1. Let P_i denote the conjunction of the i highest priority integrity constraints

$$P_i = C_1 \wedge C_2 \wedge \cdots C_i.$$

The the set of *prioritized integrity constraints* is specified as

$$ICP = \{\overset{\text{H}}{\Box}(P_i \supset \Box P_i) : i \leq n\}.$$

Definition 6.11 *M is a revision model for K with prioritized integrity constraints $C_1, \cdots, C_n$ iff* M is a revision model for K and $M \models ICP$.

Theorem 6.19 *Let M be a revision model for K with prioritized integrity constraints $C_1, \cdots, C_n$. Then $K_A^{*^M} \models C_i$ whenever A is consistent with the conjunction of all constraints C_j, $j \leq i$.*

Corollary 6.20 *Let M be a revision model for K with prioritized integrity constraints C. Then $K \models C$.*

Corollary 6.21 *If $KB \not\models C$ then $\{O(KB)\} \cup ICP$ is CO*-inconsistent.*

Example 6.12 Let $KB = \{\texttt{chair(Derek,DCS)}\}$ and $C = \{C_1, C_2\}$, where

$$C_1 = \texttt{chair(x,d)} \wedge \texttt{chair(y,d)} \supset \texttt{x=y},$$

$$C_2 = \texttt{chair(x,DCS)} \equiv \texttt{teachnocourse(x)}.$$

From $\{O(KB)\} \cup ICP$ we can derive in CO*

$$\texttt{teachnocourse(Ken)} \wedge \texttt{chair(Maria,DCS)} \xrightarrow{\text{KB}}$$

$$\texttt{chair(x,DCS)} \equiv \texttt{x = Maria}.$$

■

It may be that certain constraints should have equal priority, that none in a certain subset of C have priority over others. It should be easy to see that the strong version of integrity constraints and the prioritized version can be combined. One needs just assert the premises IC for this subset of equal-priority constraints and adjust the set ICP to reflect this. We leave the details to the reader.

6.3.3 Epistemic Entrenchment

In (Fagin, Ullman and Vardi 1983) the tags on sentences are not meant for integrity constraints alone. In fact, the only sentences tagged with 0 are called integrity constraints, and anything that can be viewed as violable (to greater or lesser degree) is not correctly called a constraint. If we adopt this perspective, we see that prioritized integrity constraints are really just prioritized sentences in a belief set reflecting the degree of epistemic entrenchment associated with each such belief. Thus, what we have termed prioritized integrity constraints can also be viewed as premises constraining belief revision to respect pragmatic concerns. Instead of such premises being in strictly conditional form, it might be the case that notions of entrenchment are more naturally specified in this way. A conditional $A \xrightarrow{\text{KB}} B$ ensures that $A \wedge B$ is more plausible (ranked lower) than $A \wedge \neg B$. Priorities can express this idea unconditionally, for if C_2 has lower priority than C_1, then $\neg C_2$ is considered more plausible than $\neg C_1$. Reviewing the examples in the last section regarding the Limit Assumption, it seemed quite natural to represent certain constraints on belief revision in a non-conditional form. For instance, the theory of rational numbers (from which we reasoned about height) ought to hold at "any" world, and is given highest priority (its negation is given lowest plausibility). Certainly beliefs about subjects like commonsense physical laws ought to have a high degree of entrenchment in general, as opposed to various contingent facts.

To say a belief C_1 is an integrity constraint of higher priority than C_2 implies that C_1 is more epistemically entrenched than C_2, and that its negation is less plausible. However, the converse is not always the case. We can have a belief C_1 be more entrenched than C_2, yet C_1 need not be an integrity constraint. To say C_1 is an integrity constraint is a much stronger assertion, for it means that all states of affairs where C_1 holds are more plausible than any state of affairs where it does not (discounting higher priorities). But this is precisely what is intended in many circumstances,

such as the ones we've examined above. Thus, not only is entrenchment expressible modally in CO*, but so is the stronger concept of integrity constraint.

Of course, one might want to express an *absolute* notion of entrenchment rather than a conditional one, without relying on integrity constraints, which can be much too strong. That is, we'd like to be able to assert or ask something like "A is more entrenched than B," or "A is more firmly held than B." Not surprisingly, entrenchment can be expressed as well. Recall Grove's ordering $\leq_G$ is in a sense the dual of entrenchment and is easier to motivate; so we start there.

When $A \leq_G B$ it is intended that A is at least as close to the actual world as B, or A is at least as plausible as B. But this just means the closest A-world is as close as the closest B-world; the rank of $\min(A)$ (hypothetically speaking) is no greater than that of $\min(B)$. This is the case exactly when every B-world can see some A-world, given our total ordering of clusters; every B-world sees $\min(B)$, which in turn sees $\min(A)$. This leads to the following obvious definitions.

Definition 6.12 Let M be a CO*-model. We say A is *at least as plausible as* B (in context M) iff $M \models \ddot{\Box}(B \supset \Diamond A)$. The *plausibility ordering* determined by M is $\leq_{PM}$ given by

$A \leq_{PM} B$ iff A is as plausible as B.

A is *more plausible* than B iff $A \leq_{PM} B$ and not $B \leq_{PM} A$.[24]

Definition 6.13 Let M be a CO*-model. The *entrenchment ordering* determined by M, denoted $\leq_{EM}$, is given by

$B \leq_{EM} A$ iff $\neg B \leq_{PM} \neg A$.

A is *at least as entrenched as* B iff $B \leq_{EM} A$. A is *more entrenched* than B iff $B \leq_{EM} A$ and not $A \leq_{EM} B$.

Corollary 6.22 *$B \leq_{EM} A$ iff $M \models \ddot{\Box}(\neg A \supset \Diamond \neg B)$.*

Of course, we must justify the use of the terms plausibility and entrenchment. Recall the Grove postulates, (G1)–(G5), and the entrenchment postulates, (E1)–(E5), as presented in Chapter 3. These results show that, as well as representing the process of revision with the subjunctive conditional $\xrightarrow{KB}$, CO* can simultaneously be used to represent constraints on the revision process in the form of entrenchment and plausibility of various sentences.

Theorem 6.23 *Let M be a CO*-model. Then $\leq_{PM}$ satisfies the Grove postulates (G1)–(G5).*

Theorem 6.24 *Let $\leq_G$ be a Grove ordering satisfying (G1)–(G5). Then there exists a CO*-model M such that the plausibility ordering $\leq_{PM}$ determined by M is $\leq_G$.*

Corollary 6.25 *Let M be a CO*-model. Then $\leq_{EM}$ satisfies the entrenchment postulates (E1)–(E5) of (Gärdenfors 1988).*

Corollary 6.26 *Let $\leq_E$ be an entrenchment ordering satisfying (E1)–(E5). Then there exists a CO*-model M such that the entrenchment ordering $\leq_{EM}$ determined by M is $\leq_E$.*

Thus we see that a given CO*-model (or theory) determines (or partially determines) an ordering of entrenchment. Not only can we constrain our theories to satisfy subjunctive premises and integrity constraints, but we can also specify directly statements of relative entrenchment and plausibility within the logical language.

[24] We use $\leq_{PM}$ to indicate greater plausibility rather than $\geq_{PM}$ to remain consistent with Grove's notation, as well as Pearl's Z-ranking, where lower ranked elements are more normal or more plausible.

6.4 Triviality

A drawback of the counterfactual logic VC for subjunctive reasoning is its inability to account for factual information that may constrain the revision process. This a direct result of failing to satisfy the criterion for consistent revision

(R4) If $\neg A \notin K$ then $K_A^+ \subseteq K_A^*$.

Gärdenfors (1988) has shown that if the postulate (RT) is added to the conditions on revision and certain others are weakened (like (R4)), then the resulting logic is precisely VC. Let $\mathbf{L}_{Cond}$ be the extension of $\mathbf{L}_{CPL}$ that includes unnested occurrences of the conditional connective $>$ and boolean combinations of such conditionals (much like the language $\mathbf{L}_C^-$ in Chapter 4).

Definition 6.14 (Gärdenfors 1988) A *belief revision system* is a pair $\langle \mathbf{K}, * \rangle$ where $\mathbf{K}$ is a collection of belief sets (in the extended language $\mathbf{L}_{Cond}$) and $*$ is a revision function mapping $\mathbf{K} \times \mathbf{L}_{Cond}$ to $\mathbf{K}$. $\mathbf{K}$ must be closed under expansions and $*$ must satisfy (RT); that is $B \in K_A^*$ iff $A > B \in K$, for all $K \in \mathbf{K}$.

By imposing certain constraints on this semantic system, in particular on $*$, Gärdenfors derives a series of stronger and stronger logics associated with various axioms. For instance, insisting that (R2) be satisfied, $A \in K_A^*$, corresponds to the axiom

$$A > A.$$

The constraint (R6) that $K_A^* = K_B^*$ iff $\vdash A \equiv B$ is captured by

$$(A > B \wedge B > A) \supset (A > C \supset B > C).$$

These can also be expressed in CO* as $A \xrightarrow{\text{KB}} A$ and

$$\Box(A \equiv B) \supset (A \xrightarrow{\text{KB}} C \supset B \xrightarrow{\text{KB}} C),$$

both of which are theorems. Some AGM postulates are too strong to be expressed in $\mathbf{L}_{Cond}$, like (R5): if $\not\vdash A$ then $K_A^* \neq Cn(\bot)$. Gärdenfors however gives a necessary (though not sufficient) axiom for (R5)

$$A > \neg A \supset B > \neg A.$$

In $\mathbf{L}_{Cond}$ the required notion of logical consistency cannot be expressed; however, in CO* this condition corresponds to

$$\Diamond A \supset \neg(A \xrightarrow{\text{KB}} \bot).$$

Postulate (R4) cannot be accommodated within this semantics (hence, neither can the stronger (R8)) and Gärdenfors provides a weak version of it.

(WR4) If $A \in K$ and $K_A^* \neq Cn(\bot)$ then $K \subseteq K_A^*$.

This condition is specified by the axiom

$$A \wedge B \supset A > B,$$

and states that if A is a belief, revision by A will cause no beliefs to be given up. Unfortunately this carries nowhere near the force of (R4), which refers to the case where A is *consistent* with K,

rather than *in* K. As discussed in the previous section, this stronger condition cannot be expressed in $\mathbf{L}_{Cond}$ for it requires the ability to capture logical consistency, or *only knowing* K. VC lacks this expressive power.

The logic determined by these conditions on a belief revision system is shown to be exactly VC.

Theorem 6.27 (Gärdenfors 1988) *A is valid with respect to the class of belief revision systems satisfying certain conditions*[25] *iff* $\vdash_{VC} A$.

Suppose we constrain a belief revision system to satisfy (R4). This requirement cannot be expressed in the logic VC, but intuitively it should characterize a certain desirable behavior. A rather discouraging result of (Gärdenfors 1986) is the following.

Theorem 6.28 (Gärdenfors 1988) *No* nontrivial *belief revision system satisfies (R2), (R4), (R5) and (RT).*[26]

A *nontrivial* system is one in which some belief set K does not contain $\neg A$, $\neg B$ or $\neg C$, for three pairwise disjoint sentences A, B, C (where a disjoint pair A, B is any such that $\neg(A \wedge B)$ is unsatisfiable).

Let's say a belief set is AB-ignorant just when it holds no belief about either A or B; that is, when none of $A \vee B$, $\neg A \vee B$, $A \vee \neg B$, $\neg A \vee \neg B$ is in K (Rott 1989). It is easy to see that a trivial belief revision system can have no AB-ignorant belief set.[27] Rott (1989) presents a proof of Gärdenfors's triviality result that is somewhat more perspicuous, and we detail his derivation here.

Proof (Rott 1989) Let K be an AB-ignorant belief set. Then

(1)	$A \vee B \in K^+_{A \vee B}$	premise
(2)	$A \notin K^+_{A \vee B}$	K is AB-ignorant
(3)	$A \vee B \in (K^+_{A \vee B})^*_{\neg A}$	$(R2), (R4), (1)$
(4)	$\neg A \in (K^+_{A \vee B})^*_{\neg A}$	$(R2)$
(5)	$B \in (K^+_{A \vee B})^*_{\neg A}$	(3), (4)
(6)	$\neg A > B \in K^+_{A \vee B}$	$(5), (RT)$
(7)	$K^+_{A \vee B} \subseteq K^+_A$	property of $+$
(8)	$\neg A > B \in K^+_A$	(6), (7)
(9)	$\neg A > \neg B \in K^+_A$	derived as is (8)
(10)	$B, \neg B \in (K^+_A)^*_{\neg A}$	(8), (9)
(11)	$(K^+_A)^*_{\neg A}$ is inconsistent	(10)
(12)	$\neg A$ is consistent	K is AB-ignorant
(13)	K^+_A is consistent	K is AB-ignorant
(14)	$(K^+_A)^*_{\neg A}$ is consistent	(12), (13)

Clearly (11) and (14) stand in direct contradiction. Hence K cannot be AB-ignorant.

[25] These conditions (some discussed above) correspond to the eight AGM postulates, though in the cases of (R4), (R5) and (R8) they are considerably weaker. See (Gärdenfors 1988) for details.

[26] Gärdenfors uses even weaker conditions than these, proving a stronger result, but this description of the theorem is sufficient for our purposes.

[27] Consider the sentences $A \wedge B$, $\neg A \wedge B$, $A \wedge \neg B$, $\neg A \wedge \neg B$. Any two of these are pairwise disjoint and by triviality at least two of the corresponding negations must be in any K.

■

Gärdenfors argues that (RT) is the culprit, and that of the premises required to prove triviality, it is most suspect and should be given up. However, the logic CO*, with the given definition of $\xrightarrow{\text{KB}}$, satisfies (RT) (by definition of $*^{\text{M}}$), and is certainly nontrivial (to see this just construct some K-revision model for an AB-ignorant K). How can this be when we have shown our revision models to satisfy the AGM postulates?

To this point we have only considered objective belief sets in the discussion of revision models. Indeed, as the triviality theorem suggests, if we permit a conditional $A \xrightarrow{\text{KB}} B$ to exist in K something has to give. In our case, it is not (RT) or nontriviality, but correspondence to the AGM postulates, in particular (R4). Consistent revision does not correspond to expansion in the presence of explicit conditionals within our model. We will show now that this is the desirable solution, that it is (R4) that should be given up, not (RT).

A simple argument against (RT), cited by Gärdenfors, is that it entails the undesirable property of monotonicity.

(M) If $K \subseteq K'$ then $K^*_A \subseteq K'^*_A$.

The implication is obvious, for if $B \in K^*_A$ then $A > B \in K^*_A$ and in K'^*_A so $B \in K'^*_A$. Gärdenfors argues that this condition should not be satisfied in general.

> Consider Victoria and her alleged father Johan. Let us assume Victoria, in her present state of belief K, believes that her own blood group is O and that Johan is her father, but she does not know anything about Johan's blood group. Let A be the proposition that Johan's blood group is AB, and C the proposition that Johan is Victoria's father. If she were to revise her beliefs by adding the proposition A, she would still believe that C, that is, $C \in K^*_A$. But, in fact, she now learns that a person with blood group AB can never have a child with blood group O. This information, which entails $C \supset \neg A$, is consistent with her present state of belief K, and thus her new belief state, call it K', is an expansion of K. If she then revises K' by adding the information that Johan's blood group is AB, she will no longer believe that Johan is her father, that is, $C \notin K'^*_A$. Thus [(M)] is violated. (Gärdenfors 1988, p.159)

While (RT) does entail (M), the antecedent of (M) is only satisfied when every conditional contained in K is in K'. If such a condition were true then we should certainly expect $K^*_A \subseteq K'^*_A$ since every conditional $A > B$ in K (which states revising by A should result in belief in B) is also in K'. There is nothing problematic with (M) in this case; typically belief sets will give rise to vastly different sets of conditionals and (M) will not apply. (M) would be discouraging if it applied to the case where the *objective* component of K is contained in K', but this is not the case.

In Gärdenfors's example, it does not appear that the trouble is actually caused by (M) or (RT), but by (R4). The original belief set contains $A > C$. When new information $C \supset \neg A$ is learned, it is simply added to K in accordance with (R4). But clearly the conditional $A > C$ should fail to hold in the revised set K'. The problem lies in (R4) insisting that $K' = K^+_{C \supset \neg A}$ and that the conditional $A > C$ should persist. This is, however, an indictment of (R4) not (M). (R4) says consistent revision ought to be identified with expansion, and in the case of objective belief sets this is warranted. However, once conditionals are added to a belief set it is not. In this example, new (objectively consistent) information $C \supset \neg A$ should cause $A > C$ to be given up (in fact, $A > \neg C$ should probably be added). $C \supset \neg A$ is consistent with the *objective component* in this case, but it is not consistent with the *entire knowledge base* K. A premise of triviality is the presence of

conditionals in K. Assuming this, clearly Gärdenfors's example does not illustrate a violation of monotonicity (M), for $K \not\subseteq K'$ due to these conditionals.

The intuitions that underlie the Ramsey test lead us to accept (RT). But these considerations also make it reasonable to accept the condition (RT')

(RT') $\neg(A > B) \in K$ iff $B \notin K^*_A$.

This enforces what Rott calls *autoepistemic omniscience*, or what has been termed *full introspection* in the context of logics of knowledge (Levesque 1986b). Given this principle, an agent has full introspective powers over its beliefs, so it can tell whether it believes A or not, as well as full introspective power with respect to revised states of beliefs. (Recall, we are discussing ideally rational agents with unbounded computational resources.) The additional constraint (RT') adds *negative introspection* to the *positive introspection* afforded by (RT), and would appear no more controvertible than (RT).

Admitting (RT) and (RT') it is easy to see that (R4) leads to the following unacceptable property of revision functions.

Proposition 6.29 *Given (R4), (RT) and (RT'), a consistent revision cannot change belief in any conditional or its negation.*

This is a simple consequence of the full introspective power of the Ramsey test, for (RT) and (RT') together ensure every K is complete with respect to its set of conditional beliefs. That is, $A > B$ or $\neg(A > B)$ is in K for every $A > B$. By (R4), any consistent revision K^*_C is identical to K^+_C, and thus any conditional $A > B \in K$ is also in K^*_C. The following example shows this to be undesirable.

Example 6.13 Suppose $KB = \{B, C\}$. Intuitively, (via $O(KB)$), $\neg(B > A)$ and $\neg(C > A)$ should be in $K = Cn(KB)$. Revising by A (presumably consistent with K) results in

$$K^*_A \supseteq \{A,\ B,\ C,\ \neg(B > A),\ \neg(C > A)\},$$

which is clearly undesirable as it violates (RT).

■

There are two ways around this problem. One is to keep (R4) and say A is inconsistent with K in this example. This seems reasonable, in the sense that, since we have conditionals in K, we ought to use whatever "logic" of $>$ is available. In this case, the presence of $\neg(B > A)$ and B in K entails $\neg A$. Though we have separate criticisms of this view (see below), the question is "Why bother?" If we adopt this perspective, there will be no interesting consistent revisions anyway, and (R4) is meaningless. (R4) has condemned itself to uselessness. The second solution is to say that (R4) simply doesn't apply to conditional knowledge bases.

Rott (1989) shows a stronger version of the triviality result that uses autoepistemic omniscience. Let $\boxtimes$ be a modal knowledge operator where $\boxtimes A$ is read as "A is known." We can define it in terms of $>$ as

$$\boxtimes A \equiv_{\mathrm{df}} \top > A.$$

Thus the correspondence to introspection becomes clear, for (RT) and (RT') ensure that

$$\boxtimes A \in K \quad \text{iff} \quad A \in K \quad \text{and}$$

$$\neg \boxtimes A \in K \quad \text{iff} \quad A \notin K.$$

Given this connection, Rott shows that no belief revision system can possess an A-ignorant belief set K.

Proof (Rott 1989) Let K be an A-ignorant belief set. Then

(1)	$\neg \boxtimes A \in K$	premise, (RT)
(2)	$A \in K_A^*$	$(R2)$
(3)	$\boxtimes A \in K_A^*$	(2), (RT)
(4)	$\boxtimes A \in K_A^+$	(3), $(R4)$
(5)	$A \supset \boxtimes A \in K$	(4), property of +
(6)	$\neg \boxtimes A \supset \neg A \in K$	(5)
(7)	$\neg A \in K$	(1), (6)

Clearly (7) stands in direct contradiction with the premise that K can be A-ignorant.

■

Again, the culprit appears to be (R4), since in step (4) of the derivation we have $\boxtimes A \in K_A^+$ merely because $\boxtimes A \in K_A^*$ and $K_A^+ = K_A^*$. This leads to step (5), the fact that $A \supset \boxtimes A$ is true in every belief set and must be a theorem of the logic. In VC this corresponds to the axiom

$$(A \supset B) \supset (A > B).$$

As a "metatheorem" or "metarule of inference" this principle is acceptable: if $A \in K$ then $\boxtimes A \in K$. But as a theorem it is inappropriate and should not be considered valid. If $\boxtimes A \in K$ then it should be that $A \supset \boxtimes A \in K$ as well, since it is a trivial propositional consequence. However, if $A \notin K$ it is not a consequence at all. It does not follow vacuously, since $\neg A$ is not necessarily in K either — we want to permit incomplete belief sets.

The sentence $A \supset \boxtimes A$ can be read as "If A is true then it is known to be true." This is precisely the kind of default rule autoepistemic logic is intended to reason with. But the point is such a sentence is a *default rule* and is intended to express one's willingness to accept $\neg A$ by default. It is not a legitimate *theorem* of autoepistemic logic (or of conditional logic) unless we restrict our belief sets to be complete. We return to the question of the validity of these sentences in the next chapter, and in the next section turn our attention to the epistemic nature of belief revision and pursue the relationship to autoepistemic logic.

Aside from these arguments that claim that (R4) is inappropriate when dealing with conditional-permitting belief sets, this conclusion is supported by the nature of our revision models. Our semantics seems quite compelling and is based on intuitions that underlie much of the earlier work on revision. In our model, the extension to conditional-permitting belief sets is quite natural, requiring no extra postulates, and automatically verifies (RT) at the expense of (R4).

6.5 A Generalization of Autoepistemic Logic

Our model for revision possesses an epistemic quality and uses the concept of what is *known* by a belief set or database to great effect. We claim any coherent notion of revision requires that a specified KB is *all that is known*, this in order to account for the effect of factual knowledge on the revision process. Indeed, we have seen that the idea of full introspection or autoepistemic omniscience can be defined quite readily within CO* in terms of the subjunctive conditional. In AI, logics of knowledge (in particular, autoepistemic logic) have been studied extensively and seem to play a crucial role in knowledge representation and default reasoning (Moore 1985; Levesque 1984a; Levesque 1990; Levesque 1986b; Lakemeyer 1991; Marek and Truszczyński 1989). Evidently, some

relationship must exist between CO* and existing logics of knowledge. Revision (or subjunctive) conditionals $A \overset{\text{KB}}{\longrightarrow} B$ have a default character in addition to their epistemic nature, as shown in various examples. For instance, from $\texttt{bird} \overset{\text{KB}}{\longrightarrow} \texttt{fly}$ one cannot infer $\texttt{bird} \wedge \texttt{penguin} \overset{\text{KB}}{\longrightarrow} \texttt{fly}$. These subjunctives seem closely related to default rules as expressed in autoepistemic logic. The "birds fly" default is written as

$$\texttt{bird} \wedge \neg \boxtimes \neg \texttt{fly} \supset \texttt{fly}$$

where $\boxtimes$ is the knowledge modality, and is read

> "If something is a bird and it is consistent to believe it flies, then it flies."

Compare this to the reading of $\texttt{bird} \overset{\text{KB}}{\longrightarrow} \texttt{fly}$

> "If a belief set is revised to accept that something is a bird, then it will include the belief that it flies."

The two are quite similar, and we can compare them by defining a knowledge operator within CO*. Recall that α is believed if $KB \supset \alpha$ is true in the revision model. This amounts to asserting that the conditional $\top \overset{\text{KB}}{\longrightarrow} \alpha$ holds, or equivalently $\Diamond\Box\alpha$.

Definition 6.15 The connective $\boxtimes$ is defined in CT4O as

$$\boxtimes\alpha \equiv_{\text{df}} \ddot{\Box}\Diamond\Box\alpha.^{28}$$

We can see the reading of $\boxtimes\alpha$ as "α is believed" is appropriate for propositional theories.

Proposition 6.30 *Let M be a preorder revision model for propositional theory K. Then $M \models \boxtimes\alpha$ iff $K \models \alpha$ for any propositional α.*

Proposition 6.35 which follows demonstrates that this definition of belief is appropriate in a more general epistemic setting, as well.

Given this definition we can show that the subjunctive conditional entails a belief in the autoepistemic version of the default rule. To see this imagine that $\texttt{bird} \overset{\text{KB}}{\longrightarrow} \texttt{fly}$ holds and that bird holds at some world w consistent with the belief set KB. By definition of $\overset{\text{KB}}{\longrightarrow}$, w satisfies $\Box(\texttt{bird} \supset \texttt{fly})$ and hence fly. That the default rule is true (nonvacuously) follows from the fact that $w \models \texttt{fly}$ implies the truth of $\neg \boxtimes \neg \texttt{fly}$, as well.

Proposition 6.31 *Let M be a preorder revision model for K such that $M \models A \overset{\text{KB}}{\longrightarrow} B$. Then $M \models \boxtimes(A \wedge \neg \boxtimes \neg B \supset B)$.*

That the converse fails to hold indicates that the subjunctive is in fact a stronger statement than the default rule. $\boxtimes(A \wedge \neg \boxtimes \neg B \supset B)$ is satisfied by any revision model where $\neg A$ is believed, but this is not true of $A \overset{\text{KB}}{\longrightarrow} B$. The subjunctive does not merely state what is true of the current belief state but also what must hold in any revised state. It is not made vacuously true by belief in $\neg A$, but will only hold if the closest A-worlds are B-worlds, a condition not enforced by the autoepistemic default statement.

---

[28] We use $\ddot{\Box}$ to emphasize the fact that $\Diamond\Box\alpha$ is true at every world in a model. In the case of CO this is unnecessary, as the shorter sequence $\ddot{\Diamond}\Box\alpha$ would suffice. In CT4O, the distinction is also unimportant if we restrict ourselves to cohesive models, but is required otherwise.

To appreciate this distinction better it befits us to compare belief revision directly to autoepistemic logic. Though several distinct versions of autoepistemic logic exist (Moore 1985; Konolige 1987; Marek, Shvarts and Truszczyński 1991), we will examine Levesque's (1990) semantic reconstruction of Moore's original formulation. Levesque provides a modal logic of belief, based on a weak S5 possible worlds semantics, in which autoepistemic expansions of a KB can be viewed as the set of beliefs that are implied by *only knowing* the KB. The logic OL contains modal operators B and O, where $\mathsf{B}\alpha$ is read as "α is believed" and $\mathsf{O}\alpha$ as "only α is believed." (Actually, O is defined in terms of B and N, where $\mathsf{N}\alpha$ is read as "At most $\neg\alpha$ is known.") The interpretations given B and O are motivated by the same considerations we discussed in Section 6.3:[29] $\mathsf{B}\alpha$ means α is true in all states of affairs that an agent considers (epistemically) possible and $\mathsf{O}\alpha$ means a world is considered possible iff it is an α-world. A model structure for OL consists of a pair $\langle \mathcal{A}, w\rangle$ where $\mathcal{A}$ is a set of possible worlds, those considered epistemically possible, and w is the "actual" world. Thus

$$\langle \mathcal{A}, w\rangle \models \mathsf{B}\alpha \quad \text{iff} \quad \mathcal{A} \subseteq \|\alpha\| \quad \text{and}$$
$$\langle \mathcal{A}, w\rangle \models \mathsf{O}\alpha \quad \text{iff} \quad \mathcal{A} = \|\alpha\|.$$

By $\|\alpha\|$ we refer to the set of worlds satisfying α among the set of all underlying logically possible worlds W, or the set of all valuations (we consider only the propositional case here, though Levesque provides a first-order treatment). Of course, the evaluation of these clauses when α contains modal operators requires appeal to the entire structure.

Given that $\mathcal{A}$ forms the set of (an agent's) accessible worlds, we can define the set of inaccessible worlds $\mathcal{I} = W - \mathcal{A}$, and view OL-structures as partitioning W as $\mathcal{A} \cup \mathcal{I}$. There is an obvious parallel to CO*, for $O(KB)$ asserts that a CO*-model is divided into a set of accessible worlds (from the point of view of KB) that satisfy KB, and a set of inaccessible worlds that do not. The difference lies in the fact that CO*-models allow a more general structure on the set $\mathcal{I}$ of inaccessible worlds, the only requirement being that they are arranged in some total preorder. This ordering ranks worlds according to closeness or plausibility. OL, on the other hand, makes no distinctions among worlds in $\mathcal{I}$, giving each equal weight, or making them "mutually accessible."[30] So we can view CO* as a generalization of OL in which less structure is imposed on inaccessible worlds; therefore such worlds can be further differentiated. See Figure 6.3.

That CO* generalizes OL implies that the type of reasoning sanctioned by OL can be duplicated in CO*. Before we demonstrate this fact, we will show that, indeed, the *structure* of OL-models can be duplicated by CO*. To constrain CO*-models to have the structure of OL-models is to ensure that all worlds fall into one of two mutually accessible clusters: the set of accessible worlds (KB-worlds) or the set of inaccessible worlds ($\neg KB$-worlds). The sentence $O(KB)$ ensures that the first set is indeed a cluster. To capture the second we must force all $\neg KB$-worlds to be mutually accessible. This is accomplished by the sentence

$$\overleftrightarrow{\Box}(\neg KB \supset \overleftrightarrow{\Box}\bot).$$

To see this imagine v is inaccessible to w where both are $\neg KB$-worlds. Then w falsifies $\overleftrightarrow{\Box}\bot$ and hence the constraint. Combining this with $O(KB)$ we get

$$\overleftrightarrow{\Box}((KB \supset (\Box KB \wedge \overleftrightarrow{\Box}\neg KB)) \wedge (\neg KB \supset \overleftrightarrow{\Box}\bot)). \tag{6.9}$$

[29] In fact, much of our account draws heavily on (Levesque 1990).

[30] Though, strictly speaking there is no accessibility relation associated with OL-structures.

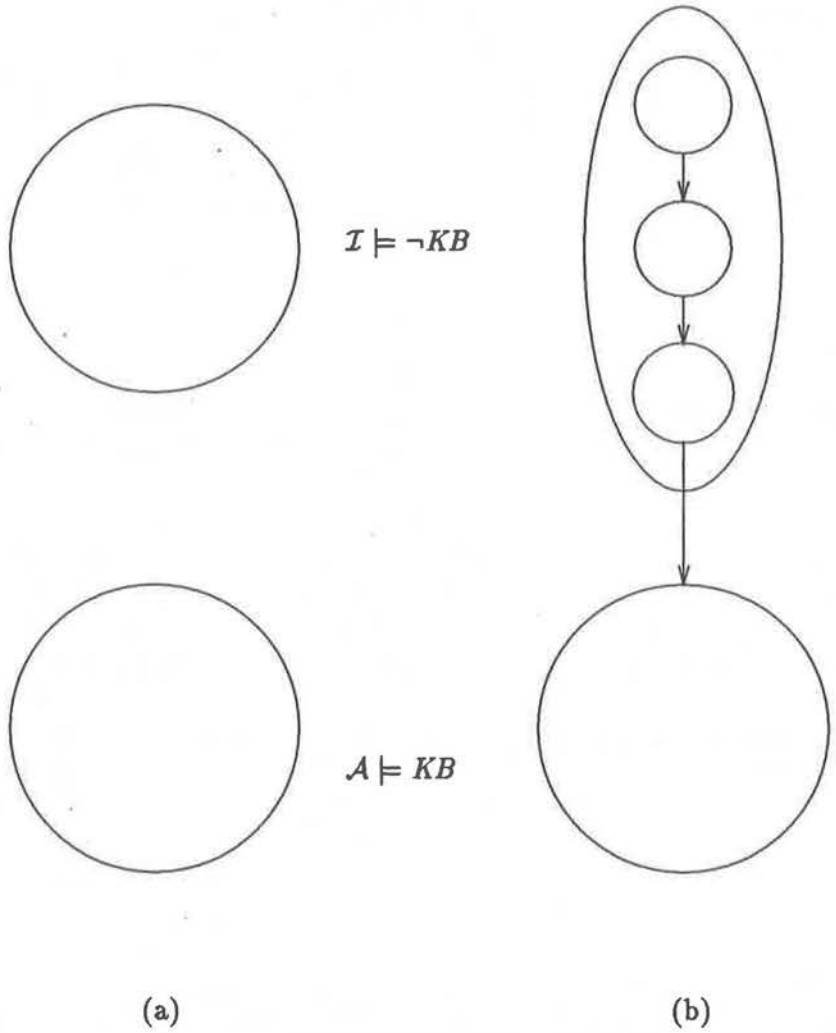

Figure 6.3: An OL-structure (a) and a CO*-structure (b) verifying O(KB) and $O(KB)$, respectively. $\mathcal{A}$ is the set of accessible worlds, and $\mathcal{I}$ the set of inaccessible worlds. CO* generalizes OL by permitting additional structure on $\mathcal{I}$ through the accessibility relation R.

We denote the sentence (6.9) by $O^+(KB)$ as it expresses an *augmented* version of only knowing $O(KB)$.

Theorem 6.32 *Let $M = \langle W, R, \varphi \rangle$ be a CO-model. Then $M \models O^+(KB)$ iff $W = \mathcal{A} \cup \mathcal{I}$ where $\mathcal{A}$ and $\mathcal{I}$ are clusters such that $\mathcal{I}$ sees $\mathcal{A}$, and $\mathcal{I}$ consists of $\neg KB$-worlds while $\mathcal{A}$ consists of KB-worlds.*

One distinction that cannot be made in CO*, but possible in OL, is the difference between only knowing a tautology and only knowing a contradiction. In CO*, $O^+(\top)$ and $O^+(\bot)$ are indistinguishable as both require models consisting of one (nonempty) cluster. Nothing differentiates $\mathcal{I}$ from $\mathcal{A}$ in this case.

Proposition 6.33 $\vdash_{CO} O^+(\top) \equiv O^+(\bot)$.

In OL this is not the case, for the semantics is defined in terms of $\mathcal{A}$ and $\mathcal{I}$ and the situation where $\mathcal{A} = \emptyset$ is different from $\mathcal{I} = \emptyset$. Thus only believing a tautology means believing nothing but tautologies, whereas only believing a contradiction entails believing every sentence. We could rule out the latter case[31] and declare by fiat that $O^+(\bot)$ really "means" $O^+(\top)$. In the results that follow however, we will primarily be interested in nontrivial belief sets (see below) and not consider belief sets that entail contradictions.

To duplicate autoepistemic reasoning in CO* we do not insist that the inaccessible worlds be indistinguishable from each other, as they are in OL. Though we can mimic OL-structures in CO*, this is not a requirement of autoepistemic reasoning. We can use the relationship illustrated in Figure 6.3 to our advantage. We wish to use the minimal set of worlds $\mathcal{A}$ as the set of epistemically possible situations. At the same time, we consider all non-minimal worlds, the set $\mathcal{I}$, to be epistemically impossible without requiring that they be mutually accessible (as they are in OL).

Definition 6.16 For any $M = \langle W, R, \varphi \rangle$ we define $\mathcal{A} \subseteq W$ to be the set of R-minimal worlds in M; that is

$$\mathcal{A} = \{w : vRw \text{ for all } v \in W\}$$

We define $\mathcal{I} = W - \mathcal{A}$.

Proposition 6.34 *For any K-revision model $M = \langle W, R, \varphi \rangle$*

(a) $M \models_w \alpha$ for each $\alpha \in K$ iff $w \in \mathcal{A}$

(b) $M \models_w \neg\alpha$ for some $\alpha \in K$ iff $w \in \mathcal{I}$

To show the correspondence to OL we must define a translation of sentences in $\mathbf{L}_{OL}$ into $\mathbf{L}_B$. While the simplest idea is to map sentences with modalities B and N to their counterparts with operators $\Box$ and $\overleftarrow{\Box}$, this can't be correct, for $\Box\alpha$ does not correspond to belief in our language. Instead, the defined connective $\boxtimes$ is a more accurate specification of belief.

Proposition 6.35 *Let M be a K-revision model. $M \models \boxtimes\alpha$ iff $M \models_w \alpha$ for each $w \in \mathcal{A}$.*

Similarly, $\mathsf{N}\alpha$, meaning α is true at all epistemically impossible worlds cannot correspond to $\overleftarrow{\Box}\alpha$ in general. We need a more "global" translation of $\mathsf{N}\alpha$, defining a connective $\overleftarrow{\boxtimes}$ that says α holds at all worlds in $\mathcal{I}$.

[31] In fact, Levesque (1981) does this by requiring $\mathcal{A}$ be nonempty.

Formulae α is true at all inaccessible worlds just when it is true at all worlds except those in the minimal cluster. In other words, it is true at all worlds inaccessible to these minimal worlds; $M \models_w \overleftarrow{\Box}\alpha$ whenever w is minimal. But if w is minimal, it can be seen by any world in M, so we want to say that $\Box\overleftarrow{\Box}\alpha$ holds at some world in M (which forces $\overleftarrow{\Box}\alpha$ to be true at each minimal world w). This points to the following definition.

Definition 6.17 The connective $\overleftarrow{\boxtimes}$ is defined in CT4O as

$$\overleftarrow{\boxtimes}\alpha \equiv_{\text{df}} \overleftrightarrow{\Diamond}\Box\overleftarrow{\Box}\alpha.$$

Proposition 6.36 *Let M be a K-revision model. $M \models \overleftarrow{\boxtimes}\alpha$ iff $M \models_w \alpha$ for each $w \in \mathcal{I}$.*

Even when using standard only knowing, $O(KB)$, instead of the augmented version, $O^+(KB)$, we can never only know a contradiction in CO*, unlike OL. Indeed, $O(\bot)$ is a theorem of CO.

Proposition 6.37 $\vdash_{CO} O(\bot)$.

This can be explained by observing that $O(KB)$ forces the set of KB-worlds to form a minimal cluster, but since the set of $\bot$-worlds is empty, $O(\bot)$ is a vacuous constraint.[32] We will only be concerned here with non-trivial belief sets.

Definition 6.18 Let $KB \subseteq \mathbf{L}_B$. KB is *trivial* iff $\vdash_{CO} O(KB) \equiv O(\bot)$.

We are now in a position to define the translation between the two languages and show that autoepistemic reasoning can be accomplished in CO*.

Definition 6.19 Let $\alpha \in \mathbf{L}_{OL}$. The *translation* of α into $\mathbf{L}_B$, denoted α^{Tr}, is defined inductively as

(1) $\alpha^{\text{Tr}} = \alpha$ for atomic propositions α
(2) $(\neg\alpha)^{\text{Tr}} = \neg(\alpha)^{\text{Tr}}$
(3) $(\alpha \supset \beta)^{\text{Tr}} = (\alpha)^{\text{Tr}} \supset (\beta)^{\text{Tr}}$
(4) $(\mathsf{B}\alpha)^{\text{Tr}} = \boxtimes(\alpha)^{\text{Tr}}$
(5) $(\mathsf{N}\alpha)^{\text{Tr}} = \overleftarrow{\boxtimes}(\alpha)^{\text{Tr}}$

Before showing the relation between OL and CO*, we will mention the relationship of OL to autoepistemic logic. Given a set of premises KB, autoepistemic logic determines the stable expansions of KB, those sets that can (roughly) be characterized as the beliefs that follow from KB together with belief sentences $\mathsf{B}\alpha$, where α is in the expansion, and $\neg\mathsf{B}\beta$, where β is not. That is, the expansion is closed under introspection. Levesque (1990) has shown that expansions of this sort are precisely those sets of sentences believed when KB is only known. Thus the primary type of autoepistemic reasoning addressed by Levesque for OL takes the form of queries: $\mathsf{O}(KB) \supset \mathsf{B}\beta$. The set of sentences satisfying this query forms the intersection of the stable expansions of KB. Intuitively, this is asking if we are justified in believing β given that KB is all the information we have about the world. Our translation of OL into CO* shows the same queries can be faithfully answered within CO*. In other words, CO* subsumes autoepistemic logic.

[32] $O(\top)$, of course, is not vacuous, and forces all worlds to be mutually accessible, so nothing is known.

Theorem 6.38 $\vdash_{OL} \mathsf{O}\alpha \supset \mathsf{B}\beta$ *iff* $\vdash_{CO*} O(\alpha^{\mathrm{Tr}}) \supset \boxtimes\beta^{\mathrm{Tr}}$.

This theorem shows that autoepistemic default rules can be specified in CO* and that only knowing in CO*, by agreeing exactly with only knowing in OL, characterizes precisely the stable expansions of an autoepistemic default theory.

Example 6.14 Suppose $KB = \{\mathtt{bird} \wedge \neg \boxtimes \neg \mathtt{fly} \supset \mathtt{fly}\}$. Then

$$\vdash_{CO*} O(KB) \supset \boxtimes(\mathtt{bird} \supset \mathtt{fly}).$$

Adding bird to KB ensures that $\boxtimes$fly is derivable from $O(KB)$. This follows from the example of Levesque (1990) showing the same derivation in OL.

■

Because CO* generalizes OL by allowing more structure on inaccessible worlds, the subjunctive representation of a default rule is in many ways more compelling than its autoepistemic counterpart. As with normative conditionals, subjunctive default rules can be used to infer new rules, and priorities on the application of rules is automatically derivable.

Example 6.15 Consider the standard "Unemployed Grad Student" example from the default reasoning literature (see Section 2.2.2). In autoepistemic logic the standard theory is written as

$$KB = \left\{ \begin{array}{l} \mathtt{adult} \wedge \neg \mathsf{B} \neg \mathtt{employed} \supset \mathtt{employed} \\ \mathtt{student} \wedge \neg \mathsf{B} \mathtt{employed} \supset \neg \mathtt{employed} \\ \mathtt{student} \wedge \neg \mathsf{B} \neg \mathtt{adult} \supset \mathtt{adult} \end{array} \right\}$$

Unfortunately, from this theory we can derive defaults stating (on the standard autoepistemic representation of rules) that student adults are both typically employed and typically unemployed, as shown by the following theorems:

$$KB \vdash_{OL} \mathtt{student} \wedge \mathtt{adult} \wedge \neg \mathsf{B} \neg \mathtt{employed} \supset \mathtt{employed} \quad \text{and}$$

$$KB \vdash_{OL} \mathtt{student} \wedge \mathtt{adult} \wedge \neg \mathsf{B} \mathtt{employed} \supset \neg \mathtt{employed}.$$

In CO* however, the natural expression of the theory doesn't give rise to this anomaly.

$$KB = \left\{ \begin{array}{l} \mathtt{adult} \xrightarrow{\mathtt{KB}} \mathtt{employed} \\ \mathtt{student} \xrightarrow{\mathtt{KB}} \neg \mathtt{employed} \\ \mathtt{student} \xrightarrow{\mathtt{KB}} \mathtt{adult} \end{array} \right\}$$

$$KB \vdash_{CO*} \mathtt{student} \wedge \mathtt{adult} \xrightarrow{\mathtt{KB}} \neg \mathtt{employed} \quad \text{but}$$

$$KB \not\vdash_{CO*} \mathtt{student} \wedge \mathtt{adult} \xrightarrow{\mathtt{KB}} \mathtt{employed}.$$

■

While it should be clear that autoepistemic statements such as

$$\mathtt{bird} \wedge \neg \boxtimes \neg \mathtt{fly} \supset \mathtt{fly}$$

cannot be interpreted as default rules *per se*, the corresponding default rule bird $\Rightarrow$ fly would seem to *justify* belief in such a statement. It should also be evident, however, that the subjunctive

$\texttt{bird} \xrightarrow{KB} \texttt{fly}$ is also not strictly a default rule. It too is a statement about an agent's belief state that is in some sense justified by the acceptance of the default $\texttt{bird} \Rightarrow \texttt{fly}$. Autoepistemic rules and subjunctives can be viewed as reason-guiding beliefs that somehow establish the principles required by default reasoning. However, subjunctives, as indicated by the previous example, are stronger statements that lead to more plausible conclusions about an agent's belief state. The exact relationship between normative defaults and subjunctives is addressed in the next chapter (but we will raise as many questions as we answer).

One property of autoepistemic rules that might make them more desirable in certain circumstances is that they can be "only known." For instance, in Example 6.14, $O(KB)$ is consistent in the case of the initial KB, and amounts to believing only that $\texttt{bird} \supset \texttt{fly}$ is true. Subjunctives (taken separately), in contrast, cannot be only known, for they are globally true sentences; if $A \xrightarrow{KB} B$ holds at some possible world, it holds at all possible worlds. Thus $O(\texttt{bird} \xrightarrow{KB} \texttt{fly})$ is inconsistent as no model can have belief in only $\texttt{bird} \xrightarrow{KB} \texttt{fly}$-worlds, since some of these must be $\texttt{bird} \wedge \neg\texttt{fly}$-worlds, contradicting the subjunctive.[33] On the other hand, a subjunctive is not made vacuously true when its antecedent is believed false, unlike an autoepistemic rule. If $\neg\texttt{bird}$ is believed, so are both autoepistemic default rules "birds fly" and "birds do not fly." A subjunctive $A \xrightarrow{KB} B$ is not affected by the falsity of A. We have yet to explore the implications of these properties in full detail, but in the next chapter we do examine the relationship between subjunctive and default reasoning more generally, and clearly the particular connections to autoepistemic logic remain a promising avenue of research.

Another aspect of the autoepistemic nature of revision we will discuss only superficially is the further generalization of autoepistemic logic afforded by CT4O*. The notion of only knowing is also meaningful for this weaker logic. Just as using $O(KB)$ in CO* allows us to construe CO* as generalizing OL, so too could using $O(KB)$ in CT4O* allow further generalization. Interestingly, however, the augmented versions $O^+(KB)$ coincide in the two logics, for two clusters cannot form a preorder and not be totally ordered (given cohesiveness).

Theorem 6.39 *Let M be a CT4O*-model such that $M \models_{CT4O*} O^+(KB)$. Then M is a CO*-model such that $M \models_{CO*} O^+(KB)$.*

Corollary 6.40 $\vdash_{CT4O*} O^+(KB) \supset \boxtimes\alpha$ *iff* $\vdash_{CO*} O^+(KB) \supset \boxtimes\alpha$.

Even more intriguing is the fact that if we restrict KB to use only the modal operators $\boxtimes$ and $\overleftarrow{\boxtimes}$, excluding formulae containing (explicit) occurrences of $\Box$ and $\overleftarrow{\Box}$, then $O(KB)$ will give rise to the same set of beliefs in both CO* and CT4O*. Thus any modal logics that lie between S4 and S4.3 give rise to exactly the same autoepistemic logics as specified by our conception of only knowing O^+ (though this is certainly not the case for "regular" only knowing O). This suggests a semantic counterpart of the *range* results of Marek, Shvarts and Truszczyński (1991) in which various nonmonotonic logics, as defined with respect to an underlying modal logic by the fixed-point operator of McDermott and Doyle (1980), are shown to be identical.[34]

[33] This is not to say we cannot only know a KB that contains subjunctives, just not those containing nothing but subjunctives.

[34] In fact, by our argument, stronger logics than S4.3 (e.g., S4.3.2) would also collapse into S4 under O^+. Thanks to Mirek Truszczyński for bringing this logic S4.3.2 to my attention. This logic appears to be important for relating default logic and nonmonotonic logic. See Section 7.2.2.

6.6 Miscellany

In this chapter, we have shown how bimodal logics such as CO* can be used to represent the process of belief revision and associated concepts. We defined a conditional $\xrightarrow{\text{KB}}$, the sentence $A \xrightarrow{\text{KB}} B$ being read "If KB is revised by A then B will be believed." Theorems 6.7 and 6.8 show that this is a coherent notion, equivalent to the AGM postulates. In fact, CO* can be viewed as the first logic for AGM revision. The expressive power of CO* enabled this characterization, for in CO* we can express the concept of *only knowing* a knowledge base. This is indeed the key requirement of any logic for revision. Theorem 6.1 and Corollary 6.3 show how the concept of only knowing can be defined in our bimodal language. We also showed how CO generalizes this process by permitting "relative consistency." CT4O generalizes CO further by determining a notion of *preorder revision*. One question that has not been answered is what set of postulates corresponds to this type of revision.

Appeal to the Ramsey test shows $\xrightarrow{\text{KB}}$ to be nothing more than a subjunctive conditional. We provided a framework for answering subjunctive queries that improves on existing logics precisely because epistemic concepts like belief and only knowing are representable. Furthermore, we respected the Ramsey test without falling prey to *triviality*, and argued that conditional belief sets should only have to satisfy (R4) vacuously, since there can be no consistent revisions of such belief sets.

We showed how a number of constraints on the revision process can be expressed in CO*, including integrity constraints of various types (Theorems 6.13, 6.16 and 6.19), plausibility of sentences (Theorems 6.23 and 6.24) and entrenchment of sentences (Corollaries 6.25 and 6.26). Finally, we demonstrated that CO* is a significant generalization of autoepistemic logic (Theorem 6.38). While autoepistemic constraints on a belief set can be expressed in CO* and interact with the revision process, they alone are not sufficient as "reason-guiding" beliefs for default reasoning; subjunctive constraints are also required. This suggests that fundamental connections exist between normative statements and their subjunctive (and autoepistemic) counterparts. While much work remains to be done in this respect, we begin addressing such relationships in the next chapter.

Chapter 7

Unification

To this point we have discussed modally-defined conditional logics extending S4 in great detail. In particular, we have concentrated on the modal logic CO*, in which concepts relating to both default reasoning and belief revision are expressible. Indeed, such diverse systems and methods as ε-entailment, rational and preferential consequence, rational closure, AGM revision and autoepistemic logic can be expressed, and extended, within CO*. We have seen the relationship that exists among the various approaches to default reasoning, as well as the connections between various models of revision (such as AGM revision, Grove's system and Lewis's counterfactual logic).

One feature of the revision conditional $\xrightarrow{\text{KB}}$ mentioned only in passing in Chapter 6 is the identity of its definition in CO* with that of the normative conditional $\Rightarrow$ given in Chapter 5. In this chapter we will suggest that default reasoning and belief revision are governed by the same logics and, in fact, that normative and subjunctive conditionals are the same conditional. The difference between normative and subjunctive reasoning is the context in which the logic is used.

We will start by reviewing some of the connections established in Chapters 4, 5 and 6, concentrating on the similarity of the processes of default reasoning and belief revision. We will then explain how the two distinguished forms of reasoning can have the same formal structure and be modeled within the same logics, requiring only a slight shift in perspective to apply the logics correctly. Straightforward revision and subjunctive reasoning is seen as revision of the *actual state of affairs*. We try to accommodate new information with our beliefs about the world. Default or normative reasoning, in contrast, is viewed as revision of some *idealized normal state of affairs*. New information (as well as previous knowledge of the world) is reconciled not with what we know, but with what we expect, or what we take normally to be the case. In this way we can account for extra knowledge of which we cannot be certain, but perhaps only reasonably safe in assuming. The fact that the same conditional connectives and logic can accurately represent subjunctives and normatives lends credence to the claim that belief revision and default reasoning are essentially the same "logical reasoning task," albeit employed for different purposes.

This notion has been suggested with varying degrees of commitment and explicitness by several people. Poole's (1988) Theorist framework implicitly adopts the revision approach to default reasoning, while in (Kraus, Lehmann and Magidor 1990) it is suggested that the difference between the interpretation of the semantics of preferential consequence and that of standard conditional logics is the reference of the latter to what would be true if the *actual* world were some way. Katsuno and Satoh (1991) have proposed a class of models called *ordered structures* that capture some of the intuitions underlying default reasoning, belief revision and conditional logics. The connection between revision and nonmonotonic reasoning has been established in its strongest form to date by Gärdenfors and Makinson (1990; 1991). We will discuss their results briefly in this chapter as well.

However, using the logic CO* (and in an analogous manner its weaker generalizations extending CT40), we will provide the first demonstration that the *logics* of revision and default reasoning reasoning are identical.

We will continue by showing that other approaches to default reasoning, including possibility theory (Dubois and Prade 1988), can also be accommodated within the CO* framework, and suggesting that CO* might provide a qualitative counterpart of even standard probabilistic (Bayesian) notions of inference. In this regard, the modal logic CO* and its weaker counterparts determine a uniform framework in which many diverse approaches to nonmonotonic inference may be developed, investigated, compared and understood, including approaches typically based on extra-logical characterizations (for instance, AGM revision, or Theorist).

7.1 On the Relation Between Subjunctive and Normative Conditionals

In Chapters 4 and 5 we developed a definition for a normative conditional based on the modal logics CT4 and CT4O. In this section, however, we will concentrate on the extension CO*, but note that the remarks that follow regarding the relationship between the subjunctive and normative conditional hold in the weaker case, given suitable adjustments. In CO* a simple modal formulation of the truth conditions for $\Rightarrow$ is

$$A \Rightarrow B \equiv_{\text{df}} \overset{\leftrightarrows}{\Box}\neg A \vee \overset{\leftrightarrows}{\Diamond}(A \wedge \Box(A \supset B)).$$

CO* places a total preorder on states of affairs, ordering worlds according to normality or typicality. In the principal case $A \Rightarrow B$ holds just when B is true at the most normal A-worlds (at least in the limit). This corresponds precisely to preferential and rational consequence and ε-entailment in the case of simple conditional theories, but extends these in a natural way, as we have seen. For instance, the language of CO* is far more expressive than that of simple, or even extended, conditionals; with the modality for inaccessibility we can axiomatize even rational closure or 1-entailment. As well, CO* makes no commitment to the Limit Assumption, providing meaningful truth conditions for $A \Rightarrow B$ even when no minimal A-worlds exist.

In Chapter 6 we developed a definition for a subjunctive conditional based on CO* as well, and the truth conditions for $\xrightarrow{\text{KB}}$ can be given as

$$A \xrightarrow{\text{KB}} B \equiv_{\text{df}} \overset{\leftrightarrows}{\Box}\neg A \vee \overset{\leftrightarrows}{\Diamond}(A \wedge \Box(A \supset B)).$$

Once again CO* places an ordering on states of affairs, but in this case worlds are ranked according to their closeness or similarity to the actual world. We were able to show that CO* stands in perfect relation to the AGM postulates for revision, yet again has the advantage of expressing revision within the logical language, and does not make the Limit Assumption.

The interesting feature of these definitions is the fact that they are identical. In the same modal logic we can define a subjunctive conditional and a normative conditional equivalently and show they behave properly (at least to the extent that the connections to existing systems demonstrate this). On this view, then, subjunctives and normatives have precisely the same (formal) truth conditions. However, subjunctive and normative reasoning are clearly two different processes. How can the two conditionals have the same truth conditions? How can the same logic be appropriate for both?

Of course, having the same *formal truth conditions* in CO* does not mean the two conditionals

have the same *actual truth conditions.* The difference is easily explained by appealing to the interpretation provided accessibility for subjunctives on the one hand, and normatives on the other. When evaluating the truth of the subjunctive $A \xrightarrow{\text{KB}} B$ we order possible worlds according to the degree to which they conform to our knowledge of the actual world. The subjunctive is true just when B is true at the *closest* A-worlds. A normative $A \Rightarrow B$ is evaluated by ranking worlds respecting the degree to which they are judged to be normal or unexceptional. The normative is true just when B is true at the *most normal* A-worlds. While these conditions are formally identical, this implies no equivalence of the "actual" (say, linguistic) truth conditions. For instance, there is no problem permitting `bird(T)` $\Rightarrow$ `fly(T)` to be true and `bird(T)` $\xrightarrow{\text{KB}}$ `fly(T)` to be false within CO* (say when `T` is some "actual" non-flying bird). In each case, the interpretation of the accessibility relation is different.

But there is also a reasonably strong argument against normatives and subjunctives having even the same formal truth conditions. That they have the same formal structure implies that the inferential relation existing between sets of normatives is the same as that for subjunctive conditionals. One rule of inference purported to distinguish these two connectives formally is modus ponens for the conditional (writing $>$ for the subjunctive):

CMP $A, A \Rightarrow B \vdash B$,

CMP $A, A > B \vdash B$.

Obviously **CMP** should not be valid for the normative conditional (in the first instance above), for if this detachment principle were valid none of our default rules could have exceptions. We must be able to assert consistently that birds normally fly while a particular bird Tweety does not, or a certain class of birds such as penguins does not. This point seems relatively incontrovertible.

In contrast, it is usually claimed that **CMP** must be valid for the subjunctive conditional. Imagine the subjunctive "If it were raining I would be carrying my umbrella" is true (say, $R > U$). This states that at the closest worlds to the actual at which R holds, U holds as well. Imagine now that R is true. Clearly the actual world is (among) the closest at which R holds. Therefore, for $R > U$ to hold U must be true. Indeed, **CMP** is valid in most subjunctive logics, such as Lewis's (1973a) VC and Stalnaker's (1968) C2.

Somewhat at odds with this assessment is the fact that **CMP** is not valid for our subjunctive (since it is not for $\Rightarrow$):

$$A, A \xrightarrow{\text{KB}} B \not\models_{CO*} B.$$

We claim here that, in fact, **CMP** is not appropriate for subjunctives. To be a valid rule of inference, the conclusion of the rule must be true in all situations where the premises hold. This includes both situations considered reasonable (by an agent) and those given little or no plausibility. This is not the case for **CMP** in CO* because of the epistemic nature of the subjunctive conditional. We have argued for an explicit reading of the Ramsey test, whereby a subjunctive is believed just when its consequent is present in the state of belief (hypothetically) achieved by obliging the antecedent. As we have noted, $R \xrightarrow{\text{KB}} U$ is true in some world just when it is true at all worlds in a CO*-model. This is due to the fact that a CO*-model is suitable for a fixed set of beliefs only. But this means that the truth of $R \xrightarrow{\text{KB}} U$ is independent of the world at which it is being evaluated, and specifically independent of the truth of R and U at any given world. If an agent's state of belief is such that $R \xrightarrow{\text{KB}} U$ is held, it does not matter what truth values are given R and U in the "real" world. It might be the case $R \wedge \neg U$ is true — all that tells us is that the agent's beliefs are mistaken (in particular, its belief in $R \supset U$). Indeed, when an agent assents to $R \xrightarrow{\text{KB}} U$, any $R \wedge \neg U$-world is a situation where **CMP** is violated. Even in worlds where I have forgotten my umbrella on a rainy

day $R \xrightarrow{\text{KB}} U$ holds *given this particular state of belief.* Imagine being shut up in a windowless office all day (or being really stupid).

Given the link between the truth conditions for $\xrightarrow{\text{KB}}$ and the Ramsey test, this suggests that whenever both $R \xrightarrow{\text{KB}} U$ and R are *believed*, U is *believed* as well. This is quite different from **CMP**, which refers to the *truth* of R and U. Indeed this "metarule" of inference is valid for CO*:

KMP If A and $A > B$ are believed, then B is believed.

KMP If $A \in KB$ and $A \xrightarrow{\text{KB}} B \in KB$ then $B \in KB$.

Using the belief notation of Section 6.5 we can write this as a derived theorem of CO*:

KMP $\vdash_{CO*} \boxtimes A \wedge \boxtimes(A \xrightarrow{\text{KB}} B) \supset \boxtimes B$.

Thus CO* and $\xrightarrow{\text{KB}}$ conform more exactly to the intuitions underlying the Ramsey test than does the logic VC. The failure of **CMP** in CO* is related to the failure of the following inference, which is valid in VC (see also Section 6.4):

$$A \wedge B \vdash_{VC} A > B$$

$$A \wedge B \not\vdash_{CO*} A \xrightarrow{\text{KB}} B.$$

However, in CO* we have the following derived theorem regarding such beliefs:

$$\vdash_{CO*} \boxtimes(A \wedge B) \supset (A \xrightarrow{\text{KB}} B).$$

This is not to say that subjunctives cannot be based on or related to the truth of facts in the actual world rather than epistemic states. There must certainly be subjunctives other than those (like $\xrightarrow{\text{KB}}$) based on the Ramsey test, conditionals that validate, say, **CMP**. For example, update semantics proposed for reasoning about revision due to changes in the world instead of changes in belief (Winslett 1988; Katsuno and Mendelzon 1991; Grahne and Mendelzon 1991) might require these rules of inference, and might validate different portions of VC. The logic of Grahne (1991) contains a connective of this sort.

Assuming now that the formal structure of normatives and subjunctives can coincide, the question remains: how do normative and subjunctive reasoning differ? We claim that the distinction lies not in the logic but in the way the logic is applied to reasoning tasks. A certain perspective is adopted when the logic is used in one instance, and this view is modified to accommodate the other the other type of reasoning. This change in view takes place essentially in the interpretation of the ordering placed on states of affairs.

Consider the evaluation of a normative conditional. To effect this process we must, to some degree, impose an ordering on worlds reflecting the extent to which they are adjudged to be uniform or unexceptional. The actual world might not be especially homogeneous, but we can imagine what it would be like if it were. In a certain respect, we are aspiring to some idealized norm, some state(s) of affairs that are deemed most normal, at which (for instance) all "defaults" are true and no exceptions exist. It is just these worlds that are minimal in our ordering, and situations that conform more exactly to these uniform states are considered more normal than those that correspond less exactly. Most certainly our knowledge won't allow us to achieve this "epistemic nirvana," but to the extent our beliefs are consistent with normality we take advantage. In this way we are able to exploit our expectations of the how the world normally (or typically, or probably) is, given our knowledge, and make more adventurous predictions. This is precisely the province of default reasoning.

Consider now the evaluation of a subjunctive conditional. Once again we rank states of affairs, but in this case we forego the ideal used in normative reasoning and allow cold, hard reality to color our perceptions. We order worlds according to their conformation to the actual state of affairs, rather than some normal state, or how they relate to the way things *are* rather than the way things *should be*. We take the actual world to be the most normal or most probable and consider how the world might change given new information.

It seems clear that the two orderings should be related. Suppose I believe birds normally fly, $\texttt{bird(x)} \Rightarrow \texttt{fly(x)}$, but I think that Aunt Martha's pet Sylvester is a cat. If I consider what would be the case were Sylvester a bird, most likely I will assent to the subjunctive $\texttt{bird(Sylvester)} \xrightarrow{\text{KB}} \texttt{fly(Sylvester)}$. What is normally the case, or what is expected, will influence how we revise our beliefs. Worlds judged to be more likely will also often be considered closer to the actual world. Conversely, the subjunctives deemed true and the revisions we typically make should affect the normative statements we admit. This might be associated with the question of how we come to accept default rules or normative conditionals, perhaps based on empirical experience or an abstraction of probabilistic data. If revising by A results in belief B frequently enough, it might be the case $A \Rightarrow B$ comes to be acknowledged. Worlds closer to the actual world will often be considered more normal than those less similar. This connection remains unexplored, but further investigation might lead to an even greater blurring of the distinction between subjunctive and normative reasoning. Indeed, the task of defining a logic in which the subjunctive and normative conditionals can be reasoned with simultaneously and interact should prove to be both feasible and worthwhile.

Another view of the relationship between default reasoning and belief revision has been put forth by Gärdenfors and Makinson (1990; 1991) and it too is compatible with our perspective and the identity of the logics governing the two processes. They propose interpreting nonmonotonic inference in terms of revision of one's expectations. Suppose we have some set of defaults KB we are willing to adopt whenever possible. For simplicity we assume $K = Cn(KB)$. When we wish to determine the nonmonotonic consequences of a sentence A the idea is to revise our expectations K to include A. The resulting theory is the belief set we are willing to accept given a belief in A. Thus $A \mathrel{|\!\sim} B$ iff $B \in K^*_A$. The nature of revision is such that we give up as few defaults as possible from K to accommodate A. This is just the sort of viewpoint adopted for default reasoning by Poole (1988) and followed by Brewka (1991).

The Theorist framework of Poole is based on a background set of defaults D, consisting of propositional sentences we are prepared to assume as true (that is, our expectations).[1] An *extension* for a set of facts F is a set $D' \cup F$ where D' is a maximal subset of D consistent with F. Roughly speaking, an extension is a possible default extension of F, a theory containing F and satisfying as many defaults as possible. One may define a (skeptical) nonmonotonic inference relation in terms of Theorist by saying $A \mathrel{|\!\sim} B$ just when B is in all extensions of A (with respect to some underlying set of defaults D).

While not explicitly stated in these terms, this is more or less the idea behind belief revision: one keeps as many original beliefs (defaults) as possible. Gärdenfors and Makinson (1990) show that (with suitable modifications) nonmonotonic reasoning based on Theorist can be interpreted as belief revision based on a belief set of expectations. They also show how various postulates for revision parallel properties of general nonmonotonic inference relations. In (Gärdenfors and Makinson 1991), they push this equivalence even further by demonstrating that *expectation orderings* (essentially orderings of entrenchment in a default setting) correspond to a class of preferential semantics in

[1] We ignore here the *constraints* of (Poole 1988), which are not relevant to the current exposition.

the sense of Shoham (1988).

On their view, somewhat like Poole, default reasoning can be viewed as revision of a set of expectations or defaults K. The default rule, or nonmonotonic inference, $A \mid\!\sim B$ is true just when $B \in K_A^*$. In this way default reasoning and belief revision can be unified and given a consistent formal treatment. This provides another interpretation of the identification of the normative with the subjunctive conditional. While subjunctive reasoning is based on revision of one's *knowledge* of the actual world, normative reasoning is based on revision of one's *expectations*. In a CO*-model of subjunctives, epistemically possible worlds are minimal in the closeness ordering, and hence it is the theory determined by these worlds (one's knowledge) that is revised. In a CO*-model of normatives, the minimal worlds (assuming some limit) are those that satisfy each conditional default or unconditional expectation. Thus it is this theory of expectations that is revised.

7.1.1 Irrelevance and Belief Revision

While the results of Gärdenfors and Makinson show a close relationship between revision and default reasoning, these are stated at the extra-logical level of postulates. One must translate statements of revision of the form $B \in K_A^*$ into statements of nonmonotonic inference $K \vdash A \mid\!\sim B$ and vice versa, showing that conditions on the revision function $*$ correspond to properties of the inference relation $\mid\!\sim$. Our results are stronger in the sense that statements of revision $A \xrightarrow{KB} B$ *are* statements of nonmonotonic inference $A \Rightarrow B$. The logics of belief revision and default reasoning are identical. Properties of the conditional $\Rightarrow$ are automatically valid for $\xrightarrow{KB}$ in any particular modal logic or modal theory. There is also the advantage of being able to express these properties directly at the object level in terms of derived theorems and inference rules rather than at the metalevel of postulates and conditions. This object level perspective has the concomitant benefit of allowing one to express interaction of, say, propositional knowledge with these postulates, or unusual interaction of the postulates, directly within the logical language (see Section 4.1.4).

In Chapter 5 we saw that reasoning by default using conditionals is somewhat problematic when irrelevant information is present in a knowledge base (which will almost always be the case). Since the notion of consequence underlying both the normative and subjunctive interpretations of our conditionals is identical this would seem to indicate that the problem of irrelevance should affect subjunctive reasoning as well. Indeed, this is the case.

Example 7.1 Consider the set of subjunctive premises

$$\{\texttt{bird} \xrightarrow{KB} \texttt{fly},\ \texttt{penguin} \xrightarrow{KB} \texttt{bird},\ \texttt{penguin} \xrightarrow{KB} \neg\texttt{fly}$$

evaluated with respect to knowledge base KB containing $\neg$bird. Clearly revising by bird results in a new KB containing fly, while revising by bird $\wedge$ penguin results in the belief $\neg$fly. However, if we were to ask if revising by bird $\wedge$ green results in the belief fly, the answer would be UNKOWN. Neither

$$\texttt{bird} \wedge \texttt{green} \xrightarrow{KB} \texttt{fly}$$

nor

$$\texttt{bird} \wedge \texttt{green} \xrightarrow{KB} \neg\texttt{fly}$$

is derivable from these premises.

■

Intuitively, as in the case of default reasoning, we expect revision by $\texttt{bird} \wedge \texttt{green}$ to result in the belief $\texttt{fly}$, since nothing in the knowledge base indicates that green birds are somehow different from birds generally when beliefs are revised. However, just as in the case of default reasoning, the logical structure of the connective $\overset{\text{KB}}{\longrightarrow}$ does not allow this inference.

The equivalence of CO* to the AGM postulates shows that this problem exists for any method of revision that respects the AGM postulates without extending their power. The problem of irrelevance has not been discussed in the literature on either subjunctive reasoning or belief revision, yet evidently it is a serious problem that needs to be addressed before systems that perform this type of reasoning can be developed. The equivalence of $\Rightarrow$ and $\overset{\text{KB}}{\longrightarrow}$ indicates that existing formalisms for belief revision are not equipped to deal with this problem, that has yet to be identified within the revision community. We note that the solutions to irrelevance we presented in Chapter 5 can be applied to reasoning with our revision conditional. However, as we mentioned at the end of that chapter, these solutions are not entirely satisfying and we do not pursue this connection here.

The extra-logical approach to nonmonotonic reasoning based on revision can be viewed as more general, since revision can be defined in terms of arbitrary *selection functions* that choose subtheories of the set of defaults being revised. Gärdenfors and Makinson (1991) provide just such a definition of *expectation inference operations*. If we think of these functions as selecting some set of worlds that is preferred given a specific antecedent then this has the effect of enforcing the Limit Assumption. But it is not the case in general that a selection function on sets of sentences need correspond to a selection function on worlds in some ordering (although Gärdenfors and Makinson seem to suggest this is what they have in mind). Thus, it may be the case that expectation inference makes no commitment to the Limit Assumption. In defense of CO* and its generalizations, we remark that the notion of ordering is such a pervasive and intuitive concept that a selection semantics that cannot be interpreted in such terms might not be interesting except from a mathematical or logical perspective. As a normative theory of reasoning, orderings seem particularly compelling.

7.2 Other Connections

7.2.1 Possibility Theory

Because we are able to unify default reasoning and belief revision within CO*, many concepts associated with revision may be applied to general nonmonotonic inference. In Chapter 6 we discussed several concepts frequently associated with revision and expressed these in our bimodal language, concepts that typically have no analogues in default reasoning. Among these are the notions of plausibility, entrenchment and integrity constraints. Given this unification these concepts are obviously applicable in a default setting and can be given a meaningful interpretation.

Plausibility is an ordering relation on sentences where A is at least as plausible as B just when $\Box(B \supset \Diamond A)$ holds. In the context of default reasoning A is more plausible just when it is more expected than B. In other words, as we are forced to "give up hope" in more and more normal (or expected) situations as our knowledge conflicts with these, we will be more willing to accept some situation where A holds than any where B holds. As we move up the ordering to less expected situations we will find a world where A holds before one satisfying B. This corresponds to the notion of degree of Lehmann (1989), since this occurs just when $A \vee B \Rightarrow A$ is true. It also equivalent to Pearl's (1990) Z-ranking of formulae, since this implies $Z(A) \leq Z(B)$.

Gärdenfors and Makinson also provide a definition of a plausibility valuation whereby sentences are mapped into a totally ordered set determining some measure of their plausibility. Intuitively,

one may think of assigning a value from the real interval $[0, 1]$ to each sentence, with higher ranked sentences being more plausible. We do not go into details here, but given two basic conditions on the plausibility mapping and an appropriate notion of inference, Gärdenfors and Makinson show that inference based on plausibility is precisely inference based on expectation. We remark in passing that this notion of plausibility corresponds to our definition of plausibility precisely.[2]

Gärdenfors and Makinson's concept of plausibility ordering is reminiscent of the *possibility theory* of (Dubois and Prade 1988). There *events* are mapped into the real interval $[0, 1]$ such that the following conditions hold:

(a) $Poss(\top) = 1$

(b) $Poss(\bot) = 0$

(c) $Poss(A \cup B) = \max(Poss(A), Poss(B))$

As observed by Gärdenfors and Makinson (1991) and Fariñas and Herzig (1991), the fact that the real unit interval is the range of the possibility function is inessential, the important quality being the total ordering of the set. Indeed the numbers have no meaning or function other than to relate the possibility measure of one sentence to another. Reading $Poss(A) = .9$ as "A is quite possible" is inappropriate, for a possibility measure assigning degree .1 to A can determine exactly the same ordering of sentences and, hence, the same inference relation.

This has lead Fariñas and Herzig (1991) to develop a modal theory of *qualitative possibility*. This logic is shown to capture the essential nature of qualitative possibility theory. Unfortunately, their logic requires a large family of modal operators, one for each element, or possibility measure, in the totally ordered set. However, this multimodal logic captures precisely the intuitions underlying entrenchment (and hence plausibility) orderings. Of course, since CO* can express notions of plausibility and entrenchment, this implies that the modal logic of possibility theory requires only two modalities, $\Box$ and $\overline{\Box}$.

Fariñas and Herzig present QPL, *qualitative possibility logic*, in terms of an ordering connective $\geq$ based on the representation theorem of Dubois and Prade (1988). That theorem states that possibility measures on events are precisely those measures that agree with certain postulates about an ordering on events. QPL, in effect, is an axiom system for these postulates. We assume a standard propositional language together with the binary connective $\geq$. $A \geq B$ is read "A has a degree of possibility as great as that of B."

Definition 7.1 (Fariñas and Herzig 1991) QPL is the smallest set of sentences containing CPL and the following axioms, and closed under the following rules of inference:

QP1 $A \geq B \wedge B \geq C \supset A \geq C$

QP2 $A \geq B \vee B \geq A$

QP3 $\neg(\bot \geq \top)$

QP4 $\top \geq A$

QP5 $A \geq B \supset (A \vee C \geq B \vee C)$

QP6 From $A \supset B$ and A infer B

QP7 From $A \equiv B$ infer $A \geq B$

[2]This can be seen by considering the clusters of a CO*-model to be the elements of the totally ordered set, ordered by accessibility.

They then present PL, *possibility logic*, as a multimodal logic in which QPL can be axiomatized without the conditional connective $\geq$. We do not give details here but to notice that the logic requires a separate modal connective for each possibility measure in the ordered set, and an indexed family of accessibility relations for each operator. It is not hard to see that the large family of modal operators is not required to give a modal presentation of QPL or possibility theory. The observation of Fariñas and Herzig that QPL is an alternative axiomatization of Lewis's conditional logic VN shows that QPL has a spheres semantics. But CO* corresponds quite naturally to systems of spheres, or total preorders on possible worlds, and axiomatizes such systems directly with two modal operators. There is no need to use more than one accessibility relation or a large class of modalities to represent QPL. The ordering on sentences $\geq$ corresponds to the notion of plausibility in CO*.

Theorem 7.1 *Let Poss be a possibility measure. Then there exists a CO*-model M such that* $\leq_{PM}$ *is the plausibility ordering determined by M and* $A \leq_{PM} B$ *iff* $Poss(A) \geq Poss(B)$.

Theorem 7.2 *Let* $\leq_{PM}$ *be the plausibility ordering determined by some CO*-model M. Then there exists a possibility measure such that* $Poss(A) \geq Poss(B)$ *iff* $A \leq_{PM} B$.

Thus our notion of plausibility for revision corresponds also to a possibility measure in a default setting.

Of course, entrenchment is just the dual of plausibility and can be equally well expressed in CO* in a default setting. In this sense, the statement that A is as entrenched as B, that is $\Box(\neg A \supset \Diamond\neg B)$, corresponds to Gärdenfors and Makinson's notion of expectation ordering.

The concept of integrity constraint can be also be given an interpretation in terms of default reasoning. When an integrity constraint C holds it means that all worlds satisfying C are more plausible, or more expected, than any world satisfying $\neg C$. Thus we can think of C as being as having the highest possible degree of expectation. No matter what information we have, we predict C by default, unless we *know* $\neg C$. Nothing short of learning $\neg C$ directly (or through certain deduction) will cause C to become unexpected.

The applicability of this notion was seen in Chapter 5 where System Z was interpreted (implicitly) in terms of integrity constraints. Any world of rank 0 is considered more normal than any world of higher rank. In this manner, the conjunction of (the material counterparts of) the default rules forms an integrity constraint. All worlds satisfying this conjunction are preferred to any world falsifying some rule. These, of course, also can be viewed as imposing further, prioritized constraints; for once this strongest constraint is violated, we try to satisfy the constraint that only lower ranked rules are falsified, preferring those worlds to others.

7.2.2 Autoepistemic Logic

The epistemic nature of revision and subjunctive reasoning lead us to posit CO* as a generalization of Levesque's (1990) logic OL of only knowing, and therefore of autoepistemic logic. Thus, autoepistemic logic can be seen as a particular logic of belief revision. But the connection between subjunctives and normatives allows one to view revision-style default rules in CO* as also making statements of normality. Thus we come full circle and interpret autoepistemic default rules as also making default or prototypical statements as was perhaps the original intent of this system when proposed in (McDermott and Doyle 1980).

However, while the relationship between autoepistemic default rules and their subjunctive or normative counterparts exists, it indicates a weakness of the autoepistemic account of defaults. For

example, from $A \xrightarrow{\text{KB}} B$ one can infer

$$\boxtimes(A \wedge \neg \boxtimes \neg B \supset B)$$

as well as the weaker $\boxtimes A \supset \boxtimes B$ (recall $\boxtimes$ is the modality of belief in CO*). The converse implication is not typically valid, for this autoepistemic rule is made true whenever $\boxtimes \neg A$ holds. The revision statement $A \xrightarrow{\text{KB}} B$ makes a much stronger assertion and seems to be a more reasonable notion of default; it says that if one *comes to believe* A then one will also believe B, rather than the weaker statement that if one *does believe* A then one will believe B.

Because of the connection with autoepistemic logic, and its ability to express even these simplistic autoepistemic default rules, CO* can be related to many other more traditional forms of default reasoning. The relationship between default logic and autoepistemic logic has drawn considerable attention over the past few years (Konolige 1987; Marek and Truszczyński 1989; Truszczyński 1991). In (Marek, Shvarts and Truszczyński 1991) a number of nonmonotonic modal logics have been proposed and investigated, and recently Truszczyński (1991) has shown how default logic can be embedded in one such logic, the nonmonotonic version of the modal system S4F. The logic S4F is an extension of S4.3 and is characterized by the class of structures having exactly two related clusters (though one may be empty). We recall that such structures correspond precisely to those CO-models that satisfy the *augmented* version of only knowing $O^+(KB)$. We concluded that augmented only knowing is too strong a restriction in general for default reasoning, and, not surprisingly, Truszczyński's result suggests to us that default logic is not suitable in many circumstances, particularly with respect to deriving priorities and new default rules. Embedding default logic in a modal logic, however, allows natural generalizations of default rules (for instance, the disjunction of rules) and hence CO* can perhaps be seen to generalize default logic as well as autoepistemic logic.

Autoepistemic and default logics have also been widely investigated in connection with logic programs. Often the semantics of logic programs are given in terms of these nonmonotonic systems. The generalization of these logics afforded by CO* might allow the semantic interpretation of logic programs in terms of this modal logic of its weaker counterparts. Indeed, the manner in which priorities on defaults are handled in CO* might provide a natural account of the semantics of stratified programs. The implementation of default and autoepistemic theories within logic programs might also suggest translations of CO* into such terms.

Because of the ties with autoepistemic logic, CO* provides the opportunity and the framework for investigating the relationship that exists between traditional approaches to default reasoning and other areas such as conditional and probabilistic representations, revision semantics and subjunctive reasoning.

7.2.3 Probabilistic Semantics

In this section we discuss only briefly the potential interpretation of CO* in terms of standard and nonstandard probabilistic semantics. The most obvious probabilistic interpretation of both the normative and subjunctive conditionals is that provided by ε-semantics. We have seen that a sentence $A \Rightarrow B$ can be understood as asserting that the conditional probability $P(B|A)$ can be made arbitrarily high, even when considering the other constraints imposed by a theory. Some have argued that the semantic content of such a statement is unclear or, worse, simply untrue (Bacchus 1990). When one says "Birds normally fly," the intent is not to assert that the proportion of nonflying birds among all birds is infinitesimal. Certainly, $P(\texttt{fly}|\texttt{bird})$ is not greater than $1 - \varepsilon$ for any ε, in the "real world." However, this misses the point of ε-semantics and the logic of high

conditional probabilities. When a default rule $A \rightarrow B$ is interpreted this way, it is not meant to assert that $P(B|A)$ vanishes. Rather it says, and this seems to be Adams's (1975) original intent, that we can make the conditional probability arbitrarily high, if required, in the presence of whatever other constraints happen to be lying around.

In the context of default rules, it is important to keep in mind that probabilistic satisfiability and ε-consistency are secondary notions compared to ε-entailment. Adams's motivation was to provide a concept of inference that preserves arbitrarily high probabilities, just as classical entailment preserves truth. In this way, no matter how high the desired probability of some conclusion, it can be achieved simply by raising the probabilities of the premises to a sufficient degree. If accepting some default rule $A \rightarrow B$ requires approximately an "empirical" probability $P(B|A) = c$, then this is perhaps a question of learning default rules, or when such normative statements should be accepted. ε-entailment ensures that no matter what degree of certainty is required for acceptance, this level can be achieved in the derived default rule.

This still leaves intact the question of the precise meaning of $A \rightarrow B$ as a probabilistic default rule. As Bacchus (1990) has argued, and we have just reiterated, we certainly do not intend that $P(B|A)$ should vanish. It is on this point, we claim, that the modal interpretation of such defaults provides a natural and compelling semantic account of the logic of arbitrarily high probabilities. The meaning of the rule $A \rightarrow B$ is just that of the normative statement $A \Rightarrow B$: if A then normally B. In Chapter 4 we demonstrated that asserting that $P(B|A)$ can be made arbitrarily high (in the presence of other constraints) is precisely stating that the most normal (probable) states of affairs where A holds also satisfy B. In this way, CT4 and its extensions (as well as the preferential semantics of Kraus, Lehmann and Magidor (1990)) determine a (more) classical semantic interpretation of rules such as $A \rightarrow B$.

In Chapter 4, we also mentioned the relationship between CT4 and the probabilistic nonmonotonic system of inference developed by Bacchus (1990). There are a number of parallels between statements such as $A \Rightarrow B$ and $P(B|A) = c$ for some constant c, and the two systems in general. Bacchus's system, unlike most probabilistic approaches, combines both objective and subjective probabilities. Thus, default rules (in the form of conditional degrees of belief) can be implicitly linked to one another, and to other propositional (or evidential) knowledge through their connection to objective probability distributions, via the principle of direct inference. The distinction between background and evidence therefore remains at an epistemological level, rather than at the formal logical level. This is similar to our modal approach and allows much desirable interaction. Naturally, with probabilities, one can express not only default rules but also the strength of such rules, and the degrees of belief accorded default conclusions. This is beyond the capability of categorical systems like our modal systems. It should prove very enlightening to pursue these connections more formally and discover to what extent CT4 and its stronger counterparts can be viewed as either fragments of, or approximations to, probabilistic default reasoning.

Within the realm of belief revision, the relationship of revision functions on deductively closed theories to the revision of probability functions (for instance, through conditionalization) has been investigated. Gärdenfors (1988) has shown a very interesting connection. If we want to revise a probability function P to include belief in A, the resulting function P_A^* should have the property $P_A^*(A) = 1$. The key postulate for such probabilistic revision is that when A is "consistent" with P, so that $P(A) > 0$, then P_A^* should be the result of conditioning P on A. In other words, $P_A^*(B) = P(B|A)$ for all B. Given this requirement and some other postulates, any probabilistic revision function determines an AGM revision function in the sense now described. If we consider the set of *beliefs* determined by P to be the set of sentences assigned probability 1 by P, this forms a standard (deductively closed) belief set. Probabilistic revision functions define AGM revision

functions on the belief sets associated with these probability assignments.

The reverse connection is of more interest here if we wish to provide a probabilistic interpretation of CO* revision models. This is also a more difficult problem, for any (consistent) belief set K can be the belief set associated with arbitrarily many probability functions P. The question of which probability functions to associate with a given belief set is addressed in great detail in (Lindström and Rabinowicz 1989). A probabilistic semantics for CO*, both in terms of belief revision and default reasoning, might draw on these types of results, and, conversely, the nature of probabilistic revision might be illuminated by probabilistic approaches to default reasoning. Again, it appears that CO* might provide a unifying framework for such investigations.

Finally, we mention the possible relationship that exists among CO*, only knowing and maximum entropy. We alluded to the similarity of only knowing to maximum entropy in Chapter 6. Because of this, it might be possible to interpret only knowing as a qualitative, symbolic counterpart of distributions of high entropy. If the relationship is a strong one, it might lead to possible connectionist implementations of default theories in Boltzmann machines, which maximize entropy of networks subject to certain constraints, or maximum likelihood models (Rumelhart, McClelland and PDP Research Group 1986, Ch.7). In our case, such constraints would be default rules that must attain some high probability. As well, such equivalence might lead to some type of qualitative or symbolic interpretation of connectionist architectures.

Chapter 8

Concluding Remarks

8.1 Summary

In this thesis, we have presented several logical models for default reasoning and belief revision, and have shown how the two types of reasoning are related. Not only were we able to show how various existing formalisms fit into our framework, but how diverse systems can be related within our framework, how they can be generalized, and how they can be improved.

Regarding default reasoning, we presented a modal framework for the representation of normative conditionals. The logics CT4 and CT4D subsume, and suggest obvious generalizations of, ε-semantics, preference logics and rational logics. Furthermore, our systems bear a qualitative resemblance to systems and methods of (more standard) probabilistic reasoning, especially in the (implicit) distinction between background and evidence. A key advantage of our modal approach is the abandonment of the Limit Assumption, an assumption made in all existing conditional approaches to default reasoning.

We discussed the problem of irrelevance for conditional logics and extended our modal logics with a modal connective for inaccessibility. With this additional expressive power, we were able to axiomatize two existing solutions to the problem of irrelevance. Furthermore, we were able to distinguish between the type of (practical) irrelevance assumed in conditional default reasoning and the traditional (statistical) notion of irrelevance, or independence.

We then moved on to propose models of belief revision based on our extended modal logics. Again, while ignoring the Limit Assumption, these models subsume the standard models of revision, and suggest various generalizations. Our system appears to be the first "classical" logic of revision. Within it a number of concepts can be expressed, such as entrenchment, plausibility, integrity constraints and notions of belief, and these can be related to revision. Furthermore, the epistemic nature of revision was established and it was shown that autoepistemic logic can, in fact, be viewed as a specialization of AGM revision. As well, subjunctive conditionals were defined in such a way to respect the Ramsey test, and avoid the classical trivialization results.

Finally, we examined in some depth the connection between default reasoning and belief revision, suggesting that formally these are the same process, and that the normative and subjunctive are the same conditional. We distinguish the two processes, however, at a practical level; the logical apparatus is the same, but we use it with a different perspective in mind in each case. This connection allows us to see the relationship between a number of disparate types of reasoning systems.

We have seen that the modal approach to default reasoning and belief revision is extremely general and powerful. It is general in the sense that we have been able to embed a number of

existing systems and concepts within our logics in a way that respects the original intuitions of these systems. However, besides demonstrating the equivalence of our logics to these systems, we have shown how the expressive power of our approach can be used to extend existing systems considerably. CO*, for instance, can be construed as the first purely logical characterization of AGM belief revision. The fact that CO* allows the expression of propositional consistency at the object level permits such a model of revision. This is important because it determines the first sound and complete proof theory for existing semantic models like that of Grove (1988). Moreover, the semantic structures underlying CO* are far more general in that they do not obey the Limit Assumption. If restricted to statements about of revision, this generality is unnecessary; but the expressive power of CO* permits other types of constraints on revision to be specified. Statements of entrenchment, plausibility, integrity constraints and autoepistemic constraints can be written in CO*, all of which interact with subjunctives or statements of revision. These, we have seen, *can* determine models that fail the Limit Assumption. Thus CO* provides two distinct and crucial extensions to the AGM theory: a syntactic extension that allows the expression of important intensional constraints; and a semantic extension that permits the violation of the Limit Assumption.

Similar remarks apply to the embedding of subjunctive reasoning and autoepistemic reasoning in CO*. With respect to subjunctive reasoning, CO* respects Lewis's (1973a) conditional axioms, but allows for the influence of factual information on the subjunctive reasoning process, an influence not representable in VC. Autoepistemic reasoning can be performed in CO*, but we allow belief sets that have a certain structure, rather than the "flat" belief sets representable in OL. Thus the degree of entrenchment of beliefs can be expressed, and this can have an impact on our autoepistemic interpretation of default rules.

Embedding preferential consequence relations and ε-semantics in CO* determines similar extensions of these systems. In these cases, we provide an *alternative* logic for existing systems in which logical formulae are not constrained to have the form of simple conditionals. As well, the fact that CO* is a modal logic suggests a number of generalizations of the embedded systems.

While CO* extends many systems of defeasible inference, perhaps the most novel aspect of this work is the unifying perspective afforded by our logics. Not only are various concepts expressible in CO*, but they are expressed in very similar ways. Indeed, we have seen that the normative and subjunctive conditional are the same connective. While this relationship has been suggested in the literature, this is the first demonstration that *the logics of default reasoning and belief revision are the same.* Besides relating default reasoning and belief revision, this view also demonstrates connections among the other embedded systems. For instance, the relationship of autoepistemic logic to revision and subjunctive reasoning is quite interesting and important, but perhaps not too surprising, since the notion of belief is crucial to each of these. In contrast, the relationship of autoepistemic logic to ε-semantics is much more novel.

Not only does the relationship between systems appear, but also the relationship between certain concepts and previously disassociated systems. Qualitative possibility reduces to plausibility. While plausibility, entrenchment and integrity constraints have been discussed in the context of belief revision, the relationship within CO* shows them to be meaningful for default reasoning as well, a connection that has yet to be explored in the literature. Furthermore, the problem of irrelevance plagues conditional representations of defaults, yet has not been discussed in the belief revision community. Since the normative and subjunctive conditional are identical, irrelevance is indeed a problem for revision also. Figure 8.1 shows in tabular form the relationship of various systems and concepts.

normative conditional $\Rightarrow$	equivalent to	revision conditional $\stackrel{\text{KB}}{\longrightarrow}$
preferential consequence	logic CO*	AGM revision
rational consequence		subjunctive reasoning
ϵ-semantics		autoepistemic logic
		possibility theory
Other concepts: Entrenchment, plausibility, Integrity constraints (weak, strong, prioritized), Belief, Knowing at most, Only knowing		

Figure 8.1: Relationship of Systems to CO*

8.2 Future Research

This section will be short, as many important potential extensions of this work have been identified throughout, especially at the ends of Chapters 4, 5 and 6; and most of Chapter 7 is devoted to further speculation. Here we will only review some of the more important prospective areas and suggest a few others.

8.2.1 On A Quantificational Extension

The account of normative and subjunctive reasoning provided here is completely propositional. We have said little about how it might be extended to the first-order case, but clearly such an extension is required if this account is to be at all adequate. The difficulties involved in extending modal logics with quantification are too numerous and intricate to discuss adequately here, and have been well-documented in the literature. These have lead some, most notably Quine, to doubt the intelligibility of quantified modal logic (and, hence, modal logic *simpliciter*).[1] In particular, problems surrounding the identity of individuals within the scope of modal operators, and the associated reference terms, are especially intractable. In a possible worlds framework, the problem is that of deciding how to identify an individual in one world with an individual in another, and how to refer to or describe this individual in different worlds.

Before looking at these problems, perhaps we should ask why we adhere to the possible worlds framework in light of such difficulties. The answer, quite briefly, is that it is the most natural and compelling account of phenomena like subjunctive and normative reasoning. The Ramsey test for subjunctives (or some variant, for different types of conditionals) is a quite plausible model. To evaluate $A \stackrel{\text{KB}}{\longrightarrow} B$ is to imagine some situation where A holds and ask if B follows. These situations are just the *counterfactual situations* or *possible worlds* that constitute our models. Of course, not any counterfactual situation will do. An agent judges some of these A-worlds to be more plausible, more likely, or more similar to the "actual world" than others. Selection functions are often used for this purpose, but we adopt the view that an ordering of worlds (represented by an accessibility relation) is an appropriate model for an ideally rational agent.

Similar remarks apply to the evaluation of a normative conditional. Given some information an agent takes to be true, the appropriate default predictions are just those additional facts true

[1] See, for example, Quine (1961; 1960). For a survey of these difficulties and attempted responses see Haack (1978, Ch.10).

at the *most normal* or *most expected* situations satisfying the known facts.

What are competing analyses of subjunctives? One is Quine's account of *dispositions* (and the associated subjunctives).[2] To say some substance x is soluble is often taken to assert the associated subjunctive, that if x were placed in water it would dissolve. Quine explains this disposition of x without appeal to intensional concepts as follows: there is some y of sufficiently similar internal structure to x such that y has been placed in water and has dissolved. That the internal structure may not be known is irrelevant, and we can imagine applying this analysis to supposedly "counterfactual" situations. However, this cannot account for statements such as

> If a large meteorite were to strike New York City, the consequences would be catastrophic.

Certainly, there is no situation of "similar internal structure" on which to base this evaluation. Indeed, Quine ultimately denies that such statements are necessary in a "regimented theoretical language" (Haack 1978).

Goodman's (1955) approach to counterfactuals is quite similar to more recent attempts to reduce default rules to material implications (see Doyle (1983) and Chapter 2). It is largely unsuccessful for the same reasons. He requires that a counterfactual be evaluated (roughly) by conjoining to the antecedent a set of relevant conditions and asking whether the consequent follows from these (together with appropriate causal laws). Of course, specifying relevant conditions is very hard, somewhat unnatural (as we saw with the qualification problem), and does not allow us to reason about counterfactuals without knowing these conditions.

It would seem that the possible worlds analysis is the only game in town. Supposing this is so, what issues arise when quantification is added to modal logic? It is impossible to convey the full scope of these problems in this space, but we illustrate some of them with a few examples, and attempt to dismiss their purported devastating implications for the enterprise of quantified modal logic.

The key problem is that of quantifying into modal contexts. Consider Quine's (1961) examples. Obviously,

> The number of planets is greater than 7

is true, as is

> 9 is greater than 7

Furthermore, we take it as given that 9 is necessarily greater than 7

> $\Box$(9 is greater than 7)

and by existential generalization

> $\exists x \Box(x$ is greater than 7)

Quine argues that this last sentence cannot possibly be true, for if the x satisfying this last sentence is 9, then x is also the number of planets, and by substitution we can infer

> $\Box$(The number of planets is greater than 7)

---

[2] See, for instance, Quine (1960).

This is surely false.

There are two senses, however, of the sentence "The number of planets is necessarily greater than 7." One is the sense described above (which is clearly a false assertion). The second sense (and this has been suggested by Smullyan (1948)) can be described as

$$\exists x \text{ ((the number of planets } = x) \text{ and } \Box(x \text{ is greater than 7}))$$

which is true (supposing mathematical truths to be necessary). In other words, the *individual*, that in this world is denoted by "the number of planets", is necessarily greater than 9. Contrast this with the first sense, that "the number of planets" denotes in each world some individual that is greater than 7.[3]

To describe this second sense requires quantifying into a modal context, something Quine finds objectionable as it smacks of *essentialism* (that is, the characteristic that one and the same "entity" can exist at different possible worlds. One objection Quine has to this approach is the consequence that all identities are necessary. So, for example,

$$x = y \text{ implies } \Box(x = y)$$

must be a theorem of any such modal approach. Frege's classic example is often used to refute the necessity of identity. We know

The morning star = the evening star

but we should not infer that

$\Box$(The morning star = the evening star)

In fact, essentialism (as we propose it) sanctions no such inference. All we can conclude is that

$$\exists x \exists y(x \text{ is the morning star and } y \text{ is the evening star and } \Box(x = y))$$

Thus, Frege's example is not problematic at all. It is not necessary that the morning and evening stars are the same object,[4] but it is necessary that the object that is denoted by "the morning star" in the actual world and the object that is so denoted by "the evening star" are the same, since they refer to a *single individual.* If this individual exists in another world (and that is a different question), how can it be anything but itself?

Before asking what this says about essentialism, we must ask how we can identify individuals as being identical across possible worlds. Technically, this is quite easy, for instance, by adopting a set of *standard names* referring to the same individuals across worlds.[5] These are also known as *rigid designators*, and are unlike constant symbols (e.g., "the morning star") that may denote differently in different worlds. Kripke (1980) often uses proper names as rigid designators, and this is not wrong if every domain object has such a name; but this often leads to confusion in examples (especially among the detractors of quantified modal logic). For instance, if we ask

What if the morning star were a bird?

[3]This assumes "9" to designate rigidly (see below).

[4]It may be preferable to state that it is not necessary that one and the same object rises in both the morning and the evening depending on the season.

[5]Levesque (1984a; 1990), for example, uses this technique.

we suppose that in such a possible world "the morning star" would not be so-called.

Pursuing this purposely bizarre example, it should be clear that what we expect is a situation such that "x is a bird" is true of that situation, where "the morning star" denotes *this* x in the actual world. Thus, we can consistently assert

$$\exists x(x = \text{the morning star and } \Diamond(x \text{ is a bird}))^6$$

As Kripke (1980) emphasizes, the "problem" of *transworld identity* of individuals is not substantial. How can it be that the morning star is a bird in some possible world? If it is a bird, what properties does it have that allow us to identify *it* as the bird that "once was" the morning star? To ask these questions is to take the term "possible world" too seriously.[7] For an object to exist at another possible world is purely a matter of stipulation, and the properties of the object are also stipulated. " 'Possible worlds' are *stipulated*, not *discovered* by powerful telescopes. There is no reason we cannot *stipulate* that, in talking about what would have happened to Nixon in a counterfactual situation, we are talking about what would have happened to *him*" (Kripke 1980, p.44).[8]

Notice that the essentialism imposed by the possible worlds approach, stated this way, is absolutely minimal. The only essential property we are committed to is identity. It is hard to see how one could not insist on the necessity of identity if one is interested in the counterfactual consequences of the morning star "turning into" a bird. Furthermore, it is hard to imagine what other essential properties there *are*, once we allow a planet to become a bird (or even vanish). If the question of essential properties is crucial to knowledge representation, it is important to note that modal logic (on the view expressed here) takes no stand on essential properties other than identity. If it is supposed that, say, the morning star could have been nothing but a celestial object, then we can assert as a premise

$$\Box(\text{The morning star is a celestial object})$$

It is not, nor should it be, a theorem of any system. Essential properties can be stipulated, just as contingent properties can.

We've seen the implications of quantification for subjunctives, but how does it influence normatives? If someone asserts that all birds normally fly, we take it to mean

$$\forall x(x \text{ is a bird} \Rightarrow x \text{ can fly})$$

Notice that the predicates "bird" and "fly" are inside the scope of the conditional (or the modality). Thus, at the most normal worlds where *any domain individual* is a bird, that individual can fly. At the most normal worlds where the morning star is a bird, it flies. At the most normal worlds where Tweety is a bird, it flies too. This is not to say that they are one and the same set of worlds. When we say *all* birds normally fly, each of them is judged independently. There need not even be

---

[6] The observant reader may ask how we came to "fix" the denotation of "bird" is this sentence. Of course, it is not fixed to denote the same domain objects in this counterfactual situation as it does in the actual world. The connective $\Diamond$ could refer to a situation in which "bird" now refers to the planets. However, considerations of comparative similarity allow one to chose the *relevantly most similar* worlds satisfying "x is a bird," presumably where "bird" has roughly the same extension, and other properties remain about the same. Naturally, context can play a role in judgements of similarity. Different situations would be used to answer the question "If the morning star were a bird, would Tweety be a planet?"

[7] Cf. Lewis's (1968) *counterpart theory*, where individuals do not exist across worlds, but have counterparts in other worlds.

[8] The analogy should be clear: read "Nixon" as "the morning star" and "him" as "it".

a most normal world where all the individuals that are birds in the actual world are birds and can fly. The most normal world satisfying "x is a bird" can be different for each individual x (though we probably imagine otherwise). Given the latitude of the logic, we can constrain our conception of normality to fit most any pattern (for instance, by asserting that all birds (in the actual world) are birds in all more normal worlds, thereby assuring that there will be a point where all "actual" birds fly).

There are other questions that need to be addressed in the quantificational domain, but none, I believe, cast doubt on the ultimate success of a possible worlds framework. One problem is that of fixing the reference of proper names, definite and indefinite descriptions, and so on. This leads to the question of what individuals should be supposed to exist at various worlds. Hirst (1991) has pointed out that most research in knowledge representation has largely ignored questions regarding the types of individuals and classes of existence that form object domains.

Another question is the appropriateness of derivations in quantified versions of the normative and subjunctive logics in their present incarnation. For instance, consider the two sentences:[9]

> People normally admire some great athlete.
> People are normally not great athletes.

The obvious interpretations of these in our logic is

$$\forall x \exists y (\top \Rightarrow (\texttt{athlete}(y) \wedge \texttt{admires}(x, y)))$$
$$\forall x (\top \Rightarrow \neg \texttt{athlete}(x))$$

These sentences stand in contradiction, one implying the existence of great athletes, the other the nonexistence, in the normal state of affairs. Are the original sentences truly inconsistent or does context play a role in resolving the conflict? Perhaps the implicature of the first sentence, that great athletes exist, should give rise to the following translation:

$$\forall x ((\exists y \texttt{athlete}(y)) \Rightarrow (\exists y (\texttt{athlete}(y) \wedge \texttt{admires}(x, y))))$$

At the most normal situations where great athletes exist, everyone admires some great athlete.

8.2.2 Other Avenues

In Chapter 4, we suggested possible generalizations of both the normative conditional logics and the definition of the conditional itself. This might lead to other interesting notions of default inference, especially systems for dealing with inconsistent default conclusions, and systems based on non-normal modal logics (for instance, as suggested in (Marek, Shvarts and Truszczyński 1991)). These generalizations should prove interesting for revision as well. *Paraconsistent revision* should fit nicely in this framework. The notion of preorder revision presented here, adopted from (Katsuno and Mendelzon 1990), also remains unexplored in any detail. Of interest would be a set of postulates, similar to the AGM postulates for total order revision. Furthermore, the relationship of our modal default and revision systems to standard probabilistic systems remains to be pursued in depth. However, at this point, the potential connections look promising.

We have said little about the form of belief revision known as *update*. While this is distinct from traditional revision, it remains to be seen if some notion of update can be defined in our framework. This would certainly facilitate further comparison, and perhaps unification, of the two ideas, and of different varieties of subjunctives determined by these.

[9] Thanks to Len Schubert for bringing this example to my attention.

The connection of revision to autoepistemic logic has been formally developed here, but we have yet to determine the extent and practical significance of the generalization afforded by CO*. Further generalizations of autoepistemic logic are also readily apparent (again, for example, involving paraconsistency) and remain to be investigated. The connection of CO* to other forms of default reasoning should also prove interesting. We also expect a strong relationship to abduction to emerge in further investigations, given the results of Gärdenfors and Makinson relating Theorist and belief revision.

Extensions of CO* itself should help further blur the distinction between normatives and subjunctives. As it stands CO* can be used in two different "modes." By unifying normatives and subjunctives to a greater degree a logic might be developed in which both types of conditionals can be represented simultaneously. In this way, we would develop a system of reasoning in which the two types of reasoning could interact and influence the other.

Of prime concern is the solution to irrelevance developed in Chapter 5. Of course, we would like to have a more logical, semantic characterization of irrelevance, one that does not rely (explicitly) on the Z-ordering of rules. Moreover, even this solution, being equivalent to 1-entailment, is not intuitive in all cases. Hence, notions such as maximum entropy should prove fruitful; and such notions might be representable (at least, qualitatively) in a modal system.

Another key problem is that of nested conditionals. These are certainly well-defined in our system, but the global nature of conditionals implies that such nesting is not particularly meaningful. The *belief revision systems* of Gärdenfors (1988) that we discussed in Section 6.4 could be used to handle, for example, the revision of *revised* belief sets, or model nested conditionals. However, we would like to extend our model with the ability to represent nesting without requiring that revision of a belief set determine a new (and substantially unrelated) revision model for the updated theory. We would like to see a single modal structure that can deal with this task.

Appendix A

Proofs of Theorems: Chapter 4

In this appendix we present proofs of various propositions, lemmas, theorems and corollaries found in Chapter 4. Note that we use $\Rightarrow$ in place of $\not\Rightarrow$ in the proofs for typographical simplicity. We begin by showing the completeness of CT4 with respect to the class of reflexive, transitive Kripke structures. This follows almost immediately from the definition of $\Box$ in terms of $\Rightarrow$, the inclusion of the S4 axioms in CT4 and the completeness of S4. However, we provide an explicit proof here based on the completeness proof of S4 found in (Hughes and Cresswell 1984). We take $M = \langle W, R, \varphi \rangle$ to be a CT4-model with worlds w, v, etc.

Proposition A.1 *Let $M = \langle W, R, \varphi \rangle$ with $w \in W$. Defining $\Box\alpha \equiv_{\text{df}} \neg\alpha \Rightarrow \alpha$, we have that $M \models_w \Box\alpha$ iff for all wRv $M \models_v \alpha$.*

Proof If $M \models_w \neg\alpha \Rightarrow \alpha$ then the first clause of the truth conditions for $\Rightarrow$ (Definition 4.7) cannot apply, since the reflexivity of R requires a world where both α and $\neg\alpha$ hold. Thus for each wRv and vRu, $M \models_u \alpha$. By transitivity and reflexivity of R, $M \models_v \alpha$ for each wRv.

■

Because of this proposition, we will treat $\Box$ and $\Diamond$, defined in the conditional language, as we would in any modal system.

Lemma A.2 (Soundness) *If $\vdash_{CT4} \alpha$ then $\models_{CT4} \alpha$.*

Proof We show soundness by demonstrating the validity of each CT4 axiom, and showing each inference rule preserves validity (on CT4-models).

K: Suppose $M \models_w \Box(A \supset B)$ and $M \models_w \Box A$. Then in any world accessible to w, $A \supset B$ and A hold. Since each world satisfies propositional tautologies, each satisfies B, and $M \models_w \Box B$.

T: Suppose $M \models_w \Box A$. As R is reflexive, wRw and $M \models_w A$.

4: Suppose $M \models_w \Box A$, wRv and vRu. As R is transitive, wRu and $M \models_u A$. Thus $M \models_v \Box A$ for any such v and $M \models_w \Box\Box A$.

C: That this axiom is valid follows immediately from the truth conditions of $\Rightarrow$ and $\Box$ (via Proposition A.1).

Nec: Suppose A is valid, so $M \models_w A$ for all $w \in W$. Then for any $w \in W$, $M \models_w \Box A$ since $M \models_v A$ for all wRv. (This holds for any model M, of course).

MP: Suppose $A \supset B$ and A are valid. Then $M \models_w A \supset B, A$ for all $w \in W$ and since worlds verify tautologies, $M \models_w B$ for all $w \in W$.

■

Proposition A.3 *Let $\Gamma \subseteq \mathbf{L}_C$ be a maximal CT4-consistent set. If $\neg(\neg\alpha \Rightarrow \alpha) \in \Gamma$ then the following set is CT4-consistent:*

$$\{\neg\alpha\} \cup \{\gamma : \neg\gamma \Rightarrow \gamma \in \Gamma\}.$$

(Recall $\neg\alpha \Rightarrow \alpha$ is abbreviated $\Box\alpha$.)

Proof This follows from Lemma 2.3 of (Hughes and Cresswell 1984).

■

Lemma A.4 (Completeness) *If $\models_{CT4} \alpha$ then $\vdash_{CT4} \alpha$.*

Proof Completeness is shown by constructing a *canonical* CT4-model which falsifies all non-theorems of CT4. We will make use of several facts without proof regarding properties of maximal consistent sets in logics extending CPL (denoted generically PMCS) and derivability in S4, as found in (Hughes and Cresswell 1984). We follow the pattern of the completeness proof of S4 found there and refer the reader to (Hughes and Cresswell 1984) for more intricate details.

A canonical model $M = \langle W, R, \varphi\rangle$ will suffice, where W consists of all maximally CT4-consistent sets of sentences, φ makes propositional variables true at those worlds which contain them, and vRw iff

$$\{\alpha : \neg\alpha \Rightarrow \alpha \in v\} \subseteq w.$$

We begin by showing $M \models_w \alpha$ iff $\alpha \in w$. The proof proceeds by structural induction on α. For atomic variables this is obvious given the definition of φ. Assume this fact holds for α and β.

(1) $M \models_w \neg\alpha$ iff $M \not\models_w \alpha$. By hypothesis, this holds iff $\alpha \notin w$ iff (by PMCS) $\neg\alpha \in w$.

(2) $M \models_w \alpha \supset \beta$ iff $M \models_w \beta$ or $M \not\models_w \alpha$. By hypothesis, this holds iff $\beta \in w$ or $\alpha \notin w$ iff (by PMCS) $\alpha \supset \beta \in w$.

(3) $M \models_w \alpha \Rightarrow \beta$ iff for each w_1 such that wRw_1 either

1. there is some w_2 such that w_1Rw_2, $M \models_{w_2} \alpha$, and for each w_3 such that w_2Rw_3, $M \not\models_{w_3} \alpha$ or $M \models_{w_3} \beta$; or
2. for every w_2 such that w_1Rw_2, $M \not\models_{w_2} \alpha$.

By hypothesis this holds iff, for each such w_1, either

1. there is some w_2 such that w_1Rw_2, $\alpha \in w_2$, and for each w_3 such that w_2Rw_3, $\alpha \in w_3$ or $\beta \in w_3$; or
2. for every w_2 such that w_1Rw_2, $\alpha \notin w_2$.

Suppose case 1 holds. (We freely use $\Box$ and $\Diamond$ as the appropriate abbreviations at this point). For all such w_3 accessible to w_2 we have by PMCS, $\alpha \supset \beta \in w_3$. This implies, by Proposition A.3, that $\Box(\alpha \supset \beta) \in w_2$. Thus, by PMCS, $\alpha \wedge \Box(\alpha \supset \beta) \in w_2$. Again by Proposition A.3, $\Diamond(\alpha \wedge \Box(\alpha \supset \beta)) \in w_1$. (NOTE: the implications hold in the reverse direction, so case 1 holds iff this sentence is in w_1.)

Suppose case 2 holds. By Proposition A.3, $\Box\neg\alpha \in w_1$. (NOTE: the implications hold in the reverse direction, so case 2 holds iff this sentence is in w_1.)
Considering both cases we have that for any w_1 such that wRw_1,

$$\Box\neg\alpha \vee \Diamond(\alpha \wedge \Box(\alpha \supset \beta)) \in w_1.$$

By Proposition A.3,

$$\Box(\Box\neg\alpha \vee \Diamond(\alpha \wedge \Box(\alpha \supset \beta))) \in w$$

and using axiom **C** and PMCS, this can be true iff $\alpha \Rightarrow \beta \in w$.

This shows every nontheorem of CT4 to be falsifiable in some Kripke structure. All that remains to show completeness is the demonstration that M is indeed a CT4-model.

Reflexivity: Suppose $\Box\alpha \in w$. By axiom **T**, $\alpha \in w$ also, and by definition of R, wRw.

Transitivity: Suppose wRv, vRu and $\Box\alpha \in w$. By axiom **4**, $\Box\Box\alpha \in w$. By definition of R, $\Box\alpha \in v$ and $\alpha \in u$, hence wRu.

■

Theorem 4.2 *The system* CT4 *is characterized by the class of CT4-models; that is,* $\vdash_{CT4} \alpha$ *iff* $\models_{CT4} \alpha$.

Proof This follows immediately from Lemmas A.2 and A.4.

■

Proposition A.5 *Let* $\alpha \in \mathbf{L}_C$ *and* $\beta \in \mathbf{L}_M$. *Let* $M = \langle W, R, \varphi\rangle$ *be a CT4-model (and hence an S4-model),* $w \in W$, *and denote the satisfaction relation viewing* M *as an L-model by* $\models^L$. *Then* $M \models^{CT4}_w \alpha$ *iff* $M \models^{S4}_w \alpha^\circ$ *and* $M \models^{S4}_w \beta$ *iff* $M \models^{CT4}_w \beta^*$.

Proof This is immediate given the truth conditions of the primitive connectives in each logic, $\Rightarrow$ in the case of CT4 and $\Box$ in the case of S4.

■

Theorem 4.3 $\vdash_{CT4} \alpha \equiv (\alpha^\circ)^*$ *and* $\vdash_{S4} \alpha \equiv (\alpha^*)^\circ$. *Also,* $\vdash_{CT4} \alpha \supset \beta$ *iff* $\vdash_{S4} \alpha^\circ \supset \beta^\circ$. *In other words,* CT4 *and* S4 *are equivalent.*

Proof We want to show that the sets of provably equivalent sentences in each logic stand in one-to-one correspondence under a mapping which preserves implication. The mappings $\circ$ and $*$ induce such an isomorphism on the Lindenbaum algebras of CT4 and S4, as can be shown via Proposition A.5 and the completeness results for the two logics.

■

Proposition 4.4 *The following are valid in* CT4.[1]

ID $A \Rightarrow A$

LLE $\Box(A \equiv B) \supset ((A \Rightarrow C) \equiv (B \Rightarrow C))$

And $((A \Rightarrow B) \wedge (A \Rightarrow C)) \supset (A \Rightarrow (B \wedge C))$

RT $(A \Rightarrow B) \supset (((A \wedge B) \Rightarrow C) \supset (A \Rightarrow C))$

Or $((A \Rightarrow C) \wedge (B \Rightarrow C)) \supset ((A \vee B) \Rightarrow C)$

RCM $\Box(B \supset C) \supset ((A \Rightarrow B) \supset (A \Rightarrow C))$

CM $((A \Rightarrow B) \wedge (A \Rightarrow C)) \supset (A \wedge B \Rightarrow C)$

Proof We use semantic means to demonstrate the validity of these sentences. We take $M = \langle W, R, \varphi \rangle$ to be a CT4-model with worlds w, v, etc.

ID: If there are no accessible A-worlds for w, $M \models_w A \Rightarrow A$ holds trivially. Otherwise, since $\Box(A \supset A)$ holds at any world (in particular A-worlds)
$M \models_w A \Rightarrow A$.

LLE: If $M \models_w \Box\neg A$ this holds trivially. Otherwise, suppose $M \models_w \Box(A \equiv B)$. For w the set of accessible A-worlds must be exactly the set of accessible B-worlds. Then clearly the truth conditions $A \Rightarrow C$ and $B \Rightarrow C$ are satisfied identically at w and $M \models_w (A \Rightarrow C) \equiv (B \Rightarrow C)$.

And: Suppose $M \models_w (A \Rightarrow B) \wedge (A \Rightarrow C)$. For each accessible v one of $M \models_v \neg\Box A$ or

$$M \models_v \Diamond(A \wedge \Box(A \supset B)) \wedge \Diamond(A \wedge \Box(A \supset C))$$

holds. In the first case the conditions for $(A \Rightarrow (B \wedge C))$ hold trivially. In the second, there exists a vRu_1 such that $M \models_{u_1} A \wedge \Box(A \supset B)$. Since wRu_1 and $M \models_{u_1} \Diamond A$,

$$M \models_{u_1} \Diamond(A \wedge \Box(A \supset C)).$$

Thus for some $u_1 R u_2$,

$$M \models_{u_2} A \wedge \Box(A \supset B \wedge C).$$

Thus $M \models_w A \Rightarrow B \wedge C$.

RT: Suppose $M \models_w A \Rightarrow B$ and $M \models_w (A \wedge B) \Rightarrow C$. For each accessible v one of $M \models_v \neg\Box A$ or $M \models_v \Diamond(A \wedge \Box(A \supset B))$ holds. In the first case the conditions for $A \Rightarrow C$ hold trivially. In the second, $M \models_v \Box\neg(A \wedge B)$ is impossible so $M \models_v \Diamond(A \wedge B \wedge \Box(A \wedge B \supset C))$. Thus for some vRu_1,

$$M \models_{u_1} A \wedge B \wedge \Box(A \wedge B \supset C).$$

Since wRu_1, $M \models_{u_1} \Diamond(A \wedge \Box(A \supset B))$. Thus for some $u_1 R u_2$, $M \models_{u_2} A \wedge \Box(A \supset B)$. But since $M \models_{u_1} \Box(A \wedge B \supset C)$, we have $M \models_{u_2} A \wedge \Box(A \supset C)$. Thus $M \models_w A \Rightarrow C$.

Or: Suppose $M \models_w (A \Rightarrow C) \wedge (B \Rightarrow C)$. For each accessible v we have

$$M \models_v \neg\Box A \vee \Diamond(A \wedge \Box(A \supset C))$$

[1] Many of the names of rules and theorems are taken or adapted from (Nute 1980; Delgrande 1987; Lehmann 1989).

and

$$M \models_v \neg\Box B \vee \Diamond(B \wedge \Box(B \supset C)).$$

(a) If $M \models_v \neg\Box A \wedge \neg\Box B$ then $M \models_v \Box\neg(A \vee B)$.
(b) Suppose $M \models_v \Diamond(B \wedge \Box(B \supset C))$. Then for some vRu we have

$$M \models_u B \wedge \Box(B \supset C)).$$

But since wRu, either $M \models_u \neg\Box A$ or $M \models_u \Diamond(A \wedge \Box(A \supset C))$. In the first case,

$$M \models_u (A \vee B) \wedge \Box(A \vee B \supset C)$$

is clearly true. In the second case, for some uRu_1, $M \models_{u_1} A \wedge \Box(A \supset C))$. As $\Box(B \supset C)$ holds at u,

$$M \models_{u_1} (A \vee B) \wedge \Box(A \vee B \supset C),$$

and $M \models_v \Diamond((A \vee B) \wedge \Box(A \vee B \supset C))$.
(c) A similar argument holds if $M \models_v \Diamond(A \wedge \Box(A \supset C))$.
Thus $\Box\neg(A \vee B) \vee \Diamond((A \vee B) \wedge \Box(A \vee B \supset C))$ holds at each such v, and $M \models_w A \vee B \Rightarrow C$.

RCM: If $M \models_w \Box(B \supset C)$ and $M \models_w A \Rightarrow B$ then $M \models_w A \Rightarrow C$ must hold since the set of B-worlds accessible to w is a subset of the set of C-worlds accessible to w.

CM: Suppose $M \models_w (A \Rightarrow B) \wedge (A \Rightarrow C)$. For each accessible v we have that either $M \models_v \neg\Box A$ or

$$M \models_v \Diamond(A \wedge \Box(A \supset B)) \wedge \Diamond(A \wedge \Box(A \supset C))$$

holds. In the first case the conditions for $A \wedge B \Rightarrow C$ hold trivially. In the second, there exists a vRu such that $M \models_u A \wedge \Box(A \supset B)$. But wRu, so $M \models_u \Diamond(A \wedge \Box(A \supset C))$. This means for some uRu_1,

$$M \models_{u_1} A \wedge \Box(A \supset B) \wedge \Box(A \supset C));$$

which implies

$$M \models_{u_1} A \wedge B \wedge \Box(A \wedge B \supset C).$$

Thus $M \models_v \Diamond(A \wedge B \wedge \Box(A \wedge B \supset C))$.

■

Proposition 4.8 $\vdash_{CT4D} \Box(\Box\neg A \vee \Diamond(A \wedge \Box(A \supset B))) \equiv \Box\neg A \vee \Diamond(A \wedge \Box(A \supset B))$.

Proof Suppose $M \models_w \Box(\Box\neg A \vee \Diamond(A \wedge \Box(A \supset B)))$. Then at every v such that wRv,

$$M \models_v \Box\neg A \vee \Diamond(A \wedge \Box(A \supset B)).$$

(a) If for every such v it is the case that $M \models_v \Box\neg A$, then $M \models_w \Box\neg A$.

(b) If for some such v it is the case that

$$M \models_v \Diamond(A \wedge \Box(A \supset B)),$$

then $M \models_w \Diamond(A \wedge \Box(A \supset B))$.

So

$$M \models_w \Box\neg A \vee \Diamond(A \wedge \Box(A \supset B)),$$

and

$$\vdash_{CT4D} \Box(\Box\neg A \vee \Diamond(A \wedge \Box(A \supset B))) \supset \Box\neg A \vee \Diamond(A \wedge \Box(A \supset B)).$$

Suppose $M \models_w \Box\neg A \vee \Diamond(A \wedge \Box(A \supset B))$.

(a) If $M \models_w \Box\neg A$ then $M \models_w \Box\Box\neg A$.

(b) Suppose $M \models_w \Diamond(A \wedge \Box(A \supset B))$. Then for some wRu we have

$$M \models_u A \wedge \Box(A \supset B).$$

For each wRv either vRu or uRv, as R is connected. If vRu then clearly

$$M \models_v \Diamond(A \wedge \Box(A \supset B)).$$

If uRv then either $M \models_v \Box\neg A$ or (since $M \models_u \Box(A \supset B)$)

$$M \models_v \Diamond(A \wedge \Box(A \supset B)).$$

Thus for each wRv, we have

$$M \models_v \Box\neg A \vee \Diamond(A \wedge \Box(A \supset B));$$

thus

$$M \models_w \Box(\Box\neg A \vee \Diamond(A \wedge \Box(A \supset B))).$$

Hence

$$\vdash_{CT4D} \Box\neg A \vee \Diamond(A \wedge \Box(A \supset B)) \supset \Box(\Box\neg A \vee \Diamond(A \wedge \Box(A \supset B))).$$

■

Proposition 4.9 *The following are valid in* CT4D.

RM $((A \Rightarrow C) \wedge (A \wedge B \not\Rightarrow C)) \supset (A \Rightarrow \neg B)$

CV $(A \not\Rightarrow B) \supset ((A \Rightarrow C) \supset (A \wedge \neg B \Rightarrow C))$

Proof We prove the validity of these sentences by semantic means using the simpler definition of $\Rightarrow$ appropriate for CT4D.

RM: Suppose $M \models_w (A \Rightarrow C) \wedge (A \wedge B \not\Rightarrow C)$. For some accessible v, we have $M \models_v A \wedge \Box(A \supset C)$, since $M \models_w \Box\neg A$ is contradicted by $A \wedge B \not\Rightarrow C$. It must be the case that $M \models_v \Box\neg(A \wedge B)$, for otherwise there would exist some vRu such that

$$M \models_u A \wedge B \wedge \Box(A \wedge B \supset C),$$

contradicting the fact that $M \models_w A \wedge B \not\Rightarrow C$. Since $M \models_v A \wedge \Box(A \supset \neg B)$, we have $M \models_w A \Rightarrow \neg B$.

CV: Suppose $M \models_w (A \not\Rightarrow B) \wedge (A \Rightarrow C)$. Then for some accessible v, we have $M \models_v A \wedge \Box(A \supset C)$. It must be the case that $M \models_v \Diamond(A \wedge \neg B)$ since $M \models_w A \not\Rightarrow B$; therefore

$$M \models_u A \wedge \neg B \wedge \Box(A \wedge \neg B \supset C)$$

for some vRu. Thus $M \models_w A \wedge \neg B \Rightarrow C$.

■

Proposition 4.11 $(A \Rightarrow B) \wedge (A \Rightarrow C) \supset (A \wedge B \Rightarrow C) \notin$ N.

Proof We construct an N-model $M = \langle W, f, \varphi \rangle$ which falsifies an instance of this sentence,, the rule **CM**. We assume a finite language of three variables A, B, C. The set of possible worlds for M is $W = \{v, w\}$, with $\varphi(A) = \{v, w\}$, $\varphi(B) = \{v\}$, and $\varphi(C) = \emptyset$. The selection function f is specified as follows:

$$f(v, \|\alpha\|) = \begin{cases} v & \text{if } \|\alpha\| = v \\ \emptyset & \text{otherwise} \end{cases}$$

$$f(w, \|\alpha\|) = \emptyset \text{ for all } \alpha$$

To show that M is indeed an N-model requires proving that each of the four conditions on the selection function f is met. This can be verified by a somewhat tedious examination of all cases involving the eight elements in the domain of f: two possible worlds crossed with four propositions over W.

Now we notice that both $M \models_v A \supset B$ and $M \models_v A \supset C$ hold since $f(v, \|A\|) \subseteq \|B\|$ and $f(v, \|A\|) \subseteq \|C\|$ (since $f(v, \|A\|) = \emptyset$). However,

$$f(v, \|A \wedge B\|) = \{v\} \not\subseteq \|C\| = \emptyset.$$

Hence, $M \models_v A \wedge B \not\Rightarrow C$, contradicting the proposed theorem **CM**.

■

Lemma A.6 (Soundness) *If* $\vdash_{P^*} \alpha$ *then* $\models_{P^*} \alpha$.

Proof The soundness of the axioms and rules which constitute P and CPL is obvious. To deal with the remaining axiom **T**, let M be some P-model and assume $M \models_w \neg A \mathrel{|\!\sim} A$. Then $\neg A \mathrel{|\!\sim} \bot$ holds (by **ID**, **RCM**, and **LLE**) and at every minimal $\neg A$-world, $\bot$ holds. By smoothness $\|\neg A\|^M = \emptyset$, and $M \models_w A$. Hence, $\vdash_{P^*} \alpha$ implies $\models_{P^*} \alpha$.

■

Lemma A.7 (Completeness) *If* $\models_{P^*} \alpha$ *then* $\vdash_{P^*} \alpha$.

Proof To show completeness, we will show that every P*-consistent set is satisfiable. Let Γ' be such a set of formulae, and let Γ be any maximal consistent extension of Γ'. Let $K = \{\alpha \mathrel{|\!\sim} \beta : \alpha \mathrel{|\!\sim} \beta \in \Gamma\}$, the set of assertions in this maximal extension. Clearly, K forms a preferential

consequence relation, and by Theorem 5 there exists a P-model W which determines this relation; that is, $W \models \alpha \mathrel{|\!\sim} \beta$ iff $\alpha \mathrel{|\!\sim} \beta \in K$. Let $\bar{W} = \langle S, \varphi, \prec \rangle$.

Now we will construct a model which satisfies Γ, namely $W' = \langle S', \varphi', \prec' \rangle$, where $W' = W \cup \{w\}$, $\prec' = \prec \cup \{\langle s, w \rangle : s \in S\}$, and $\varphi'(A) = \varphi(A)$ except when variable $A \in \Gamma$, in which case $\varphi'(A) = \varphi(A) \cup \{w\}$. Clearly, W' is a P-model as $\prec'$ is still a strict partial order and smoothness is unaffected by the addition of w to W. We want to show that $W' \models_w \Gamma$, and hence that Γ (and Γ') is satisfiable. We proceed by structural induction on formulae in Γ. Obviously, for any atomic variable A, $A \in \Gamma$ iff $W' \models_w A$. Assume this property holds for α and β.

(a) $\neg\alpha \in \Gamma$ iff $\alpha \notin \Gamma$ iff $W' \not\models_w \alpha$ iff $W' \models_w \neg\alpha$.

(b) $\alpha \supset \beta \in \Gamma$ iff $\alpha \notin \Gamma$ or $\beta \in \Gamma$ iff $W' \not\models_w \alpha$ or $W' \models_w \beta$ iff $W' \models_w \alpha \supset \beta$.

(c) Suppose then that $\alpha \mathrel{|\!\sim} \beta \in \Gamma$. Then $W \models \alpha \mathrel{|\!\sim} \beta$, so at every minimal α-world in S, β is true. Now consider two cases:

1. Suppose $||\alpha||^W \neq \emptyset$. By smoothness, this is true iff there exists a minimal α-world. Clearly every such world is still minimal in W' (since $s \prec' w$ for all $s \in S$). So this holds iff $W' \models \alpha \mathrel{|\!\sim} \beta$.
2. Suppose $||\alpha||^W = \emptyset$. In such a case, $W \models \alpha \mathrel{|\!\sim} \neg\alpha$, and by definition of W, this can be true iff $\alpha \mathrel{|\!\sim} \neg\alpha \in \Gamma$. By axiom **T** and the fact that Γ is maximal, this is equivalent to $\neg\alpha \in \Gamma$, and by the inductive hypothesis, $\neg\alpha \in \Gamma$ iff $W' \models_w \neg\alpha$ iff $||\alpha||^{W'} = \emptyset$, and (trivially) iff $W' \models \alpha \mathrel{|\!\sim} \beta$.

■

Theorem 4.16 $\models_{P*} \alpha$ *iff* $\vdash_{P*} \alpha$.

Proof This follows immediately from Lemmas A.6 and A.7.

■

Theorem 4.17 $\models_{R*} \alpha$ *iff* $\vdash_{R*} \alpha$.

Proof The proof proceeds exactly as that of Theorem 4.16, except we construct a new model $W' = \langle S', \varphi', \prec' \rangle$ given a $W = \langle S, \varphi, \prec \rangle$ which is ranked by a total order $\langle \Omega, < \rangle$ and a function $f : S \mapsto \Omega$ (where $f(s) < f(t)$ iff $s \prec t$), and where model W determines the appropriate rational consequence relation. The only thing to verify is that the new model W' can be ranked. Define the total order $\langle \Omega', <' \rangle$ such that $\Omega' = \Omega \cup \{\sigma\}$ and $<' = < \cup \{\langle \omega, \sigma \rangle : \omega \in \Omega\}$. Let $f' : S' \mapsto \Omega'$ such that $f' = f \cup \{\langle w, \sigma \rangle\}$. It is not hard to verify that f' ranks W'; that is, $f'(s) <' f'(t)$ iff $s \prec' t$.

■

Theorem 4.18 *Let KB be a set of conditional assertions and A an assertion. A is preferentially entailed by KB iff* $KB \vdash_{CT4} A$.

Proof This follows immediately from Theorem 4.21 which follows.

■

Theorem 4.19 *Let KB be a set of conditional assertions and A an assertion. A is rationally entailed by KB iff* $KB \vdash_{CT4D} A$.

Proof This follows immediately from Theorem 4.22 which follows.

■

Theorem 4.21 *Let* $A \in \mathbf{L}_C^-$. *Then* $\vdash_{CT4-} A$ *iff* $\vdash_{P*} A$.

Proof That all axioms and rules of P* are valid in CT4 can be shown by semantic means quite readily (see, for example, Theorem 4.4). From this it follows that all P*-valid sentences are CT4-valid. To prove the converse, we will show by structural induction on formulae that all P*-satisfiable sentences are CT4-satisfiable. For atomic variables this is obvious. We assume α and β are P*-satisfiable only if they are CT4-satisfiable (which means iff they are CT4-satisfiable, since P*-validity implies CT4-validity).

(a) $\neg\alpha$ is P*-satisfiable iff α is not P*-valid iff (by inductive hypothesis) α is not CT4-valid iff $\neg\alpha$ is CT4-satisfiable.

(b) $\alpha \supset \beta$ is P*-satisfiable iff α is not P*-valid or β is P*-satisfiable iff (by inductive hypothesis) α is not CT4-valid or β is CT4-satisfiable iff $\alpha \supset \beta$ is CT4-satisfiable.

(c) Suppose $\alpha \mathrel{|\!\sim} \beta$ is P*-satisfiable (note that α and β must be free of the connective $\mathrel{|\!\sim}$). This is the case only there exists a preferential model $M = \langle S, \varphi, \prec\rangle$ which satisfies $\alpha \mathrel{|\!\sim} \beta$; that is, where $M \models_s \beta$ at all α-minimal worlds $s \in S$.
Let $M' = \langle S, R, \varphi\rangle$ be such that $R = \succ \cup \{\langle s, s\rangle : s \in S\}$. As R is the reflexive closure of (the inversion of) partial order $\prec$, M' is clearly as CT4-model. We want to show that M' satisfies $\alpha \Rightarrow \beta$, so consider two cases:

1. Suppose $||\alpha||^M = \emptyset$. Then $||\alpha||^{M'} = \emptyset$, and M' trivially satisfies $\alpha \Rightarrow \beta$ (at any world s).
2. Suppose $||\alpha||^M \neq \emptyset$. For some $s \in S$, we want to show that $M' \models_s \alpha \Rightarrow \beta$, so for any $s \in S$, let sRw. If $M' \not\models_w \Box\neg\alpha$, then there exists a w_1 such that wRw_1 and $M' \models_{w_1} \alpha$. By the smoothness of M, either w_1 is α-minimal in $\prec$ or there is a w_2 such that $w_2 \prec w_1$ and w_2 is α-minimal. In either case, one of these (say w_1) is such that wRw_1, and w_1 is α-minimal. By definition of M, β holds at all α-minimal worlds, hence at w_1. As well, $\alpha \supset \beta$ holds at all w_2 such that w_1Rw_2 (by minimality of w_1).
So for all w such that sRw, $M' \models_w \Box\neg\alpha \vee \Diamond(\alpha \wedge \Box(\alpha \supset \beta))$; hence, $M' \models_s \alpha \Rightarrow \beta$.

Thus any P*-satisfiable sentence is CT4-satisfiable and the theorems of P* are identical to those of CT4-.

■

Theorem 4.22 *Let $A \in \mathbf{L}_C^-$. Then $\vdash_{CT4D^-} A$ iff $\vdash_{R^*} A$.*

Proof That all axioms and rules of R* are valid in CT4D can be shown by semantic means quite readily (see, for example, Theorem 4.9). From this it follows that all R*-valid sentences are CT4D-valid. To prove the converse, we will show by structural induction on formulae that all R*-satisfiable sentences are CT4D-satisfiable. For atomic variables this is obvious. We assume α and β are R*-satisfiable only if they are CT4D-satisfiable (which means iff they are CT4D-satisfiable, since R*-validity implies CT4D-validity).

(a) $\neg\alpha$ is R*-satisfiable iff α is not R*-valid iff (by inductive hypothesis) α is not CT4D-valid iff $\neg\alpha$ is CT4D-satisfiable.

(b) $\alpha \supset \beta$ is R*-satisfiable iff α is not R*-valid or β is R*-satisfiable iff (by inductive hypothesis) α is not CT4D-valid or β is CT4D-satisfiable iff $\alpha \supset \beta$ is CT4D-satisfiable.

(c) Suppose $\alpha \mathrel{|\!\sim} \beta$ is R*-satisfiable (note that α and β must be free of the connective $\mathrel{|\!\sim}$). This is the case only if there exists a ranked model $M = \langle S, \varphi, \prec\rangle$ which satisfies $\alpha \mathrel{|\!\sim} \beta$; so, $M \models_s \beta$ at all α-minimal worlds s. Assume M is ranked by $f : S \mapsto \Omega$ and total order $\langle\Omega, <\rangle$, so that $f(s) < f(t)$ iff $s \prec t$. By the properties of ranked models, all α-minimal worlds s have the same rank, say $f(s) = \kappa$. So for all $w \in ||\alpha||^M$, $f(w) > \kappa$.

Let $M' = \langle S, R, \varphi\rangle$ be such that sRt iff $f(s) \geq f(t)$. Clearly R is reflexive and transitive. Suppose sRt and sRu: either $f(t) < f(u)$ or $f(u) < f(t)$ or $f(t) = f(u)$, so either uRt or tRu (or both). Hence, R is a connected relation, and M' is a CT4D-model. We want to show that M' satisfies $\alpha \Rightarrow \beta$, so consider two cases:

1. Suppose $||\alpha||^M = \emptyset$. Then $||\alpha||^{M'} = \emptyset$, and M' trivially satisfies $\alpha \Rightarrow \beta$ (at any world s).
2. Suppose $||\alpha||^M \neq \emptyset$. Let $s \in S$ be such that $f(s) \geq \kappa$ (where κ is the rank of all α-minimal worlds). Consider any w where sRw. (i) If $f(w) \geq \kappa$, then there exists a w_1 such that wRw_1 (let $f(w_1) = \kappa$) and $M' \models_{w_1} \alpha$. Furthermore, w_1 is α-minimal in $\prec$. By definition of M, β holds at all α-minimal worlds, hence at w_1. As well, $\alpha \supset \beta$ holds at all w_2 such that w_1Rw_2 (by minimality of w_1). So $M' \models_w \Diamond(\alpha \wedge \Box(\alpha \supset \beta))$. (ii) If $f(w) < \kappa$, then by the definition of κ, for all w_1 such that wRw_1, $M' \not\models_{w_1} \alpha$, and $M' \models_w \Box\neg\alpha$.
   So for all w such that sRw, $M' \models_w \Box\neg\alpha \vee \Diamond(\alpha \wedge \Box(\alpha \supset \beta))$; hence, $M' \models_s \alpha \Rightarrow \beta$.

Thus any R*-satisfiable sentence is CT4D-satisfiable and the theorems of R* are identical to those of CT4D-.

■

Theorem 4.26 *T^P is CT4-consistent iff every non-empty subset of T is confirmable.*

Proof Suppose every non-empty subset of T is confirmable. We can then construct a CT4-model M of T^P. We omit details here, noting that the model Z_T, as defined in Section 5.4.2, is such a CT4-model (see Corollary 5.17).

Suppose some subset $S \subseteq T$ is not confirmable. Then for every $\alpha \Rightarrow \beta \in S$, the set

$$\{\alpha\} \cup \{\gamma \supset \delta : \gamma \Rightarrow \delta \in S\}$$

is propositionally unsatisfiable. Let $|S| = n$. Choose any $A \Rightarrow B$ in S.

If T^P is consistent, there is some CT4-model M with world w such that $M \models_w T^P$. Since $\Diamond A \in T^P$, we have $M \models_v A \wedge \Box(A \supset B)$ for some wRv. By the definition of $\Rightarrow$, it must be the case that

$$M \models_v \Box\neg\alpha \vee \Diamond(\alpha \wedge \Box(\alpha \supset \beta))$$

for each $\alpha \Rightarrow \beta \in S$. If $M \models_v \Box\neg\alpha$ for each, then we would have a valuation satisfying A and $\alpha \supset \beta$ for each $\alpha \Rightarrow \beta$ in S, contradicting the fact that S is not confirmable. So $M \models_v \Diamond(\alpha \wedge \Box(\alpha \supset \Box))$ for at least one member of S (distinct from $A \Rightarrow B$). Let's call this member $C \Rightarrow D$.

This implies that, for some vRu, $M \models_u C \wedge \Box(C \supset D)$. Since vRu, we have $M \models_u A \supset B$, and by a similar argument we must have $M \models_u \Diamond(\alpha \wedge \Box(\alpha \supset \beta))$ for at least one member of S (distinct from *both* $A \Rightarrow B$ and $C \Rightarrow D$). Repeating this argument $n-1$ times, we find that there must exist some world satisfying $\alpha \supset \beta$ for each $\alpha \Rightarrow \beta$ in S, as well as the antecedent of one such conditional. This, however, contradicts the fact that S is not confirmable. Thus, T^P is inconsistent.

■

Theorem 4.28 *Let $T \subseteq \mathbf{L}_C$ be a finite set of simple conditionals. $T^\varepsilon \cup \{A \rightarrow B\}$ is substantively inconsistent iff $T^P \cup \{\Diamond A\}$ is CT4-consistent while $T^P \cup \{\Diamond A, A \Rightarrow B\}$ is CT4-inconsistent.*

Proof $T^\varepsilon \cup \{A \rightarrow B\}$ is substantively inconsistent iff $T^\varepsilon \cup \{A \rightarrow B\}$ is not ε-consistent, while $T^\varepsilon \cup \{A \rightarrow \top\}$ is ε-consistent. We notice that $A \Rightarrow \top$ is a theorem of CT4; so by Theorems 4.23 and 4.26, the result follows.

■

Theorem 4.29 *Let $T \subseteq \mathbf{L}_C$ be a finite set of simple conditionals such that $T^P \cup \{\Diamond A\}$ is CT4-consistent. $T^P \vdash_{CT4} A \Rightarrow B$ iff T^ε ε-entails $A \rightarrow B$.*

Proof Since T^P is CT4-consistent we know (by Theorem 4.26) that T^ε is ε-consistent.

By the definition of ε-entailment, T^ε ε-entails $A \rightarrow B$ iff $T^\varepsilon \cup \{A \rightarrow \neg B\}$ is substantively inconsistent; by Theorem 4.28, this holds iff $T^P \cup \{\Diamond A, A \Rightarrow \neg B\}$ is CT4-inconsistent. By Theorem 4.26, this holds iff some non-empty subset of $T \cup \{A \Rightarrow \neg B\}$ is not confirmable. As in the construction used in the proof of Theorem 4.26, this is true exactly when any model of $T^P \cup \{\Diamond A\}$ entails $A \Rightarrow B$. Since $\neg\Box A$ entails $A \Rightarrow B$ and $\neg\Box A \vee \Diamond A$ is CT4-valid, we have $T^P \vdash_{CT4} A \Rightarrow B$.

■

Corollary 4.30 *Let $T \subseteq \mathbf{L}_C$ be a finite set of simple conditionals including strict sentences. T^P is CT4-inconsistent iff T^ε is not ε-consistent.*

Proof The proof is almost identical to that of Theorem 4.26.

■

Corollary 4.31 *Let $T \subseteq \mathbf{L}_C$ be a finite set of simple conditionals including strict sentences, such that $T^P \cup \{\Diamond A\}$ is CT4-consistent. $T^P \vdash_{CT4} A \Rightarrow B$ iff T^ε ε-entails $A \rightarrow B$.*

Proof The proof follows that of Theorem 4.29.

■

Appendix B

Proofs of Theorems: Chapter 5

In this appendix we present proofs of various propositions, lemmas, theorems and corollaries found in Chapter 5.

Lemma B.1 *The following are derived theorems of CO:*

T* $\boxminus A \supset A$

5* $\diamondminus A \supset \boxminus\diamondminus A$

B* $A \supset \boxminus\diamondminus A$

4* $\boxminus A \supset \boxminus\boxminus A$

Proof (a) **T*** follows directly from **T** and the definition of $\boxminus$.

(b) **5*** $\diamondminus A \supset \boxminus\diamondminus A$:

(1)	$\diamondminus(\Box A \wedge \boxminus B) \supset \boxminus(A \vee B)$	instance of **H**
(2)	$\neg\boxminus(A \vee A) \supset \neg\diamondminus(\Box A \wedge \boxminus A)$	(1), subst. A for B
(3)	$\neg\boxminus A \supset \neg\diamondminus\boxminus A$	(2), defn of $\boxminus$
(4)	$\diamondminus A \supset \boxminus\diamondminus A$	(3), subst. $\neg A$ for A

(c) **B*** $A \supset \boxminus\diamondminus A$:

(1)	$\diamondminus A \supset \boxminus\diamondminus A$	instance of **5***
(2)	$A \supset \diamondminus A$	contrapostive of **T***
(3)	$A \supset \boxminus\diamondminus A$	(1), (2)

(d) **4*** $\boxminus A \supset \boxminus\boxminus A$:

(1)	$\diamondminus\boxminus A \supset \boxminus A$	contrapostive of **5***
(2)	$\boxminus\diamondminus\boxminus A \supset \boxminus\boxminus A$	(1), **Nec**
(3)	$\boxminus A \supset \boxminus\diamondminus\boxminus A$	instance of **B***
(4)	$\boxminus A \supset \boxminus\boxminus A$	(2), (3)

■

Lemma 5.3 *Any instance of a Humberstone schema* **H*** *is derivable in CO.*

Proof Recall the Humberstone schemata

H* $D(\Box\alpha \wedge \overleftarrow{\Box}\beta) \supset B(\alpha \vee \beta)$,
where D is any sequence of the connectives $\Diamond$ and $\overleftarrow{\Diamond}$ having length ≥ 0, and B is any such sequence of $\Box$ and $\overleftarrow{\Box}$.

Using axiom **4***, it is easy to show that from each of $\Diamond\Diamond A$, $\overleftarrow{\Diamond}\Diamond A$, $\Diamond\overleftarrow{\Diamond}A$, and $\overleftarrow{\Diamond}\overleftarrow{\Diamond}A$, one can derive $\overleftarrow{\Diamond}A$. Thus, an easy inductive proof on the length of any "diamond" sequence D demonstrates that $D(\alpha) \supset \overleftarrow{\Diamond}\alpha$ is derivable. Similarly, we can show using **4*** that each of $\Box\Box A$, $\overleftarrow{\Box}\Box A$, $\Box\overleftarrow{\Box}A$, and $\overleftarrow{\Box}\overleftarrow{\Box}A$ is derivable from $\overleftarrow{\Box}A$; and hence, $\overleftarrow{\Box}\alpha \supset B(\alpha)$ is derivable for any "box" sequence B. Together with axiom **H** this shows any instance of the Humberstone schema to be derivable in CO.

■

Lemma B.2 (Soundness) *If* $\vdash_{CO} \alpha$ *then* $\models_{CO} \alpha$.

Proof We show soundness by demonstrating the validity of each CO axiom, and showing each inference rule preserves validity (on CO-models).

K: Suppose $M \models_w \Box(A \supset B)$ and $M \models_w \Box A$. Then in any world accessible to w, $A \supset B$ and A hold. Since each world satisfies propositional tautologies, each satisfies B, and $M \models_w \Box B$.

K': Suppose $M \models_w \overleftarrow{\Box}(A \supset B)$ and $M \models_w \overleftarrow{\Box}A$. Then in any world inaccessible to w, $A \supset B$ and A hold. Since each world satisfies propositional tautologies, each satisfies B, and $M \models_w \overleftarrow{\Box}B$.

4: Suppose $M \models_w \Box A$, wRv and vRu. As R is transitive, wRu and $M \models_u A$. Thus $M \models_v \Box A$ for any such v and $M \models_w \Box\Box A$.

S Assume $M \models_w A$ and not wRv. By the requirement that R be totally-connected, it must be that vRw and $M \models_v \Diamond A$. Hence $M \models_w \overleftarrow{\Box}\Diamond A$ and $A \supset \overleftarrow{\Box}\Diamond A$ is valid.

Nec: Suppose A is valid, so $M \models_w A$ for all $w \in W$. Then for any $w \in W$, $M \models_w \Box A$ since $M \models_v A$ for all wRv. Similarly for $\overleftarrow{\Box}A$. (This holds for any model M, of course).

MP: Suppose $A \supset B$ and A are valid. Then $M \models_w A \supset B, A$ for all $w \in W$ and since worlds verify tautologies, $M \models_w B$ for all $w \in W$.

■

Lemma B.3 (Completeness) *If* $\models_{CO} \alpha$ *then* $\vdash_{CO} \alpha$.

Proof To show completeness it is sufficient to show that α is falsifiable for any non-theorem α. Letting Γ be some maximal CO-consistent set (MCS) which contains $\neg\alpha$, we will construct a model $M = \langle W, R, \varphi\rangle$ which falsifies α. This technique is employed in (Humberstone 1983). The model is constructed with W consisting of MCS's and two relations R and $\overline{R}$ over W, where $\overline{R}$ is intended to represent the complement of R. Ultimately, R and $\overline{R}$ will be mutually exclusive and exhaustive on $W \times W$.

The construction proceeds as follows: We start at stage 0 by adding Γ to W, so that $W = \{\Gamma\}$ and $R = \overline{R} = \emptyset$. At each following stage i, for each set Λ added to W at stage $i-1$ we do the following: (a) For each formula $\Diamond\beta \in \Lambda$ add a MCS Λ' where $\{\beta\} \cup \{\gamma : \Box\gamma \in \Lambda\} \subseteq \Lambda'$, and add $\langle \Lambda, \Lambda' \rangle$ to R; and (b) similarly, for each formula $\overleftarrow{\Diamond}\beta \in \Lambda$ add a MCS Λ' where $\{\beta\} \cup \{\gamma : \overleftarrow{\Box}\gamma \in \Lambda\} \subseteq \Lambda'$, and add $\langle \Lambda, \Lambda' \rangle$ to $\overline{R}$. That such MCS's exist is claimed without proof (see (Humberstone 1983)). Now let M be the totality of this (typically infinite) construction. Evaluating the truth conditions of $\overleftarrow{\Box}$ with respect to $\overline{R}$ (as if $\overline{R}$ were the complement of R), we can show the following, assuming $\varphi(A) = \{w : A \in w\}$ for atomic A.

Lemma B.4 *$M \models_w \beta$ iff $\beta \in w$.*

Proof We proceed by induction on the structure of β. For atomic β, this follows by the definition of φ. Assuming this for α and β, clearly it holds for both $\neg\alpha$ and $\alpha \supset \beta$ by standard properties of maximal consistent sets. Now suppose $\Box\beta \in w$. By the construction of M, for all wRv, $M \models_v \beta$, therefore $M \models_w \Box\beta$. If $\Box\beta \notin w$, then $\Diamond\neg\beta \in w$. By the construction of M, there is some wRv such that $M \models_v \neg\beta$, therefore $M \not\models_w \Box\beta$. The same argument holds for $\overleftarrow{\Box}\beta$, assuming $\overline{R}$ to be the complement of R.

■

Now we have a "structure" which falsifies α, as $\neg\alpha \in \Gamma$ and by the above, $M \models_\Gamma \neg\alpha$. However, M is not a CO-model, since R is neither reflexive, transitive nor strongly-connected, and $\overline{R}$ is not the complement of R. We now show that R and $\overline{R}$ can be extended in such a way that R does possess the desired properties and $\overline{R}$ is the complement of R, while not changing the fact that $M \models_w \beta$ iff $\beta \in w$.

Suppose that $\langle w, v \rangle \notin R$ and $\langle w, v \rangle \notin \overline{R}$, and that it cannot be "consistently" added to either of R or $\overline{R}$. Then there must be some $\Box\beta \in w$, $\beta \notin v$ and some $\overleftarrow{\Box}\gamma \in w$, $\gamma \notin v$. Both w and v must be some finite "distance" away from out starting point Γ, say m and n "steps", respectively. Following the "path" which lead to the addition of w to W, we have $M \models_\Gamma D_1(\Box\beta \wedge \overleftarrow{\Box}\gamma)$ where D_1 is a string of m $\Diamond$'s and $\overleftarrow{\Diamond}$'s (depending on how w was added). Similarly, $M \models_\Gamma D_2(\neg\beta \wedge \neg\gamma)$ where D_2 is the string of n $\Diamond$'s and $\overleftarrow{\Diamond}$'s corresponding to how w was added. But this sentence is equivalent to $\neg B_2(\beta \vee \gamma)$, where B_2 is formed by replacing $\Diamond$ and $\overleftarrow{\Diamond}$ with $\Box$ and $\overleftarrow{\Box}$ (respectively) in D_2. This means both $D_1(\Box\beta \wedge \overleftarrow{\Box}\gamma) \in \Gamma$ and $\neg B_2(\beta \vee \gamma) \in \Gamma$, contradicting the Humberstone schema. Since Γ is consistent, $\langle w, v \rangle$ can be added to either R or $\overline{R}$ without affecting the truth of formulae at any world in W, and hence R and $\overline{R}$ can be extended to complement one another, making valuation of $\overleftarrow{\Box}$ with respect to $\overline{R}$ the same as valuation with respect to the standard truth conditions.

We can ensure that R is reflexive, as well. Adding wRw affects the truth of some sentence only if there is some β such that $\Box\beta \in w$ and $\beta \notin w$; but this contradicts the axiom **T** and the fact that w is a MCS.

For transitivity, suppose wRv and vRt. Adding wRt can only affect truth if there is some $\Box\beta \in w$ and $\beta \notin t$. Since $\Box\beta \in w$, by axiom 4, $\Box\Box\beta \in w$. This means $\Box\beta \in v$ and $\beta \in t$, contradicting the assumption.

For total-connectedness, suppose $w\overline{R}v$. Adding vRw can affect truth only if there is some $\Box\beta \in v$ and $\beta \notin w$. If $\beta \notin w$, then $\neg\beta \in w$ and by axiom **S**, $\overleftarrow{\Box}\Diamond\neg\beta \in w$. Now since $w\overline{R}v$, $\Diamond\neg\beta \in v$, and $\Box\beta \notin v$, contradicting the original assumption.

It is clear that there may be some interaction during these "steps" whereby certain pairs of worlds are moved from the set $\overline{R}$ to R; but, clearly nothing in principle stops one from constructing a suitable model with the appropriate constraints being fulfilled by the relations. In fact, if we insist that R be completed maximally before we complete $\overline{R}$, there need not be any interaction. For instance, at the step where we decide to add each pair of worlds to R or $\overline{R}$, we can consider the union of the family of all possible relations R on $W \times W$ that respect on restrictions on accessibilty; we take this set to be R and let $\overline{R}$ then be $W \times W - R$. Hence, we can construct a CO-model which falsifies the non-theorem α.

■

Theorem 5.4 *The system* CO *is characterized by the class of CO-models; that is,* $\vdash_{CO} \alpha$ *iff* $\models_{CO} \alpha$.

Proof This follows immediately from Lemmas B.2 and B.3.

■

Corollary 5.5 *The logic KO (Humberstone's (1983)* K^2 *+ (*)) has the following axiomatic basis.*

K $\Box(A \supset B) \supset (\Box A \supset \Box B)$

K′ $\overline{\Box}(A \supset B) \supset (\overline{\Box} A \supset \overline{\Box} B)$

H $\overset{\leftrightarrow}{\Diamond}(\Box A \wedge \overline{\Box} B) \supset \overset{\leftrightarrow}{\Box}(A \vee B)$

H1 $(\Box A \wedge \overline{\Box} B) \supset (A \vee B)$

H2 $\overset{\leftrightarrow}{\Diamond}(\Box A \wedge \overline{\Box} B) \supset (A \vee B)$

H3 $(\Box A \wedge \overline{\Box} B) \supset \overset{\leftrightarrow}{\Box}(A \vee B)$

Nes *From* A *infer* $\overset{\leftrightarrow}{\Box} A$.

MP *From* $A \supset B$ *and* A *infer* B.

Proof That each of these rules is sound for KO-models is easy to verify. To show completeness, we show that each of the (infinite set of) axioms due to Humberstone is derivable from these.

Humberstone (1983) axiomatizes KO with the axioms shown here except with the infinite set of schemata **H*** replacing **H**, **H1**, **H2** and **H3**. We showed **H*** to be derivable in CO, but the only axiom we used distinct from these was **T**. However, there **T** was only used in the derivation of **T***, $\overset{\leftrightarrow}{\Box} A \supset A$, which was itself used to derive **H***. That **T*** is derivable from these axioms is evident if we consider an instance of **H1**: $(\Box A \wedge \overline{\Box} A) \supset (A \vee A)$.

■

Theorem 5.7 *The system* CO* *is characterized by the class of CO*-models; that is,* $\vdash_{CO*} \alpha$ *iff* $\models_{CO*} \alpha$.

Proof Given the soundness of the axioms of CO, we need only show that axiom **NB** is sound on the class of CO*-models. Let $M \models_w \overleftarrow{\Box}\alpha$ for some CO*-model M. We assume α is a propositional, falsifiable sentence. Since this is so, there is some world v which satisfies $\neg\alpha$. Since $M \models_w \overleftarrow{\Box}\alpha$, v must be accessible to w, in which case $M \models_w \neg\Box\alpha$.

We now turn our attention to completeness. Let w be any maximal consistent set of propositional sentences (which can be viewed as a propositional valuation, or possible world). For any (CO*-) MCS Γ denote by Γ^+ the set $\{\gamma : \Box\gamma \in \Gamma\}$, and by Γ^- the set $\{\gamma : \overleftarrow{\Box}\gamma \in \Gamma\}$. We will show that either

1. There is some β^+ such that $\Diamond\beta^+ \in \Gamma$ and $w \cup \{\beta^+\} \cup \Gamma^+$ is consistent; or
2. There is some β^- such that $\overleftarrow{\Diamond}\beta^- \in \Gamma$ and $w \cup \{\beta^-\} \cup \Gamma^-$ is consistent.

Suppose that the first condition is false. Then for each such β^+, some finite subset of this set is inconsistent, say $w_1 \cup \{\beta^+\} \cup \Gamma_1^+$. This mean (using the set to denote the conjunction of its elements) that $\beta^+ \wedge \Gamma_1^+ \supset \neg w_1$ is derivable. Now either $\Box\beta^+ \in \Gamma$ or $\Diamond\neg\beta^+ \in \Gamma$. In the latter case, by the fact that condition (1) is false, we can also show that some $\neg\beta^+ \wedge \Gamma_2^+ \supset \neg w_2$ is derivable (using the same argument). From the fact that both $\Box(\beta^+ \vee \neg\beta^+)$ and $\Box(\Gamma_1^+ \wedge \Gamma_2^+)$ are both in Γ, we know that $\Box\neg(w_1 \wedge w_2)$ is in Γ. In the case that $\Box\beta^+ \in \Gamma$ we can also derive this fact, since together with $\Box\Gamma_1^+$, this implies $\Box\neg w_1$. So whenever condition (1) is false, $\Box\neg(w_1 \wedge w_2) \in \Gamma$.

An entirely analogous argument shows that if the second condition is false then for some finite subsets of w, $\overleftarrow{\Box}\neg(w_3 \wedge w_4) \in \Gamma$. However, this implies both $\Box\neg(w_1 \wedge w_2 \wedge w_3 \wedge w_4)$ and $\overleftarrow{\Box}\neg(w_1 \wedge w_2 \wedge w_3 \wedge w_4)$ are in Γ. But $\neg(w_1 \wedge w_2 \wedge w_3 \wedge w_4)$ is falsifiable and this contradicts the fact that Γ is consistent (by axiom **NB**). So one of the two conditions must hold for any propositional MCS w.

To show completeness of CO*, we note that for any MCS Γ we can perform the construction associated with the previous theorem in such a way that every propositional valuation is represented as a possible world in the falsifying model. At stage one of the construction, we just have to add enough MCS's to cover every propositional MCS. That this can be performed without violating the truth values given to sentences in the model follows immediately from the above disjunctive condition.

■

Lemma B.5 (Soundness) *If* $\vdash_{CT4O} \alpha$ *then* $\models_{CT4O} \alpha$.

Proof The proof is exactly like that of Lemma B.2.

■

Lemma B.6 *If* **(Completeness)** $\models_{CT4O} \alpha$ *then* $\vdash_{CT4O} \alpha$.

Proof The proof is exactly like that of Lemma B.3.

■

Theorem 5.9 *The system* CT4O *is characterized by the class of CT4O-models; that is,* $\vdash_{CT4O} \alpha$ *iff* $\models_{CT4O} \alpha$.

Proof This follows immediately from Lemmas B.5 and B.6.

■

Theorem 5.10 *The system* CT4O* *is characterized by the class of CT4O*-models; that is,* $\vdash_{CT4O*} \alpha$ *iff* $\models_{CT4O*} \alpha$.

Proof The proof is entirely analagous to that of 5.7.

■

Proposition 5.11 *Let* M *be a CO-model. Then* $M \models A \Rightarrow B$ *iff* $M \models_w A \Rightarrow B$ *for some* w .

Proof The proof lies in the observation that $M \models_w \boxdot\alpha$ iff $M \models \boxdot\alpha$ and $M \models_w \diamondsuit\!\!\!\!-\alpha$ iff $M \models \diamondsuit\!\!\!\!-\alpha$. Given the definition of $\Rightarrow$ as the disjunction of two such statements, the proposition follows.

■

Proposition 5.12 *Let* $M = \langle W, R, \varphi\rangle$ *be a CO-model. Then for any* w, $M \models_w A > B$ *iff* $M \models A > B$. *If* $M \models A > B$ *then there exists a cluster* C *in* W *such that* $\|A \wedge B\| \subseteq C$ *and no A-world is strictly more normal than* C. *Furthermore,* $\|A \wedge B\|$ *must be nonempty.*

Proof That $M \models_w A > B$ iff $M \models A > B$ is immediate given the definition of $A > B$ as the conjunction of two statements with main connectives $\boxdot$ and $\diamondsuit\!\!\!\!-$.

If there exist no $A \wedge B$-worlds then $A > B$ cannot hold, as it entails $\diamondsuit\!\!\!\!-A$ and $\boxdot(A \supset \Diamond(A \wedge B))$. So $\|A \wedge B\|$ is nonempty. Suppose there is no cluster C such that $\|A \wedge B\| \subseteq C$. Then there must be two $A \wedge B$-worlds, say w and v, such that wRv and not vRw. This implies $M \models_w A \wedge \neg\boxdot(A \supset \neg B)$. However, from $M \models A > B$ we know that $M \models_w A \supset \boxdot(A \supset \neg B)$, contradicting the assumption. So such a cluster C exists. By a similar argument, there can be no A-world w such that C is inaccessible to w.

■

Proposition 5.13 *Let* $M = \langle W, R, \varphi\rangle$ *be a CO-model. If* $M \models A > B$ *and* $M \models A \Rightarrow B$ *then there exists a cluster* C *in* W *such that* $\|A \wedge B\| \cup \mathcal{N} = C$, *where each world in* $\mathcal{N}$ *satisfies* $\neg A$, *and no A-world is strictly more normal than* C.

Proof By Proposition 5.12, there exists a cluster such that $\|A \wedge B\| \subseteq C$ and no A-world is strictly more normal than C. Since $\|A \wedge B\| \neq \emptyset$, the truth of $A \Rightarrow B$ ensures that $\diamondsuit\!\!\!\!-(A \wedge \Box(A \supset B))$ holds. So for some w, $M \models_w A \wedge \Box(A \supset B)$ holds. But w cannot be strictly more normal than C, so C is accessible to w and we have for $v \in C$, $M \models_v A \supset B$.

■

Proposition 5.14 $A \Rightarrow B, A > B \vdash_{CO} \overset{\infty}{\Diamond}(A \wedge B \wedge \alpha) \supset (A \wedge \alpha \Rightarrow B)$.

Proof This follows immediately from Proposition 5.13.

■

Proposition 5.15 *Let $A \wedge B \wedge \alpha$ be propositionally satisfiable. Then*

$$A \Rightarrow B, A > B \vdash_{CO*} A \wedge \alpha \Rightarrow B$$

Proof This follows immediately from Proposition 5.14 and the fact that $\models_{CO*} \overset{\infty}{\Diamond}\alpha$ for any satisfiable propositional α.

■

Corollary 5.16 *Z_T is a CO*-model.*

Proof That the relation R of Z_T is transitive and totally-connected is obvious given the range of the Z-ranking, the natural numbers.

■

Corollary 5.17 $Z_T \models_{CO*} T$.

Proof A property of 1-entailment is that $\alpha \vdash_1 \beta$ if $\alpha \rightarrow \beta \in T$. Thus for each such rule in T, we have

$$Z(\alpha \wedge \beta) < Z(\alpha \wedge \neg\beta).$$

So there is some w such that $Z_T \models_w \alpha \wedge \beta$ and for no wRv is it the case that $Z_T \models_w \alpha \wedge \neg\beta$. Thus $Z_T \models_{CO*} T$.

■

Corollary 5.18 *$Z_T \models_{CO*} \alpha \Rightarrow \beta$ iff $\alpha \vdash_1 \beta$ whenever α is satisfiable.*

Proof Suppose $\alpha \vdash_1 \beta$. Then we have

$$Z(\alpha \wedge \beta) < Z(\alpha \wedge \neg\beta).$$

As in the proof of Proposition 5.17, this means $Z_T \models_{CO*} \alpha \Rightarrow \beta$.

Suppose $Z_T \models_{CO*} \alpha \Rightarrow \beta$. Then we have some world w such that $Z_T \models_w \alpha \wedge \beta$ and for no wRv is it the case that $Z_T \models_w \alpha \wedge \neg\beta$. By the definition of R, clearly

$$Z(\alpha \wedge \beta) < Z(\alpha \wedge \neg\beta),$$

and $\alpha \vdash_1 \beta$.

■

Lemma 5.19 *If $M \models_{CO*} T$, then $Z_T \leq M$.*

Proof Let $M = \langle W, R\rangle$ be some model of the finite ruleset T, where T can be partitioned as $T = T_0 \cup \cdots \cup T_n$. It is easy to verify that for each rule $r \in T$ there must be some world $w_r \in W$ which verifies r and for all w such that $w_r R w$, w does not falsify r (this follows from the truth conditions of $\Rightarrow$, and the fact that no conditional in T can be satisfied "vacuously" since T is ε-consistent and M is a CO*-model). For any set of worlds S, define

$$\min S = \{w \in S : \forall v \in S, vRw\}.$$

For each $k \leq n$ we define $\min_k = \min\{w_r : r \in T_k\}$. In other words, $\min_k$ is the subset of verifying worlds for rules of rank k which is "minimal" (or most normal) in R. (That we have some latitude in choosing w_r for each r is of no consequence.) Somewhat loosely, we will use $\min_k$ to refer to some arbitrary element of that set. Notice that if $\min_k Rw$ then w satisfies all rules r of rank k.

Lemma B.7 *For each $j \leq n$, $\min\{\min_k : n \geq k \geq j\} = \min_j$. In other words, $\min_{k+1} R \min_k$ and not $\min_k R \min_{k+1}$ for all $k < n$.*

Proof Suppose $\min_l \subseteq \min\{\min_k : k \geq j\}$, for some $l > j$. Then $\min_k R \min_l$ for all $k \geq j$. By definition of $\min_k$, $\min_l$ falsifies no rules r of rank greater than j. Since $\min_l$ verifies some rule of rank l, it must be that it falsifies some rule of rank $l-1$. That is $\min_l$ falsifies some rule of rank $\geq j$. This contradiction implies that $\min_j = \min\{\min_k : k \geq j\}$.

■

As a corollary to this lemma, we have that $Z(\min_j) = j$, since for all $k \geq j$, $\min_k R \min_j$ (which means that $\min_j$ falsifies no rules of rank $\geq j$), and it falsifies some rule of rank $j-1$.

Now let $w \in W$ and suppose that $\{v : wRv\} \not\subseteq \{v : wR_Z v\}$. Then there exists a $v \in W$ such that such that wRv, but not $wR_Z v$; hence, $Z(v) > Z(w) = k$. So v falsifies some rule r such that $Z(r) = k' \geq k$. So clearly $vR \min_{k'}$ and not $\min_{k'} Rv$, which entails $vR \min_k$ and not $\min_k Rv$ (by the above lemma). This means $wR \min_k$ and not $\min_k Rw$. By the corollary and the definition of Z_T, $wR_Z \min_k$ and $\min_k R_Z w$, so $N(w, R_Z, R)$. Hence, $Z_T \leq M$.

■

Lemma 5.20 *If $M \models_{CO*} T$ and $M \leq Z_T$, then $M = Z_T$.*

Proof As an auxiliary lemma consider

Lemma B.8 *If there is some $w \in W$ such that $Z(w) = k$ and not $\min_k Rw$, then $M \not\leq Z_T$.*

Proof We will show this by induction on k. Suppose $k = 0$, Clearly, not $N(\min_k, R, R_Z)$ as $\min_k Rw$ implies $Z(w) = 0$, and $\min_k R_Z w$ and $wR_Z \min_k$. But if there is some $Z(w) = 0$ and not $\min_k Rw$ then $\{w : \min_k Rw\} \not\subseteq \{w : \min_k R_Z\}$. So $M \not\leq Z_T$.

Assume this holds for all $n \leq k-1$. Suppose there is some $Z(w) = k$ and not $\min_k Rw$. Then $\{w : \min_k Rw\} \not\subseteq \{w : \min_k R_Z w\}$. If $N(\min_k, R, R_Z)$ then there is some v such that $Z(v) = j < k$ and $vR\min_k$. But if $M \leq Z_T$, by the inductive hypothesis, for some $j < k$, $\min_j Rv$ and thus $\min_j R\min_k$, contradicting a previous auxiliary lemma. So not $N(\min_k, R, R_Z)$ and hence $M \not\leq Z_T$.

■

Let $M = \langle W, R\rangle$ such that $M \models_{CO*} T$. Suppose $M \leq Z_T$ and $M \neq Z_T$. If not $N(w, R, R_Z)$ for all $w \in W$, then by definition of $\leq$, $M = Z_T$. So assume there exists some $w \in W$ such that $N(w, R, R_Z)$ holds. Then for some v, vRw, wRv and not vR_Zw, which implies $j = Z(v) < Z(w) = k$. Now $v \notin \min_j$ (with respect to M) since vRw and $Z(w) > j$ (see Lemma 5.19). So not $\min_j Rv$ and, by the above lemma, $M \not\leq Z_T$.

■

Theorem 5.21 *$T \models_{\leq} \alpha \Rightarrow \beta$ iff $\alpha \vdash_1 \beta$ with respect to T.*

Proof By Lemmas 5.19 and 5.20, the unique $\leq$-minimal model of T is Z_T (modulo a given set of possible worlds). Clearly, $T \models_{\leq} \alpha \Rightarrow \beta$ iff β is true at the "minimal" α-worlds in Z_T. But clearly, this is the same criterion for evaluating the truth of $\alpha \vdash_1 \beta$. (Here we assume that α is a consistent propositional formula. If not, we can define $\alpha \vdash_1 \beta$ to be "trivially" true.)

■

Lemma 5.22 *$Clos(T)$ is consistent iff T is, and is "categorical" in the sense that there is a unique CO*-model which satisfies it, namely Z_T.*[1]

Proof We will show that any model of $Clos(T)$ must be equivalent to Z_T (modulo some fixed set of possible worlds). It is easy to verify that in fact Z_T satisfies $Clos(T)$ and hence that $Clos(T)$ is consistent iff T is. We proceed by induction on the rank of possible worlds to show that if $M = \langle W, R\rangle$ satisfies $Clos(T)$, then vRw iff $Z(v) \geq Z(w)$.

Let $Z(w) = 0$. Suppose $Z(v) = 0$. By the sentence $\neg R^{\wedge}_{-1} > R^{\wedge}_0$ it must be the case that all worlds which falsify no rule are mutually accessible (i.e. all worlds of rank 0). Hence vRw. Suppose $Z(v) > 0$. Then v falsifies some rule $Z(r) \geq 0$. If wRv then for every u such that $Z(u) = 0$, uRv. As v falsifies some rule, M cannot satisfy T (nor $Clos(T)$). So it must be that not wRv. Since R must be totally-connected, vRw.

Let this property hold for all $s \in W$ such that $Z(s) < i$. That is $Z(t) \geq Z(s)$ iff tRs, and $Z(t) > Z(s)$ iff not sRt. Let $Z(w) = i$. Since $\neg R^{\wedge}_{i-1} > R^{\wedge}_i$ is in $Clos(T)$, M must be such that all worlds of rank i are mutually accessible. So if $Z(v) = i$ then vRw. Suppose $Z(v) > i$. We must proceed by induction again to show that not wRv. For the case of $Z(v) = i+1$, it must be that v falsifies some rule $Z(r) = i$. But in order for M to satisfy r there must be

[1] Again, it categorical if we ignore "duplicate" worlds (associated with the same valuation), which contribute nothing to the truth or falsity of formulae in Z_T.

some world of rank i which verifies r and sees no world which falsifies r. If wRv, then uRv for all worlds of rank i (by the original inductive hypothesis), and M cannot satisfy r. So it can't be the case that wRv. Now assume this is true for all worlds x such that $j > Z(x) > i$, and that $Z(v) = j$. v must falsify some rule $Z(r) = j - 1$. In order for M to satisfy r there must be some world of rank $j - 1$ which verifies r and sees now world which falsifies r. If wRv, then by the inductive hypothesis, uRv for all $Z(u) = j - 1$ (since uRw), and M cannot satisfy R. So it can't be that wRv. Since R is totally-connected, vRw.

■

Theorem 5.23 *$Cl(T) \models_{CO*} \alpha \Rightarrow \beta$ iff $T \models_{\leq} \alpha \Rightarrow \beta$.*

Proof By Lemma 5.22, $Clos(T) \models_{CO*} \alpha \Rightarrow \beta$ iff $Z_T \models_{CO*} \alpha \Rightarrow \beta$. But this is precisely the same condition for evaluating $T \models_{\leq} \alpha \Rightarrow \beta$.

■

Theorem 5.27 *Let M be a CO*-model. The c-relevance relation determined by M satisfies (I0), (I1), (I3), (I4) and (I5).*

Proof We let $M = \langle W, R, \varphi \rangle$ be a CO*-model, with worlds w, v, etc.

(I0) Suppose $\vdash e \supset (p \equiv q)$. Then $\vdash (p \wedge e) \equiv (q \wedge e)$; and

$$\vdash_{CO*} (p \wedge e \Rightarrow r) \equiv (q \wedge e \Rightarrow r),$$

$$\vdash_{CO*} (p \wedge e \Rightarrow \neg r) \equiv (q \wedge e \Rightarrow \neg r)$$

both hold. So $p\mathcal{R}_e r$ iff $q\mathcal{R}_e r$.

(I1) $p\mathcal{R}_e r$ iff not $p\mathcal{I}_e r$ holds by definition of $\mathcal{R}_e$ and $\mathcal{I}_e$.

(I3) A simple derivation in CO* shows that $e \Rightarrow r$ is equivalent to $e \wedge \top \Rightarrow r$. So $\top \mathcal{I}_e r$.

(I4) If $M \models e \Rightarrow r$ then $M \models \neg r \wedge e \not\Rightarrow r$ (since r is contingent on e). If $M \models e \not\Rightarrow r$ (since r is contingent on e) then $M \models r \wedge e \Rightarrow r$. In either case, there is some q such that $q\mathcal{R}_e r$.

(I5) If $p\mathcal{R}_e r$ and $\not\vdash \neg(p \wedge q \wedge e)$ then, by the definition of relevance, $(p \wedge q)\mathcal{R}_e r$, since $\vdash p \wedge q \supset p$.

■

Proposition 5.28 *C-relevance is nontrivial, in general. That is, there are CO*-models M and sentences p and r, both contingent on e and pairwise contingent, such that $p\mathcal{I}_e r$ where $\mathcal{I}_e$ is determined by M.*

Proof This is simple to verify by considering any CO*-model with a most normal set of e-worlds, each satisfying r, and having at least one p-world in this set of normal e-worlds. Then $e \Rightarrow r$ holds in this model, and for any sentence q implied by p, $q \wedge e \Rightarrow r$ holds as well.

■

Proposition 5.29 *Let M be a CO*-model and $\mathcal{I}_e$ and $\mathcal{R}_e$ the c-relevance relation determined by M. If $M \models e > r$ and $M \models e \Rightarrow r$ then $p\mathcal{I}_e r$ for all p such that $\not\vdash \neg(p \wedge e \wedge r)$.*

Proof This follows at once from Proposition 5.15.

■

Appendix C

Proofs of Theorems: Chapter 6

In this appendix we present proofs of various propositions, lemmas, theorems and corollaries found in Chapter 6.

Theorem 6.1 *Let K be finitely specified by KB and $M = \langle W, R, \varphi \rangle$ be a CT4O-model such that for some $v, w \in W$, $M \models_w O(KB)$ and $M \models_v \boxminus\Diamond KB$. Then M is a preorder revision model for K.*

Proof Let $w \in W$ be a KB-world and W' be the connected fragment of M containing w; that is,

$$W' = \{v \in W : wRv \text{ or } vRw\}.$$

Since M satisfies $O(KB)$, we have $M \models_w \Box KB$. This means $M \models_v KB$ for each v such that wRv. Thus, every minimal world in W' satisfies KB. Furthermore, $M \models_w \boxminus\neg KB$, so no KB-world can be inaccessible to w. Hence, every KB-world is in W' and is minimal in W.

Since M satisfies $\boxminus\Diamond KB$, for every world v we have $M \models_v \Diamond KB$. This means vRw for each $v \in W$ so that $W = W'$. Hence all and only KB-worlds are minimal in M and M is a preorder revision model.

■

Theorem 6.2 *Let K be finitely specified by KB and $M = \langle W, R, \varphi \rangle$ be a CO-model such that for some $v, w \in W$, $M \models_w O(KB)$ and $M \models_v \boxminus\Diamond KB$. Then M is a total order revision model for K.*

Proof The proof is exactly the same as that of Theorem 6.1.

■

Corollary 6.3 *Let $M = \langle W, R, \varphi \rangle$ be a CO-model such that $M \models_w O(KB)$ for some $w \in W$. Then M is a total order revision model for K.*

Proof The proof is exactly the same as that of Theorem 6.1, except that the condition $\boxminus\Diamond KB$ is not required to show that $W = W'$, since this is true of any CO-model (which must be totally connected).

■

Proposition 6.4 $M \models_{CO} \boxplus\neg A \vee \diamondsuit\!\!\!\!\diamondsuit(A \wedge \Box(A \supset B))$ *iff*
$M \models_{CO} \boxplus(\Box\neg A \vee \Diamond(A \wedge \Box(A \supset B)))$.

Proof Let $M = \langle W, R, \varphi\rangle$ be a CO-model. Suppose $M \models \boxplus(\Box\neg A \vee \Diamond(A \wedge \Box(A \supset B)))$. Then at every $v \in W$

$$M \models_v \Box\neg A \vee \Diamond(A \wedge \Box(A \supset B)).$$

(a) If for every such v it is the case that $M \models_v \Box\neg A$, then $M \models \boxplus\neg A$.

(b) If for some such v it is the case that

$$M \models_v \Diamond(A \wedge \Box(A \supset B)),$$

then $M \models \diamondsuit\!\!\!\!\diamondsuit(A \wedge \Box(A \supset B))$.

So

$$M \models \boxplus\neg A \vee \diamondsuit\!\!\!\!\diamondsuit(A \wedge \Box(A \supset B)),$$

and

$$\vdash_{CO} \Box(\Box\neg A \vee \Diamond(A \wedge \Box(A \supset B))) \supset \Box\neg A \vee \Diamond(A \wedge \Box(A \supset B)).$$

Suppose $M \models \boxplus\neg A \vee \diamondsuit\!\!\!\!\diamondsuit(A \wedge \Box(A \supset B))$.

(a) If $M \models \boxplus\neg A$ then $M \models \boxplus\Box\neg A$.

(b) Suppose $M \models \diamondsuit\!\!\!\!\diamondsuit(A \wedge \Box(A \supset B))$. Then for some $u \in W$ we have

$$M \models_u A \wedge \Box(A \supset B).$$

For each $v \in W$ either vRu or uRv, as R is connected. If vRu then clearly

$$M \models_v \Diamond(A \wedge \Box(A \supset B)).$$

If uRv then either $M \models_v \Box\neg A$ or (since $M \models_u \Box(A \supset B)$)

$$M \models_v \Diamond(A \wedge \Box(A \supset B)).$$

Thus for each $v \in W$, we have

$$M \models_v \Box\neg A \vee \Diamond(A \wedge \Box(A \supset B));$$

thus

$$M \models \boxplus(\Box\neg A \vee \Diamond(A \wedge \Box(A \supset B))).$$

Hence

$$\vdash_{CO} \boxplus\neg A \vee \diamondsuit\!\!\!\!\diamondsuit(A \wedge \Box(A \supset B)) \supset \boxplus(\Box\neg A \vee \Diamond(A \wedge \Box(A \supset B))).$$

■

Proposition C.1 *Let M be a CT4O-model. Then $M \models_w A \overset{\text{KB}}{\longrightarrow} B$ for some w iff $M \models A \overset{\text{KB}}{\longrightarrow} B$.*

Proof This is a simple consequence of the definition of $\overset{\text{KB}}{\longrightarrow}$ and the truth conditions for $\overset{\text{KB}}{\Box}$ and $\overset{\text{KB}}{\Diamond}$.

■

Proposition 6.5 *The following are derived theorems and inference rules in* CT4O *(assuming as a premise $O(KB)$ wherever KB is mentioned).*

RCM *From $B \supset C$ infer $(A \overset{\text{KB}}{\longrightarrow} B) \supset (A \overset{\text{KB}}{\longrightarrow} C)$*

And $(A \overset{\text{KB}}{\longrightarrow} B) \wedge (A \overset{\text{KB}}{\longrightarrow} C) \supset (A \overset{\text{KB}}{\longrightarrow} B \wedge C)$

ID $A \overset{\text{KB}}{\longrightarrow} A$

KC $(A \overset{\text{KB}}{\longrightarrow} B) \supset (KB \wedge A \supset B)$

CK $\overset{\text{KB}}{\Diamond}(KB \wedge A) \supset (\overset{\text{KB}}{\Box}(KB \wedge A \supset B) \supset (A \overset{\text{KB}}{\longrightarrow} B))$

Cons $\overset{\text{KB}}{\Box}\neg A \equiv (A \overset{\text{KB}}{\longrightarrow} \bot)$

LLE *From $A \equiv B$ infer $(A \overset{\text{KB}}{\longrightarrow} C) \supset (B \overset{\text{KB}}{\longrightarrow} C)$*

KCI $(A \wedge B \overset{\text{KB}}{\longrightarrow} C) \supset (A \overset{\text{KB}}{\longrightarrow} (B \supset C))$

Proof We demonstrate the derivability of these sentences with semantic arguments and appeal to the characterization result Theorem 5.9. We assume $M = \langle W, R, \varphi \rangle$ is a CT4O-model with various elements v, w, etc. in W. We freely use Proposition C.1.

RCM: Suppose $B \supset C$ is valid and $M \models A \overset{\text{KB}}{\longrightarrow} B$. By definition of $\overset{\text{KB}}{\longrightarrow}$, at every world $\Box\neg A$ holds or $\Diamond(A \wedge \Box(A \supset B)$ holds. At any world where the latter holds (since $\vdash_{CT4O} B \supset C$) it must be that $\Diamond(A \wedge \Box(A \supset C)$ holds as well. Thus $M \models A \overset{\text{KB}}{\longrightarrow} C$.

And: Suppose $M \models (A \overset{\text{KB}}{\longrightarrow} B) \wedge (A \overset{\text{KB}}{\longrightarrow} C)$. Then at every world $\Box\neg A$ holds, or $\Diamond(A \wedge \Box(A \supset B)$ and $\Diamond(A \wedge \Box(A \supset C)$ hold. At any world v in the latter case, there must be some accessible vRu such that $M \models_u A \wedge B$ and $M \models_u \Box(A \supset B)$. By the fact that $M \not\models_u \Box\neg A$, we have $M \models_u \Diamond(A \wedge \Box(A \supset C)$. This means

$$M \models_u \Diamond(A \wedge \Box(A \supset B \wedge C),$$

which implies

$$M \models_v \Diamond(A \wedge \Box(A \supset B \wedge C).$$

Thus $M \models A \overset{\text{KB}}{\longrightarrow} B \wedge C$.

ID: That $A \overset{\text{KB}}{\longrightarrow} A$ is valid follows from the validity of $\Box(A \supset A)$ and $\Box\neg A \vee \Diamond A$ (using e.g. **Nec** and the definition of $\Diamond$).

KC: Assume $M \models O(KB)$ and $M \models A \overset{\text{KB}}{\longrightarrow} B$. Suppose $M \models_w KB \wedge A$. Since $M \models O(KB)$, by definition of $O(KB)$ every world in M can see w. If $M \not\models_w B$ then at no world can $\Diamond(A \wedge \Box(A \supset B)$ hold. Since $\Box\neg A$ is false at all worlds, the fact that $M \models A \overset{\text{KB}}{\longrightarrow} B$ is contradicted. Thus $M \models_w B$.

CK: Assume $M \models O(KB)$, $M \models \overset{\infty}{\Diamond}(KB \wedge A)$ and $M \models \overset{\infty}{\Box}(KB \wedge A \supset B)$. Then there exists an $KB \wedge A$-world w such that all worlds see w. Furthermore, any A-world v which w sees is a $KB \wedge A$-world since $M \models O(KB)$; and, since $M \models \overset{\infty}{\Box}(KB \wedge A \supset B)$, for any such v we have $M \models_v B$. Thus every world sees a world (namely w) such that $M \models_w \Diamond(A \wedge \Box(A \supset B))$. Hence $M \models A \overset{\text{KB}}{\longrightarrow} B$.

Cons: That $\overset{\infty}{\Box}\neg A \equiv (A \overset{\text{KB}}{\longrightarrow} \bot)$ is valid follows immediately from the definition of $\overset{\text{KB}}{\longrightarrow}$.

LLE: Assume $A \equiv B$ is valid, and suppose $M \models A \overset{\text{KB}}{\longrightarrow} C$. If $\overset{\infty}{\Box}\neg A$ holds at some world w, $\overset{\infty}{\Box}\neg B$ also holds at w. If $\Diamond(A \wedge \Box(A \supset C)$ holds at some w, then so does $\Diamond(B \wedge \Box(B \supset C)$.

KCI: Suppose $M \models A \wedge B \overset{\text{KB}}{\longrightarrow} C$. Then, at any world w, either $\Box\neg(A \wedge B)$ holds or

$$\Diamond((A \wedge B) \wedge \Box(A \wedge B \supset C))$$

holds. In the first case either (a) $\Box\neg A$ holds; or (b) $\Box(A \supset \neg B)$ holds, which means $\Box(A \supset (B \supset C))$ does as well. In the second case, there is an accessible wRv such that $M \models_v A \wedge B$ and $M \models_v \Box(A \wedge B \supset C)$. This means $M \models_v A$ and $M \models_v \Box(A \supset (B \supset C)$. So

$$M \models_w \Diamond(A \wedge \Box(A \supset (B \supset C)).$$

Thus $M \models A \overset{\text{KB}}{\longrightarrow} (B \supset C)$.

■

Proposition 6.6 CKI *is derivable in CO.*

CKI $\neg(A \overset{\text{KB}}{\longrightarrow} \neg B) \supset ((A \overset{\text{KB}}{\longrightarrow} (B \supset C)) \supset (A \wedge B \overset{\text{KB}}{\longrightarrow} C))$

Proof The proof is demonstrated by semantic means, using the characterization result Theorem 5.4. Let $M = \langle W, R, \varphi \rangle$ be a CO-model with $w \in W$. Suppose

$$M \models \neg(A \overset{\text{KB}}{\longrightarrow} \neg B) \text{ and } M \models A \overset{\text{KB}}{\longrightarrow} (B \supset C).$$

By the first condition, using the simple definition of $\Rightarrow$ (see Proposition 6.4), we have $M \not\models \overset{\infty}{\Diamond}(A \wedge \Box(A \supset B))$. Thus $M \models \overset{\infty}{\Box}(\neg A \vee \Diamond(A \wedge B))$. By the second condition,

$$M \models \overset{\infty}{\Box}\neg A \vee \overset{\infty}{\Diamond}(A \wedge \Box(A \supset (B \supset C))).$$

If $\overset{\infty}{\Box}\neg(A \wedge B)$ holds, $A \wedge B \overset{\text{KB}}{\longrightarrow} C$ holds trivially. If not, then

$$M \models_w A \wedge \Box(A \supset (B \supset C))$$

for some $w \in W$. But by the first condition, $M \models_w \Diamond(A \wedge B)$. This means at some world v, where wRv,

$$M \models_v A \wedge B \wedge \Box(A \wedge B \supset C).$$

Thus $M \models \overset{\infty}{\Diamond}(A \wedge B \wedge \Box(A \wedge B \supset C))$; that is, $A \wedge B \overset{\text{KB}}{\longrightarrow} C$.

■

Theorem 6.7 *Let M be a full revision model and $*^M$ the revision function determined by M. Then $*^M$ satisfies postulates (R1) through (R8).*

Proof This follows almost immediately from Propositions 6.5 and 6.6.

■

Theorem 6.8 *Let $*$ be a revision function satisfying postulates (R1) through (R8). Then for any theory K there exists a full revision model M such that $K^*_A = K^{*^M}_A$ for all A.*

Proof The proof of the theorem will be similar to the proof of the representation result given by Grove (1988), and uses the observation that a system of spheres can be considered to be a totally preordered relation on possible worlds.

Let $*$ be an AGM revision operator. We use several well-known facts about such operators (see (Gärdenfors 1988)).

We will construct a full revision model $M = \langle W, R, \varphi \rangle$ which corresponds to $*$ for any given consistent K. M is given by:

1. $W = \{w : w \text{ is a maximal consistent set of } \mathbf{L}_{CPL}\}$. We use w to also denote the obvious propostional valuation associated with the set of formulae (so $w \models \alpha$ iff $\alpha \in w$; and $w \in \|S\|$ iff $S \subseteq w$ for any set of sentences S).
2. vRw iff there exists a sentence $A \in w \cap v$ such that $K^*_A \subseteq w$, or there is no consistent A such that $K^*_A \subseteq v$. We define $w \leq v$ iff vRw, and $w < v$ iff vRw and not wRv.
3. $\varphi(w, A) = 1$ iff $A \in \varphi$ for any atomic A.

We will show M is a K-revision model by demonstrating each of the necessary properties.

(a) R is reflexive: That $w \leq w$ for all $w \in W$ follows easily in either the case where $w \in \|K^*_A\|$ or not, for some consistent A.

(b) R is totally-ordered: Suppose $w \not\leq v$ and $v \not\leq w$. Then for some consistent A and B, $v \in \|K^*_A\|$ and $w \in \|K^*_B\|$. So $A \vee B \in v \cap w$, and (by (R7) and (R8)) we have $K^*_{A\vee B} = K^*_A$ or $K^*_{A\vee B} = K^*_B$ or $K^*_{A\vee B} = K^*_A \cap K^*_B$. Thus one of v or w is in $K^*_{A\vee B}$, contradicting the fact that neither $v \leq w$ or $w \leq v$.

(c) R is transitive: Suppose $u \leq v$ and $v \leq w$. Assume there are consistent B, C such that $B \in w$ and $v \in \|K^*_B\|$, $C \in v$ and $u \in \|K^*_C\|$ (otherwise $u \leq w$ holds trivially). Now $\neg C \notin K^*_B$, so by (R7) and (R8) we have $\neg C \notin K^*_{B\vee C}$. This means (by (R8)) that $K^*_{(B\vee C)\wedge C} = (K^*_{B\vee C})^+_C$. In other words, $K^*_C = (K^*_{B\vee C})^+_C$. Hence, $\|K^*_C\| \subseteq \|K^*_{B\vee C}\|$ and $u \in \|K^*_{B\vee C}\|$. But $B \vee C \in w \cap u$, so $u \leq w$.

(d) All K-worlds are minimal: Suppose $w \in \|K\|$. For any world v there exists some $A \in w \cap v$ such that $K \cup \{A\}$ is consistent (e.g. if A is a tautology). By (R3) and (R4), $K^*_A = K^+_A$. Since $w \in \|K^+_A\|$, we have $w \in \|K^*_A\|$. Then by definition of $\leq$, $w \leq v$.

(e) Only K-worlds are minimal: Suppose $w \in \|K\|$, while $v \notin \|K\|$. Choose some $A \in v \cap w$ (such an A exists, for example, any tautology). Since $A \in w$, we have $K \cup \{A\}$ is consistent, and by (R3) and (R4), $K^*_A = K^+_A$. But $v \notin \|K\|$ so $v \notin \|K^+_A\|$. Hence $v \notin \|K^*_A\|$ and $v \not\leq w$. $v \in \|K^*_A\|$.

(f) All truth valuations are in W: This is obvious given the definition of M.

Thus M is a K-revision model. It remains to be shown that the revision function determined by M is $*$. For inconsistent A this is clear, since $K^*_A = Cn(\bot)$ in both cases (for $*$ and $*^M$). So assume A is consistent. We must show

$$B \in K^*_A \text{ iff there is some } w \in W,\ M \models_w A \wedge \Box(A \supset B).$$

Lemma C.2 *If $v \leq w$, $w \in \|K^*_A\|$, and $v \models A$, then $w \leq v$.*

Proof Since $A \in w \cap v$, $w \in \|K^*_A\|$. The definition of $\leq$ ensures $v \leq v$.

■

Lemma C.3 *$w \in \|K^*_A\|$ iff $w \in \min(A)$, where*
$\min(A) = \{v \in W : v \models A$ and if $u < v$ then $u \not\models A\}$.

Proof Suppose $w \in \|K^*_A\|$ and $v \in \|A\|$. By Lemma C.2 $w \leq v$, hence $w \in \min(A)$. (Thus the existence of $\min(A)$ is demonstrated.)

Suppose $w \in \min(A)$ and $w \notin \|K^*_A\|$. Since A is consistent, there is some $v \in \|K^*_A\|$. By supposition, $w \leq v$, since $v \models A$. This means there is a $B \in w \cap v$ such that $w \in \|K^*_B\|$. But then we have

$$w \in \|(K^*_B)^+_A\| = \|K^*_{A \wedge B}\| = \|K^*_{B \wedge A}\| = \|(K^*_A)^+_B\| \subseteq \|K^*_A\|.$$

Thus $\|K^*_A\| = \min(A)$.

■

Suppose $B \in K^*_A$. Then for all $w \in \|K^*_A\|$, $w \models B$ and (by Lemma C.3) for all $w \in \min(A)$, $w \models B$. By definition of $\min(A)$,

$$M \models_w A \wedge \Box(A \supset B).$$

So $M \models A \overset{\text{KB}}{\longrightarrow} B$.

Suppose $M \models A \overset{\text{KB}}{\longrightarrow} B$. Then there is a v such that

$$M \models_v A \wedge \Box(A \supset B).$$

By definition of $\min(A)$, vRw for all $w \in \min(A)$. So for $w \in \|K^*_A\|$ we have $w \models A \supset B$; and then $B \in K^*_A$.

Thus, $K^*_A = K^{*M}_A$.

■

Proposition 6.9 *The following theorems are derivable in CO.*

And $(A \overset{\text{KB}}{\longrightarrow} B) \wedge (A \overset{\text{KB}}{\longrightarrow} C) \supset (A \overset{\text{KB}}{\longrightarrow} B \wedge C)$

Or $(A \overset{\text{KB}}{\longrightarrow} C) \wedge (B \overset{\text{KB}}{\longrightarrow} C) \supset (A \vee B \overset{\text{KB}}{\longrightarrow} C)$

RT $(A \overset{\text{KB}}{\longrightarrow} B) \supset ((A \wedge B \overset{\text{KB}}{\longrightarrow} C) \supset (A \overset{\text{KB}}{\longrightarrow} C))$

CM $(A \overset{\text{KB}}{\longrightarrow} B) \wedge (A \overset{\text{KB}}{\longrightarrow} C) \supset (A \wedge B \overset{\text{KB}}{\longrightarrow} C)$

RM $(A \overset{\text{KB}}{\longrightarrow} C) \wedge \neg(A \wedge B \overset{\text{KB}}{\longrightarrow} C) \supset A \overset{\text{KB}}{\longrightarrow} \neg B$

CV $\neg(A \overset{\text{KB}}{\longrightarrow} B) \supset ((A \overset{\text{KB}}{\longrightarrow} C) \supset (A \wedge \neg B \overset{\text{KB}}{\longrightarrow} C))$

Proof The proofs of these theorems are almost identical to those given for the analogous theorems in CT4D. We refer to the the proofs of Propositions 4.4 and 4.9.

■

Theorem 6.13 *Let M be a revision model for K with weak integrity constraints C. Then $K_A^{*^M} \models C$ for all A consistent with C.*

Proof Let $M = \langle W, R, \varphi \rangle$ be as specified, with $M \models \boxminus(C \supset \Box C)$. Since A is consistent with C, there exists some $A \wedge C$-world $w \in W$. Since $M \models_w \Box C$, we have $M \models_w A \wedge \Box(A \supset C)$. Thus $M \models A \overset{\text{KB}}{\longrightarrow} C$.

■

Corollary 6.14 *Let M be a revision model for K with weak integrity constraints C. Then $K \models C$.*

Proof Since C is satisfiable, there is some C-world $w \in W$, and $M \models_w \Box C$. Since all K-worlds are minimal in M, wRv for any K-world v, and $M \models_v C$. Thus $K \models C$.

■

Corollary 6.15 *If $KB \not\models C$ then $\{O(KB)\} \cup WIC$ is CO^*-inconsistent.*

Proof This is immediate given Corollaries 6.3 and 6.14.

■

Theorem 6.16 *Let M be a revision model for K with strong integrity constraints C. Let $\mathcal{S}$ be the collection of maximal subsets $S_i \subseteq C$ such that S_i is consistent with A. For some $S \subseteq \mathcal{S}$, it is the case that $K_A^{*^M} \models \bigvee S$.*

Proof Let $S_i \subseteq C$ be a maximal subset of constraints consistent with A. Using, as usual, S_i to denote the conjunction of its members, there must be some world w satisfying $A \wedge S_i$. Now $M \models IC$ so

$$M \models \boxminus((S_i \wedge \overline{S_i}) \supset \Box(\overline{S_i} \supset S_i)).$$

Since S_i is maximal $M \models_w S_i \wedge \overline{S_i}$, so $M \models_w \Box(\overline{S_i} \supset S_i)$. This implies $M \models A \overset{\text{KB}}{\longrightarrow} \overline{S_i} \supset S_i$ for each $S_i \in \mathcal{S}$. However, since $\mathcal{S}$ is the set of maximal constraint sets consistent with A, it must be that

$$\vdash A \supset \bigvee\{\overline{S_i} : S_i \in \mathcal{S}\}.$$

Thus

$$M \models A \xrightarrow{\text{KB}} \bigvee\{S_i \in \mathcal{S}\};$$

and, in particular, this holds for some subset of $S \subset \mathcal{S}$.

■

Corollary 6.17 *Let M be a revision model for K with strong integrity constraints C. Then $K \models C$.*

Proof One member of IC is $\Box(C \supset \Box C)$. Thus, this result follows immediately from Corollary 6.14.

■

Corollary 6.18 *If $KB \not\models C$ then $\{O(KB)\} \cup IC$ is CO*-inconsistent.*

Proof This is immediate given Corollaries 6.3 and 6.17.

■

Theorem 6.19 *Let M be a revision model for K with prioritized integrity constraints $C_1, \cdots, C_n$. Then $K_A^{*^M} \models C_i$ whenever A is consistent with the conjunction of all constraints C_j, $j \leq i$.*

Proof Let $M = \langle W, R, \varphi \rangle$ be as specified. Since A is consistent with P_i, there exists some $A \wedge P_i$-world $w \in W$. Since $M \models ICP$, we have $M \models_w \Box P_i$. Thus $M \models_w A \wedge \Box(A \supset P_i)$. Thus $M \models A \xrightarrow{\text{KB}} P_i$.

■

Corollary 6.20 *Let M be a revision model for K with prioritized integrity constraints C. Then $K \models C$.*

Proof One member of ICP is $\Box(C \supset \Box C)$. Thus, this result follows immediately from Corollary 6.14.

■

Corollary 6.21 *If $KB \not\models C$ then $\{O(KB)\} \cup ICP$ is CO*-inconsistent.*

Proof This is immediate given Corollaries 6.3 and 6.20.

■

Corollary 6.22 *$B \leq_{EM} A$ iff $M \models \Box(\neg A \supset \Diamond \neg B)$.*

Proof Let M be a CO*-model determining the entrenchment ordering $\leq_{EM}$ and plausibility ordering $\leq_{PM}$. By definition, $B \leq_{EM} A$ iff $\neg B \leq_{PM} \neg A$ iff $M \models \breve{\Box}(\neg A \supset \Diamond \neg B)$.

■

Theorem 6.23 *Let M be a CO*-model. Then $\leq_{PM}$ satisfies the Grove postulates (G1)–(G5).*

Proof We take $M = \langle W, R, \varphi \rangle$ to be a CO*-model with worlds w, v, etc.

(G1) Assume A and B are satisfiable (if not, $A \leq_{PM} B$ or $B \leq_{PM} A$). Suppose $A \not\leq_{PM} B$. Then for some w, $M \models_w B \wedge \neg \Diamond A$, so for all wRv, $M \models_v \neg A$. This means, for any A-world u, uRv and $M \models_u \Diamond B$. Hence, $B \leq_{PM} A$.

(G2) Suppose $A \leq_{PM} B$ and $B \leq_{PM} C$. Then M satsifies both $\breve{\Box}(B \supset \Diamond A)$ and $\breve{\Box}(C \supset \Diamond B)$; thus, $\breve{\Box}(C \supset \Diamond A)$ holds as well.

(G3) If $\vdash A \supset B \vee C$ then $\breve{\Box}(A \supset B \vee C)$ and, hence, $\breve{\Box}(A \supset \Diamond(B \vee C))$ are both CO*-valid. This implies $\breve{\Box}(A \supset \Diamond B)$ or $\breve{\Box}(A \supset \Diamond C)$. Thus $B \leq_{PM} A$ or $C \leq_{PM} A$.

(G4) Suppose A is consistent with K. Since all K-worlds are minimal in M, there is some A-world minimal in M and $\breve{\Box}(B \supset \Diamond A)$ for all B. Hence, $B \leq_{PM} A$ for all B.

(G5) If $\vdash \neg A$ the $\breve{\Box} \neg A$ is CO*-valid. Hence, $\breve{\Box}(A \supset \Diamond B)$ and $B \leq_{PM} A$ for all B.

■

Theorem 6.24 *Let $\leq_G$ be a Grove ordering satisfying (G1)–(G5). Then there exists a CO*-model M such that the plausibility ordering $\leq_{PM}$ determined by M is $\leq_G$.*

Proof The proof uses the technique of Grove (1988), which shows that $\leq_G$ can model a revision function. For the ordering $\leq_G$, let a *cut* $\mathcal{C}$ be any set of sentences satisfying the following closure property:

> If $A \in \mathcal{C}$ and $A \leq_G B$, then $B \in \mathcal{C}$.

Thus if we think of $\leq_G$ as assigning a degree of plausibility to sentences, a cut contains all sentences with at most a specified degree of plausibility (recall $A \leq_G B$ means A is *more* plausible).

It is easy to verify that cuts are totally ordered under set inclusion. Let $\mathcal{C}$ and $\mathcal{D}$ be two cuts with $A \in \mathcal{C}$ and $B \in \mathcal{D}$. For any pair of sentences A, B we have $A \leq_G B$ or $B \leq_G A$, so either $A \in \mathcal{D}$ or $B \in \mathcal{C}$. So $\mathcal{C} \subseteq \mathcal{D}$ or $\mathcal{D} \subseteq \mathcal{C}$.

Now we define a model $M = \langle W, R, \varphi \rangle$ where W is (as usual) the set of all maximal consistent sets of propositional sentences, φ is given by set membership of atoms, and R is defined as:

> wRv iff for every cut $\mathcal{C}$, $v \cap \mathcal{C} \neq \emptyset$ implies $w \cap \mathcal{C} \neq \emptyset$.

Since cuts are nested, it is easy to see R is reflexive, transitive and totally connected.

If B is unsatisfiable, by (G5), $A \leq_G B$ for all A, and in CO* we have that $\breve{\Box}(B \supset \Diamond A)$ is valid. So assume B is satisfiable, $A \leq_G B$, and $M \models_w B$. Let the set of cuts intersecting w be

$$\mathcal{S} = \{\mathcal{C} : \mathcal{C} \cap w \neq \emptyset\}.$$

Then $\mathcal{C} = \bigcap \mathcal{S}$ is clearly a cut, and $\mathcal{C} \cap w \neq \emptyset$ (since cuts are nested).

Consider the set $\{\neg D : D \in \mathcal{C}\}$. If this set is consistent with A, it can be extended to a maximal set v which includes A. Clearly, $\mathcal{C} \cap v = \emptyset$. As $\mathcal{C} \cap w \neq \emptyset$, any cut which intersects v contains $\mathcal{C}$, and intersects w as well. Hence, wRv. (Moreover, not vRw.)

Suppose $\{\neg D : D \in \mathcal{C}\}$ is inconsistent with A. Then for some $D_1, \ldots D_n \in \mathcal{C}$,

$$\vdash D_1 \wedge \ldots \wedge D_n \supset \neg A.$$

In other words,

$$\vdash A \supset D_1 \vee \ldots \vee D_n.$$

Using (G3), we see that $D_1 \leq_G A$ or $\ldots D_n \leq_G A$. This means $A \in \mathcal{C}$. Now let $\mathcal{D}$ be any cut smaller than $\mathcal{C}$, i.e., any $\mathcal{D} \subset \mathcal{C}$. Now consider the set $\{\neg D : D \in \mathcal{D}\}$. As shown above, if this set is consistent with A, it can be extended maximally to include A, determining a world v such that $\mathcal{D} \cap v = \emptyset$. If it is inconsistent with A, as discussed above, it must be that $A \in \mathcal{D}$. But since $A \leq_G B$, this implies $B \in \mathcal{D}$, contradicting the fact that $\mathcal{D} \subset \mathcal{C}$. Hence, there exists a world v satisfying A such that whenever $\mathcal{D} \subset \mathcal{C}$ it must be that $\mathcal{D} \cap v = \emptyset$. Thus wRv.

From this, we conclude that, for any B-world $w \in W$, there exists an A-world v such that wRv. Hence, $M \models \overset{\cdot\cdot}{\Box}(B \supset \Diamond A)$.

■

Corollary 6.25 *Let M be a CO*-model. Then $\leq_{EM}$ satisfies the entrenchment postulates (E1)–(E5) of (Gärdenfors 1988).*

Proof Theorem 6.23, together with the definition of $\leq_{EM}$ and the identity of $B \leq_G A$ with $\neg B \leq_E \neg A$, ensures this result.

■

Corollary 6.26 *Let $\leq_E$ be an entrenchment ordering satisfying (E1)–(E5). Then there exists a CO*-model M such that the entrenchment ordering $\leq_{EM}$ determined by M is $\leq_E$.*

Proof Theorem 6.24, together with the definition of $\leq_{EM}$ and the identity of $B \leq_G A$ with $\neg B \leq_E \neg A$, ensures this result.

■

Proposition 6.30 *Let M be a preorder revision model for propositional theory K. Then $M \models \boxtimes\alpha$ iff $K \models \alpha$ for any propositional α.*

Proof Let $M = \langle W, R, \varphi \rangle$ be a CT4O model with worlds w, v, u. Suppose $M \models \boxtimes\alpha$. Then at every world w, $M \models_w \Diamond\Box\alpha$, and at some world v, $M \models_v \Box\alpha$. Since all K-worlds are minimal in M, vRu for each K-world u, and hence each K-world satisfies α.

Suppose $K \models \alpha$. Since all and only K-worlds are minimal in M, we have for any K-world v that $M \models_v \Box\alpha$. Since each world w can see each such v, we have $M \models_w \Diamond\Box\alpha$ for all $w \in W$. Thus $M \models \boxtimes\alpha$.

■

Proposition 6.31 *Let M be a preorder revision model for K such that $M \models A \xrightarrow{\text{KB}} B$. Then $M \models \boxtimes(A \wedge \neg \boxtimes \neg B \supset B)$.*

Proof Let $M \models_w A$ for some K-world w in M. Since $M \models A \xrightarrow{\text{KB}} B$, we have $M \models_w \Diamond(A \wedge \Box(A \supset B))$. Hence $M \models_w B$, and $M \models_v A \wedge \neg \boxtimes \neg B \supset B$ holds at any K-world v and $M \models \boxtimes(A \wedge \neg \boxtimes \neg B \supset B)$.

■

Theorem 6.32 *Let $M = \langle W, R, \varphi \rangle$ be a CO-model. Then $M \models O^+(KB)$ iff $W = \mathcal{A} \cup \mathcal{I}$ where $\mathcal{A}$ and $\mathcal{I}$ are clusters such that $\mathcal{I}$ sees $\mathcal{A}$, and $\mathcal{I}$ consists of $\neg KB$-worlds while $\mathcal{A}$ consists of KB-worlds.*

Proof Suppose $M \models O^+(KB)$. We will show two facts:

(a) If $M \models_w KB$ then vRw for all $v \in W$.

(b) If $M \models_w \neg KB$ then for all $v \in W$, vRw if $M \models_v \neg KB$, and not vRw if $M \models_v KB$.

To show (a), suppose $M \models_w KB$ and for some v, not vRw. Since R is connected, wRv. If $M \models_v KB$ then $M \not\models_v \overline{\Box} \neg KB$, and $M \not\models O^+(KB)$. If $M \models_v \neg KB$ then $M \not\models_w \Box KB$, and $M \not\models O^+(KB)$.

To show (b), suppose $M \models_w \neg KB$. If $M \models_v KB$ and vRw for some v, then $M \not\models_v \Box KB$, and $M \not\models O^+(KB)$. If $M \models_v \neg KB$ and not vRw for some v, then $M \not\models_v \overline{\Box} \bot$, (since w is inaccessible) and $M \not\models O^+(KB)$.

■

Proposition 6.33 $\vdash_{CO} O^+(\top) \equiv O^+(\bot)$.

Proof

$$\begin{aligned} O^+(\bot) &\equiv (\bot \supset (\Box\bot \wedge \overline{\Box}\top)) \wedge (\bot \supset \overline{\Box}\top) \\ &\equiv \top \supset \overline{\Box}\bot \\ &\equiv (\top \supset (\Box\top \wedge \overline{\Box}\bot)) \wedge (\top \supset \overline{\Box}\bot) \\ &\equiv O^+(\top) \end{aligned}$$

■

Proposition 6.35 *Let M be a K-revision model. $M \models \boxtimes\alpha$ iff $M \models_w \alpha$ for each $w \in \mathcal{A}$.*

Proof Let $M = \langle W, R, \varphi \rangle$ be a K-revision model with worlds w, v, u. $M \models \boxtimes\alpha$ iff, at every world $u \in W$, $M \models_u \Diamond\Box\alpha$, iff at some world v, $M \models_v \Box\alpha$. Since $\mathcal{A}$ forms the set of worlds minimal in M, vRw for each $w \in \mathcal{A}$, Thus, this holds iff $M \models_w \alpha$ for each $w \in \mathcal{A}$.

■

Proposition 6.36 *Let M be a K-revision model. $M \models \overleftarrow{\boxtimes}\alpha$ iff $M \models_w \alpha$ for each $w \in \mathcal{I}$.*

Proof Let $M = \langle W, R, \varphi \rangle$ be a K-belief model with worlds w, v, u. $M \models \overleftarrow{\boxtimes}\alpha$ iff, at some world $u \in W$, $M \models_u \Box\overleftarrow{\Box}\alpha$. Since uRv for any R-minimal world v, this holds exactly when $M \models_v \overleftarrow{\Box}\alpha$ for each $v \in \mathcal{A}$. For any $v \in \mathcal{A}$ and $w \in \mathcal{I}$, vRw is false, so this holds iff $M \models_w \alpha$ for all $w \in \mathcal{I}$.

■

Proposition 6.37 $\vdash_{CO} O(\bot)$.

Proof

$$\begin{aligned} O(\bot) &\equiv \overleftarrow{\Box}(\bot \supset (\Box\bot \wedge \overleftarrow{\Box}\top)) \\ &\equiv \overleftarrow{\Box}\top \\ &\equiv \top \end{aligned}$$

■

We will now present several auxiliary lemmas from which the proof of Theorem 6.38 will follow immediately, demonstrating the subsumption of OL by CO*.

Lemma C.4 *Let $KB \subseteq \mathbf{L}_{OL}$ be such that KB^{Tr} is a nontrivial belief set. If $\vdash_{OL} \mathsf{O}KB \supset \mathsf{B}\alpha$ then $\vdash_{CO*} O(KB^{\mathrm{Tr}}) \supset \boxtimes\alpha^{\mathrm{Tr}}$.*

Proof Let $M = \langle W, R, \varphi \rangle$ be a nontrivial CO*-model of $O(KB^{\mathrm{Tr}})$, where $W = \mathcal{A} \cup \mathcal{I}$, as specified in Chapter 6. We will construct an OL-model $\overline{M}$ which verifies the "same" sentences as M. Let $\overline{M}$ be the OL-model $\langle \overline{\mathcal{A}}, \overline{\mathcal{I}} \rangle$ where

$$\overline{\mathcal{A}} = \{w* : w \in \mathcal{A}\} \text{ and } \overline{\mathcal{I}} = \{w* : w \in \mathcal{I}\},$$

taking $w*$ to be the valuation associated with world w in M. We recall that Levesque's (1990) structures take the valuations themselves to be possible worlds, thus eliminating the need for a valuation mapping. Clearly $\overline{M}$ is a (maximal) OL-model, as every valuation is represented in W. Define the inverse image of a valuation $w*$ to be

$$f(w*) = \{v \in W : v* = w*\}.$$

Lemma C.5 $\overline{M} \models_{w*} \alpha$ *iff* $M \models_v \alpha^{\mathrm{Tr}}$ *for all* $v \in f(w*)$.

Proof We proceed by structural induction on α (and we take v to be an arbitrary member of $f(w*)$). We use Propositions 6.35 and 6.36 in parts (d) and (e).

(a) For atomic α we have: $\overline{M} \models_{w*} \alpha$ iff $\alpha \in w*$ iff $\alpha \in \varphi(v)$ iff $M \models_v \alpha$. We assume now that this property holds for α and β.

(b) For $\neg\alpha$ we have: $\overline{M} \models_{w*} \neg\alpha$ iff $\overline{M} \not\models_{w*} \alpha$ iff $M \not\models_v \alpha$ iff $M \models_v \neg\alpha$.
(c) For $\alpha \supset \beta$ we have: $\overline{M} \models_{w*} \alpha \supset \beta$ iff $\overline{M} \not\models_{w*} \alpha$ or $\overline{M} \models_{w*} \beta$ iff $M \not\models_v \alpha$ or $M \models_v \beta$ iff $M \models_v \alpha \supset \beta$.
(d) For $\mathsf{B}\alpha$ we have: $\overline{M} \models_{w*} \mathsf{B}\alpha$ iff $\overline{M} \models_{u*} \alpha$ for all $u* \in \overline{\mathcal{A}}$ iff $M \models_u \alpha$ for all $u \in \mathcal{A}$ iff $M \models_v \boxtimes\alpha$ for any $v \in W$.
(e) For $\mathsf{N}\alpha$ we have: $\overline{M} \models_{w*} \mathsf{N}\alpha$ iff $\overline{M} \models_{u*} \alpha$ for all $u* \in \overline{\mathcal{I}}$ iff $M \models_u \alpha$ for all $u \in \mathcal{I}$ iff $M \models \overleftarrow{\boxtimes}\alpha$ iff $M \models_v \overleftarrow{\boxtimes}\alpha$ for all $v \in W$.

■

Since we have $M \models O(KB^{\mathrm{Tr}})$, by the lemma above, $\overline{M} \models_{w*} KB$ iff $w* \in \overline{\mathcal{A}}$. So $\overline{M} \models \mathsf{O}(KB)$. Furthermore, by the lemma above, if $\overline{M} \models \mathsf{B}\alpha$, then $M \models \boxtimes\alpha$.

Now suppose $\vdash_{OL} \mathsf{O}\alpha \supset \mathsf{B}\beta$. For any CO*-model M, where $M \models O(KB^{\mathrm{Tr}})$, we can construct an OL-model $\overline{M}$ as above satisfying $\mathsf{O}(KB)$. Furthermore, by supposition and the completeness of OL, $\overline{M} \models \mathsf{B}\alpha$ and, by the lemma, $M \models \boxtimes\alpha$. By the completeness theorem for CO*, $\vdash_{CO*} O(KB^{\mathrm{Tr}}) \supset \boxtimes\alpha^{\mathrm{Tr}}$.

■

Lemma C.6 *Let $KB \subseteq \mathbf{L}_{OL}$ be such that KB^{Tr} is a nontrivial belief set. If $\vdash_{CO*} O(KB^{\mathrm{Tr}}) \supset \boxtimes\alpha^{\mathrm{Tr}}$ then $\vdash_{OL} \mathsf{O}KB \supset \mathsf{B}\alpha$.*

Proof Let $\overline{M} = \langle \mathcal{A}, \mathcal{I} \rangle$ be an OL-model satisfying $\mathsf{O}(KB)$. We will construct a CO*-model which satisfies the "same" sentences. Let $M = \langle W, R, \varphi \rangle$ be a CO*-model such that

1. $W = \mathcal{A} \cup \mathcal{I}$
2. $\varphi(A) = \{w \in W : w \models A\}$
3. $R = \{\langle v, w \rangle : v, w \in \mathcal{A}\} \cup \{\langle v, w \rangle : v, w \in \mathcal{I}\} \cup \{\langle v, w \rangle : v \in \mathcal{I}, w \in \mathcal{A}\}$

Clearly M is a CO*-model and W can be partitioned as $\mathcal{A} \cup \mathcal{I}$ in the usual sense. Recall that OL-structures take the valuations themselves to be possible worlds.

Lemma C.7 $\overline{M} \models_w \alpha$ *iff* $M \models_v \alpha^{\mathrm{Tr}}$.

Proof We proceed by structural induction on α. We use Propositions 6.35 and 6.36 in parts (d) and (e).
(a) For atomic α we have: $\overline{M} \models_w \alpha$ iff $\alpha \in w$ iff $w \in \varphi(\alpha)$ iff $M \models_w \alpha$. We assume now that this property holds for α and β.
(b) For $\neg\alpha$ we have: $\overline{M} \models_w \neg\alpha$ iff $\overline{M} \not\models_w \alpha$ iff $M \not\models_w \alpha$ iff $M \models_w \neg\alpha$.
(c) For $\alpha \supset \beta$ we have: $\overline{M} \models_w \alpha \supset \beta$ iff $\overline{M} \not\models_w \alpha$ or $\overline{M} \models_w \beta$ iff $M \not\models_w \alpha$ or $M \models_w \beta$ iff $M \models_w \alpha \supset \beta$.
(d) For $\mathsf{B}\alpha$ we have: $\overline{M} \models_w \mathsf{B}\alpha$ iff $\overline{M} \models_u \alpha$ for all $u \in \mathcal{A}$ iff $M \models_u \alpha$ for all $u \in \mathcal{A}$ iff $M \models_v \boxtimes\alpha$ for any $v \in W$.
(e) For $\mathsf{N}\alpha$ we have: $\overline{M} \models_w \mathsf{N}\alpha$ iff $\overline{M} \models_u \alpha$ for all $u* \in \mathcal{I}$ iff $M \models_u \alpha$ for all $u \in \mathcal{I}$ iff $M \models_v \overleftarrow{\boxtimes}\alpha$ for all $v \in W$.

■

Since we have $\overline{M} \models \mathsf{O}(KB)$, by the lemma above, $M \models_w KB^{\mathrm{Tr}}$ iff $w \in \mathcal{A}$. So $M \models O(KB^{\mathrm{Tr}})$. Furthermore, if $M \models \boxtimes\alpha^{\mathrm{Tr}}$ then $\overline{M} \models \mathsf{B}\alpha$.

Now suppose $\vdash_{CO*} O(KB^{\mathrm{Tr}}) \supset \boxtimes\alpha^{\mathrm{Tr}}$. For any OL-model $\overline{M}$ satisfying $\mathsf{O}(KB)$, we can construct a CO*-model M as above satisfying $O(KB^{\mathrm{Tr}})$. By assumption and the completeness of CO*, $M \models \boxtimes\alpha^{\mathrm{Tr}}$, and by construction, it must be that $\overline{M} \models \mathsf{B}\alpha$. Hence, by the completeness of OL, $\vdash_{OL} \mathsf{O}(KB) \supset \mathsf{B}\alpha$.

■

Theorem 6.38 $\vdash_{OL} \mathsf{O}KB \supset \mathsf{B}\alpha$ *iff* $\vdash_{CO*} O^{+}(KB^{\mathrm{Tr}}) \supset \boxtimes\alpha^{\mathrm{Tr}}$.

Proof This follows immediately from Lemmas C.4 and C.6.

■

Theorem 6.39 *Let M be a CT4O*-model such that $M \models_{CT4O*} O^{+}(KB)$. Then M is a CO*-model such that $M \models_{CO*} O^{+}(KB)$.*

Proof This is obvious given the fact that a CT4O*-model satisfying $O^{+}(KB)$ must be totally connected.

■

Corollary 6.40 $\vdash_{CT4O*} O^{+}(KB) \supset \boxtimes\alpha$ *iff* $\vdash_{CO*} O^{+}(KB) \supset \boxtimes\alpha$.

Proof Immediate given Theorem 6.39.

■

Appendix D

Proofs of Theorems: Chapter 7

In this appendix we present proofs of various propositions, lemmas, theorems and corollaries found in Chapter 7.

Theorem 7.1 *Let Poss be a possibility measure. Then there exists a CO*-model M such that $\leq_{PM}$ is the plausibility ordering determined by M and $A \leq_{PM} B$ iff $Poss(A) \geq Poss(B)$.*

Proof Let *Poss* be a possibility measure. We construct a CO*-model M such that $\leq_{PM}$ respects *Poss*. We let $M = \langle W, R, \varphi \rangle$ be (as usual) determined by the set W of maximal consistent propositional sentences. We define the ***rank*** of $w \in W$ to be

$$r(w) = \min\{Poss(A) : A \in w\}.$$

We let wRv iff $r(w) \leq r(v)$.

It is easy to verify that M is a CO*-model. We can also show using the technique found in the proof of Theorem 6.24 (i.e. using *cuts* on the possibilty ordering) that every B-world in W sees some A-world in W whenever $Poss(B) \leq Poss(A)$.

■

Theorem 7.2 *Let $\leq_{PM}$ be the plausibility ordering determined by some CO*-model M. Then there exists a possibility measure such that $Poss(A) \geq Poss(B)$ iff $A \leq_{PM} B$.*

Proof Let $M = \langle W, R, \varphi \rangle$ be a CO*-model. Since W consists of a set of totally ordered clusters, we can assign a rank $r(w)$ to each world w from the interval $[0, 1]$ such that $r(w) \leq r(v)$ iff wRv. If W has a set of minimal elements in R, we can ensure that these get rank 1.

Now define the possibilty of A to be

$$Poss(A) = \sup\{r(w) : M \models_w A\}.$$

If no sentence gets assigned possibilty 1 under this scheme, W has no minimal elements. In this case, we alter this definition somewhat:

$$Poss(A) = \begin{cases} \sup\{r(w) : M \models_w A\} & \text{if } M \not\models \boxtimes A \\ 1 & \text{otherwise} \end{cases}$$

We let $Poss(A) = 0$ for any falsehood. It is easy to verify that *Poss* is a possibilty measure.

Suppose that $Poss(A) \geq Poss(B)$. In the case where $A \equiv \top$ or $B \equiv \bot$ it is easy to see that this holds iff $A \leq_{PM} B$. For both $\Box(B \supset \Diamond\top)$ and $\Box(\bot \supset \Diamond A)$ are CO*-valid for any A, B. Suppose both sentences are contingent. Then this holds iff, for every B-world, there exists an A-world which has rank at least as great as that B-world. Thus, for each B-world w, there exists an A-world v such that wRv. Hence, this is true iff $A \leq_{PM} B$.

■

Bibliography

Achinger, J. and Jankowski, A. W. 1986. On decidable consequence operators. *Studia Logica*, 45(4):415–424.

Adams, E. W. 1975. *The Logic of Conditionals*. D.Reidel, Dordrecht.

Alchourrón, C., Gärdenfors, P., and Makinson, D. 1985. On the logic of theory change: Partial meet contraction and revision functions. *Journal of Symbolic Logic*, 50:510–530.

Anderson, A. R. and Belnap, N. D. 1962. The pure calculus of entailment. *Journal of Symbolic Logic*, 27:19–52.

Appiah, A. 1985. *Assertion and Conditionals*. Cambridge University Press, Cambridge.

Åqvist, L. 1967. Good Samaritans, contrary-to-duty imperatives, and epistemic obligations. *Noûs*, 1:361–379.

Asher, N. and Morreau, M. 1991. Commonsense entailment: A modal theory of nonmonotonic reasoning. In *Proceedings of the Twelfth International Joint Conference on Artificial Intelligence*, pages 387–392, Sydney.

Bacchus, F. 1988. Representing and reasoning with uncertainty. Research Report CS–88–31, University of Waterloo.

Bacchus, F. 1989. A modest, but semantically well founded, inheritance reasoner. In *Proceedings of the Eleventh International Joint Conference on Artificial Intelligence*, pages 1104–1109, Detroit.

Bacchus, F. 1990. *Representing and Reasoning with Probabilistic Knowledge*. MIT Press, Cambridge.

Beatty, H. 1972. On evaluating deontic logics. *Journal of Philosophical Logic*, 1:439–444.

Bell, J. 1990. The logic of nonmonotonicity. *Artificial Intelligence*, 41:365–374.

Bell, J. L. and Machover, M. 1977. *A Course in Mathematical Logic*. North-Holland, Amsterdam.

Belnap, N. D. 1977. A useful four-valued logic. In Epstein, G. and Dunn, J. M., editors, *Modern Uses of Multiple-Valued Logic*, pages 8–37. Reidel, Boston.

Besnard, P. 1988. Axiomatizations in the metatheory of nonmonotonic inference systems. In *Proceedings of Canadian Society for Computational Studies of Intelligence Conference*, pages 117–124, Edmonton.

Bonner, A. J. 1988. A logic for hypothetical reasoning. In *Proceedings of the Seventh National Conference on Artificial Intelligence*, pages 480–484, St. Paul.

Bonner, A. J., McCarty, L. T., and Vadaparty, K. 1989. Expressing database queries with intuitionistic logic. In *Proceedings of the North American Conference on Logic Programming*, pages 831–850.

Bossu, G. and Siegel, P. 1985. Saturation nonmonotonic reasoning and the closed-world assumption. *Artificial Intelligence*, 25:13–63.

Boutilier, C. 1988. Default reasoning with the conditional logic E. Master's thesis, University of Toronto, Toronto.

Boutilier, C. 1989. A semantical approach to stable inheritance reasoning. In *Proceedings of the Eleventh International Joint Conference on Artificial Intelligence*, pages 1134–1139, Detroit.

Boutilier, C. 1990. Conditional logics of normality as modal systems. In *Proceedings of the Eighth National Conference on Artificial Intelligence*, pages 594–599, Boston.

Boutilier, C. 1991a. Default priorities as epistemic entrenchment. Technical Report KRR-TR-91-2, University of Toronto, Toronto.

Boutilier, C. 1991b. Inaccessible worlds and irrelevance: Preliminary report. In *Proceedings of the Twelfth International Joint Conference on Artificial Intelligence*, pages 413–418, Sydney.

Boutilier, C. 1991c. A modal analysis of subjunctive queries. In *Proceedings of Workshop on Nonstandard Queries and Nonstandard Answers*, volume 1, pages 13–32, Toulouse.

Boutilier, C. 1991d. On the semantics of stable inheritance reasoning. *Computational Intelligence.* (To appear).

Brewka, G. 1991. *Nonmonotonic Reasoning: Logical Foundations of Commonsense.* Cambridge University Press, Cambridge.

Bull, R. A. 1966. That all normal extensions of S4.3 have the finite model property. *Zeitschrift für Mathematische Logik und Grundlagen der Mathematik*, 12:341–344.

Carnap, R. 1950. *Logical Foundations of Probability.* University of Chicago Press, Chicago.

Chellas, B. F. 1975. Basic conditional logic. *Journal of Philosophical Logic*, 4:133–153.

Chellas, B. F. 1980. *Modal Logic: An Introduction.* Cambridge University Press, Cambridge.

Cherniak, C. 1986. *Minimal Rationality.* MIT Press, Cambridge.

Chisholm, R. M. 1966. *Theory of Knowledge.* Prentice-Hall, Englewood Cliffs.

Clark, K. 1978. Negation as failure. In Gallaire, H. and Minker, J., editors, *Logic and Databases*, pages 119–140. Plenum, New York.

Cosmadakis, S. S. and Papadimitriou, C. H. 1983. Updates of relational views. In *Proceedings of SIGACT-SIGMOD Symposium on Principles of Database Systems*, pages 317–331.

Czelakowski, J. and Mrmalinowski, G. 1985. Key notions of Tarski's methodology of deductive systems. *Studia Logica*, 44(4):321–351.

Dalal, M. 1988. Investigations into a theory of knowledge base revision: Preliminary report. In *Proceedings of the Seventh National Conference on Artificial Intelligence*, pages 475–479, St. Paul.

Davis, M. 1980. The mathematics of non-monotonic reasoning. *Artificial Intelligence*, 13:73–80.

de Kleer, J. 1986. An assumption-based TMS. *Artificial Intelligence*, 28:127–161.

de Kleer, J. and Williams, B. C. 1987. Diagnosing multiple faults. *Artificial Intelligence*, 32:97–130.

Delgrande, J. P. 1986. A propositional logic for natural kinds. In *Proceedings of the Canadian Society for Computational Studies of Intelligence Conference*, pages 44–48, Montreal.

Delgrande, J. P. 1987. A first-order logic for prototypical properties. *Artificial Intelligence*, 33:105–130.

Delgrande, J. P. 1988. An approach to default reasoning based on a first-order conditional logic: Revised report. *Artificial Intelligence*, 36:63–90.

Delgrande, J. P. 1990. A semantics for a class of inheritance networks. In *Proceedings of the Canadian Society for Computational Studies of Intelligence Conference*, pages 54–60, Ottawa.

Dennett, D. C. 1978. *Brainstorms.* Bradford, Cambridge.

Doyle, J. 1979. A truth maintenance system. *Artificial Intelligence*, 12:231–272.

Doyle, J. 1983. Methodological simplicity in expert system construction: The case for judgements and reasoned assumptions. In Shafer, G. and Pearl, J., editors, *Readings in Uncertain Reasoning*, pages 689–693. Morgan Kaufmann, San Mateo. 1990.

Doyle, J. 1991. Rational belief revision: Preliminary report. In *Proceedings of Second International Conference on Principles of Knowledge Representation and Reasoning*, pages 163–174, Cambridge.

Dubois, D. and Prade, H. 1988. *Possibility Theory: An Approach to Computerized Processing of Uncertainty.* Plenum, New York.

Dummett, M. A. E. and Lemmon, E. J. 1959. Modal logics between S4 and S5. *Zeitschrift für Mathematische Logik und Grundlagen der Mathematik*, 5:250–264.

Etherington, D. W. 1987a. Relating default logic and circumscription. In *Proceedings of the Tenth International Joint Conference on Artificial Intelligence*, pages 489–494, Milan.

Etherington, D. W. 1987b. A semantics for default logic. In *Proceedings of the Tenth International Joint Conference on Artificial Intelligence*, pages 495–498, Milan.

Etherington, D. W. 1988. *Reasoning with Incomplete Information: Investigations of Non-monotonic Reasoning.* Pitman, London.

Etherington, D. W., Mercer, R. E., and Reiter, R. 1985. On the adequacy of predicate circumscription for closed-world reasoning. *Computational Intelligence*, 1:11–15.

Etherington, D. W. and Reiter, R. 1983. On inheritance hierarchies with exceptions. In *Proceedings of the Third National Conference on Artificial Intelligence*, pages 104–108, Washington.

Fagin, R., Ullman, J. D., and Vardi, M. Y. 1983. On the semantics of updates in databases: Preliminary report. In *Proceedings of SIGACT-SIGMOD Symposium on Principles of Database Systems*, pages 352–365.

Fariñas, del Cerro, L. and Herzig, A. 1991. A modal analysis of possibility theory. In *Proceedings of Workshop on Nonstandard Queries and Nonstandard Answers*, volume 2, pages 181–188, Toulouse.

Gabbay, D. 1985. Theoretical foundations for non-monotonic reasoning in expert systems. In Apt, K. R., editor, *Logics and Models of Concurrent Systems*, pages 439–457. Springer-Verlag, Berlin.

Gärdenfors, P. 1978a. Conditionals and changes in belief. *Acta Philosophica Fennica*, 30:381–404.

Gärdenfors, P. 1978b. On the logic of relevance. *Synthese*, 37(3):351–367.

Gärdenfors, P. 1984. Epistemic importance and minimal changes of belief. *Australasian Journal of Philosophy*, 62(2):136–157.

Gärdenfors, P. 1986. Belief revisions and the ramsey test for conditionals. *The Philosophical Review*, 95:81–93.

Gärdenfors, P. 1987. Variations on the ramsey test: More triviality results. *Studia Logica*, 46(4):321–327.

Gärdenfors, P. 1988. *Knowledge in Flux: Modeling the Dynamics of Epistemic States.* MIT Press, Cambridge.

Gärdenfors, P. 1990. Belief revision and nonmonotonic logic: Two sides of the same coin? In *Proceedings of the European Conference on Artificial Intelligence*, pages 768–773.

Gärdenfors, P. and Makinson, D. 1991. Nonmonotonic inference based on expectations. To appear.

Garey, M. R. and Johnson, D. S. 1979. *Computers and Intractability: A Guide to the Theory of NP-Completeness.* Freeman, New York.

Geffner, H. 1989. Default reasoning: Causal and conditional theories. Technical Report 137, Department of Computer Science, UCLA, Los Angeles.

Gelfond, M., Przymusinska, H., and Przymusinski, T. 1989. On the relationshp between circumscription and negation as failure. *Artificial Intelligence*, 38:75–94.

Gentzen, G. 1934. Investigations into logical deduction. In Szabo, M., editor, *The Collected Works of Gerhard Gentzen*, pages 68–129. North Holland, Amsterdam. 1968.

Gettier, E. 1963. Is knowledge justified true belief? *Analysis*, 23:121–123.

Ginsberg, M. L. 1986. Counterfactuals. *Artificial Intelligence*, 30(1):35–79.

Glymour, C. 1988. Artificial intelligence is philosophy. In Fetzer, J. H., editor, *Aspects of Artificial Intelligence*, pages 195–207. Kluwer, Dordrecht.

Goldszmidt, M., Morris, P., and Pearl, J. 1990. A maximum entropy approach to nonmonotonic reasoning. In *Proceedings of the Eighth National Conference on Artificial Intelligence*, pages 646–652, Boston.

Goldszmidt, M. and Pearl, J. 1989. On the consistency of defeasible databases. In *Proceedings of the 5th Workshop on Uncertainty in AI*, pages 131–141, Windsor.

Goldszmidt, M. and Pearl, J. 1990. On the relation between rational closure and system Z. In *Third International Workshop on Nonmonotonic Reasoning*, pages 130–140, South Lake Tahoe.

Goodman, N. 1955. *Fact, Fiction, and Forecast.* Harvard University Press, Harvard.

Grahne, G. 1991. Updates and counterfactuals. In *Proceedings of Second International Conference on Principles of Knowledge Representation and Reasoning*, pages 269–276, Cambridge.

Grahne, G. and Mendelzon, A. O. 1991. Updates and subjunctive queries. In *Proceedings of Workshop on Nonstandard Queries and Nonstandard Answers*, volume 1, pages 33–52, Toulouse.

Grove, A. 1988. Two modellings for theory change. *Journal of Philosophical Logic*, 17:157–170.

Haack, S. 1978. *Philosophy of Logics.* Cambridge University Press, Cambridge.

Halpern, J. Y. and Moses, Y. O. 1985. A guide to the modal logics of knowledge and belief: A preliminary draft. In *Proceedings of the Ninth International Joint Conference on Artificial Intelligence*, pages 480–490, Los Angeles.

Halpern, J. Y. and Rabin, M. O. 1987. A logic to reason about likelihood. *Artificial Intelligence*, 32:379–405.

Hanks, S. and McDermott, D. 1986. Default reasoning, nonmonotonic logics and the frame problem. In *Proceedings of the Fifth National Conference on Artificial Intelligence*, pages 328–333, Philadelphia.

Hansson, B. 1969. An analysis of some deontic logics. *Noûs*, 3:373–398.

Harman, G. 1986. *Change in View.* MIT Press, Cambridge.

Hayes, P. J. 1974. Some problems and non-problems in representation theory. In Brachman, R. J. and Levesque, H. J., editors, *Readings in Knowledge Representation*, pages 4–22. Morgan Kaufmann, Los Altos. 1985.

Hintikka, J. 1962. *Knowledge and Belief.* Cornell University Press, Ithaca.

Hirst, G. 1991. Existence assumptions in knowledge representation. *Artificial Intelligence*, 49:199–242.

Horty, J. F. 1991. Moral dilemmas and nonmonotonic logic. In *Workshop on Deontic Logic in Computer Science*, Amsterdam. To appear.

Horty, J. F., Thomason, R. H., and Touretzky, D. S. 1987. A skeptical theory of inheritance in nonmonotonic semantic networks. Technical Report CMU-CS-87-175, Carnegie-Mellon University, Pittsburgh.

Hughes, G. E. and Cresswell, M. J. 1968. *An Introduction to Modal Logic.* Methuen, London.

Hughes, G. E. and Cresswell, M. J. 1984. *A Companion to Modal Logic.* Methuen, London.

Humberstone, I. L. 1983. Inaccessible worlds. *Notre Dame Journal of Formal Logic*, 24(3):346–352.

Israel, D. 1980. What's wrong with non-monotonic logic? In *Proceedings of the First National Conference on Artificial Intelligence*, pages 99–101, Stanford.

Jackson, F. 1987. *Conditionals.* Blackwell, Oxford.

Jackson, P. 1989. On the semantics of counterfactuals. In *Proceedings of the Eleventh International Joint Conference on Artificial Intelligence*, pages 1382–1387, Detroit.

Johnson-Laird, P. 1983. *Mental Models.* Harvard University Press, Cambridge.

Katsuno, H. and Mendelzon, A. O. 1990. Propositional knowledge base revision and minimal change. Technical Report KRR-TR-90-3, University of Toronto, Toronto.

Katsuno, H. and Mendelzon, A. O. 1991. On the difference between updating a knowledge database and revising it. In *Proceedings of Second International Conference on Principles of Knowledge Representation and Reasoning*, pages 387–394, Cambridge.

Katsuno, H. and Satoh, K. 1991. A unified view of consequence relation, belief revision and conditional logic. In *Proceedings of the Twelfth International Joint Conference on Artificial Intelligence*, pages 407–412, Sydney.

Kautz, H. A. 1986. The logic of persistence. In *Proceedings of the Fifth National Conference on Artificial Intelligence*, pages 401–405, Philadelphia.

Kelly, K. T. 1988. AI and effective epistemology. In Fetzer, J. H., editor, *Aspects of Artificial Intelligence*, pages 309–322. Kluwer, Dordrecht.

Kolaitis, P. and Papadimitriou, C. 1988. Some computational aspects of circumscription. In *Proceedings of the Seventh National Conference on Artificial Intelligence*, pages 465–469, St. Paul.

Konolige, K. 1987. On the relation between default theories and autoepistemic logic. *Artificial Intelligence.*

Kraus, S., Lehmann, D., and Magidor, M. 1990. Nonmonotonic reasoning, preferential models and cumulative logics. *Artificial Intelligence*, 44:167–207.

Kripke, S. A. 1963. Semantical analysis of modal logic I, normal propositional calculi. *Zeitschrift für Mathematische Logik und Grundlagen der Mathematik*, 9:67–96.

Kripke, S. A. 1980. *Naming and Necessity.* Harvard University Press, Harvard.

Kyburg, Jr., H. E. 1961. *Probability and the Logic of Rational Belief.* Wesleyan University Press, Middletown.

Ladner, R. E. 1977. The computational complexity of provability in systems of modal propositional logic. *Siam Journal of Computing*, 6(3):467–480.

Lakemeyer, G. 1987. Tractable meta-reasoning in propositional logics of belief. In *Proceedings of the Tenth International Joint Conference on Artificial Intelligence*, pages 402–408, Milan.

Lakemeyer, G. 1991. On the relation between explicit and implict belief. In *Proceedings of Second International Conference on Principles of Knowledge Representation and Reasoning*, pages 368–375, Cambridge.

Lamarre, P. 1991. S4 as the conditional logic of nonmonotonicty. In *Proceedings of Second International Conference on Principles of Knowledge Representation and Reasoning*, pages 357–367, Cambridge.

Lehmann, D. 1989. What does a conditional knowledge base entail? In *Proceedings of First International Conference on Principles of Knowledge Representation and Reasoning*, pages 212–222, Toronto.

Lehmann, D. and Magidor, M. 1990. What does a conditional knowledge base entail? Technical Report TR-90-10, Hebrew University, Jerusalem.

Levesque, H. 1986a. Making believers out of computers. *Artificial Intelligence*, 30:81–107.

Levesque, H. J. 1981. The interaction with incomplete knowledge bases: A formal treatment. In *Proceedings of the Seventh International Joint Conference on Artificial Intelligence*, pages 240–245, Vancouver.

Levesque, H. J. 1984a. Foundations of a functional approach to knowledge representation. *Artificial Intelligence*, 23:155–212.

Levesque, H. J. 1984b. A logic of implicit and explicit belief. In *Proceedings of the Fourth National Conference on Artificial Intelligence*, pages 198–202, Austin.

Levesque, H. J. 1986b. Knowledge representation and reasoning. *Annual Review of Computer Science*, 1:255–287.

Levesque, H. J. 1990. All I know: A study in autoepistemic logic. *Artificial Intelligence*, 42:263–309.

Levesque, H. J. and Brachman, R. J. 1985. A fundamental tradeoff in knowledge representation and reasoning. In Brachman, R. J. and Levesque, H. J., editors, *Readings in Knowledge Representation*, pages 42–70. Morgan Kaufmann, Los Altos.

Levi, I. 1980. *The Enterprise of Knowledge.* MIT Press, Cambridge.

Levinson, S. C. 1983. *Pragmatics.* Cambridge University Press, Cambridge.

Lewis, D. 1968. Counterpart theory and quantified modal logic. *Journal of Philosophy*, 65.

Lewis, D. 1973a. *Counterfactuals.* Blackwell, Oxford.

Lewis, D. 1973b. Counterfactuals and comparative possibility. *Journal of Philosophical Logic*, 2:418–446.

Lewis, D. 1976. Probabilities of conditionals and conditional probability. *Philosophical Review*, 85:297–315.

Lifschitz, V. 1985a. Closed-world databases and circumscription. *Artificial Intelligence*, 27:229–235.

Lifschitz, V. 1985b. Computing circumscription. In *Proceedings of the Ninth International Joint Conference on Artificial Intelligence*, pages 121–127, Los Angeles.

Lifschitz, V. 1986a. On the satisfiability of circumscription: Research note. *Artificial Intelligence*, 28:17–27.

Lifschitz, V. 1986b. Pointwise circumscription: Preliminary report. In *Proceedings of the Fifth National Conference on Artificial Intelligence*, pages 406–410, Philadelphia.

Lifschitz, V. 1987. Formal theories of action: Preliminary report. In *Proceedings of the Sixth National Conference on Artificial Intelligence*, pages 966–972, Milan.

Lin, F. and Shoham, Y. 1989. Argument systems: a uniform basis for nonmonotonic reasoning. In *Proceedings of First International Conference on Principles of Knowledge Representation and Reasoning*, pages 245–255, Toronto.

Lindström, S. and Rabinowicz, W. 1989. On the probabilitic representation of non-probabilistic belief revision. *Journal of Philosophical Logic*, 18:69–101.

Loui, R. 1987a. Defeat among arguments: A system of defeasible inference. *Computational Intelligence*, 3:100–106.

Loui, R. 1987b. Real rules of inference. *Communication and Cognition.*

Łukaszewicz, W. 1985. Two results on default logic. In *Proceedings of the Ninth International Joint Conference on Artificial Intelligence*, pages 459–461, Los Angeles.

Łukaszewicz, W. 1988. Considerations on default logic. *Computational Intelligence*, 4:1–16.

MacLennan, B. 1988. Logic for the new AI. In Fetzer, J. H., editor, *Aspects of Artificial Intelligence*, pages 309–322. Kluwer, Dordrecht.

Makinson, D. and Gärdenfors, P. 1990. Relations between the logic of theory change and nonmonotonic logic. In Fuhrmann, A. and Morreau, M., editors, *The Logic of Theory Change*, pages 185–205. Springer-Verlag, Berlin.

Makinson, D. C. 1966. There are infinitely many diodorean modal functions. *Journal of Symbolic Logic*, 31:406–408.

Makinson, D. C. 1989. General thory of cumulative inference. In Reinfrank, M., deKleer, J., Ginsberg, M. L., and Sandewall, E., editors, *Nonmonotonic Reasoning*, pages 1–18. Springer-Verlag, Berlin. 1990.

Marek, W., Shvarts, G. F., and Truszczyński, M. 1991. Modal nonmonotonic logics: Ranges, characerization, computation. In *Proceedings of Second International Conference on Principles of Knowledge Representation and Reasoning*, pages 395–404, Cambridge.

Marek, W. and Truszczyński, M. 1989. Relating autoepistemic and default logics. In *Proceedings of First International Conference on Principles of Knowledge Representation and Reasoning*, pages 276–288, Toronto.

McArthur, G. 1988. Reasoning about knowledge and belief: A survey. *Computational Intelligence*, 4(3):223–243.

McCarthy, J. 1977. Epistemological problems in artificial intelligence. In Brachman, R. J. and Levesque, H. J., editors, *Readings in Knowledge Representation*, pages 24–30. Morgan Kaufmann, Los Altos. 1985.

McCarthy, J. 1980. Circumscription – A form of non-monotonic reasoning. *Artificial Intelligence*, 13:27–39.

McCarthy, J. 1986. Applications of circumscription to formalizing commonsense reasoning. *Artificial Intelligence*, 28:89–116.

McCarthy, J. and Hayes, P. 1969. Some philosophical problems from the standpoint of artificial intelligence. *Machine Intelligence*, 4:463–502.

McDermott, D. 1982. Nonmonotonic logic II: nonmonotonic modal theories. *Journal of the Association for Computing Machinery*, 29:33–57.

McDermott, D. 1987. A critique of pure reason. *Computational Intelligence*, 3(3):151–160.

McDermott, D. and Doyle, J. 1980. Nonmonotonic logic I. *Artificial Intelligence*, 13:41–72.

Mercer, R. E. 1988. Using default logic to derive natural language presupposition. In *Proceedings of the Canadian Society for Computational Studies of Intelligence Conference*, pages 14–21, Edmonton.

Minker, J. 1983. On definite databases and the closed world assumption. In *Proceedings of the 6th Conference on Automated Deduction*, pages 292–308.

Minsky, M. 1974. A framework for representing knowledge. In Brachman, R. J. and Levesque, H. J., editors, *Readings in Knowledge Representation*, pages 246–262. Morgan Kaufmann, Los Altos. 1985.

Moore, R. C. 1982. The role of logic in knowledge representation and commonsense reasoning. In *Proceedings of the Second National Conference on Artificial Intelligence*, pages 428–433, Pittsburgh.

Moore, R. C. 1985. Semantical considerations for nonmonotonic logic. *Artificial Intelligence*, 25:75–94.

Nakamura, A. 1980. The computational complexity of satisfiability in systems of modal logic. In *Proceedings of the Fifth IBM Symposium on Mathematical Foundations of Computing Sciences: Computational Complexity*, Hiroshima.

Nebel, B. 1989. A knowledge level analysis of belief revision. In *Proceedings of First International Conference on Principles of Knowledge Representation and Reasoning*, pages 301–311, Toronto.

Newell, A. 1982. The knowledge level. *Artificial Intelligence*, 18:87–127.

Nute, D. 1975. Counterfactuals. *Notre Dame Journal of Formal Logic*, 16(4):476–482.

Nute, D. 1980. *Topics in Conditional Logic.* D.Reidel, Dordrecht.

Nute, D. 1984a. A non-monotonic logic based on conditional logic. Research Report 01-0007, Univeristy of Georgia, Athens.

Nute, D. 1984b. Non-monotonic reasoning and conditionals. Research Report 01-0002, Advanced Computational Methods Center, Univeristy of Georgia, Athens.

Ono, H. and Nakamura, A. 1980. On the size of refutation Kripke models for some linear modal and tense logics. *Studia Logica*, 39(4):325–333.

Padgham, L. 1989. Negative reasoning using inheritance. In *Proceedings of the Eleventh International Joint Conference on Artificial Intelligence*, pages 1086–1092, Detroit.

Pappas, G., editor 1979. *Justification and Knowledge.* D. Reidel, Dordrecht.

Patel-Schneider, P. F. 1987. Decidable, logic-based knowledge representation. Technical Report 201/87, University of Toronto, Toronto.

Pearl, J. 1988. *Probabilistic Reasoning in Intelligent Systems: Networks of Plausible Inference.* Morgan Kaufmann, San Mateo.

Pearl, J. 1989. Probabilistic semantics for nonmonotonic reasoning: A survey. In *Proceedings of First International Conference on Principles of Knowledge Representation and Reasoning*, pages 505–516, Toronto.

Pearl, J. 1990. System Z: A natural ordering of defaults with tractable applications to default reasoning. In Vardi, M., editor, *Proceedings of Theoretical Aspects of Reasoning about Knowledge*, pages 121–135. Morgan Kaufmann, San Mateo.

Perlis, D. and Minker, J. 1986. Completeness results for circumscription. *Artificial Intelligence*, 28:29–42.

Pollock, J. L. 1984. *The Foundations of Philosophical Semantics.* Princeton University Press, Princeton.

Pollock, J. L. 1986. *Contemporary Theories of Knowledge.* Rowman & Littlefield, Totowa.

Pollock, J. L. 1987. Defeasible reasoning. *Cognitive Science*, 11:481–518.

Poole, D. 1988. A logical framework for default reasoning. *Artificial Intelligence*, 36:27–47.

Poole, D. 1989. Explanation and prediction: An architecture for default and abductive reasoning. *Computational Intelligence*, 5:97–110.

Poole, D. 1991. The effect of knowledge on belief: Conditioning, specificity and the lottery paradox in default reasoning. *Artificial Intelligence*, 49:281–307.

Poole, D. L. 1985. On the comparison of theories: Preferring the most specific explanation. In *Proceedings of the Ninth International Joint Conference on Artificial Intelligence*, pages 144–147, Los Angeles.

Pratt, V. R. 1976. Semantical considerations on floyd-hoare logic. In *IEEE Symposium on Foundations of Computer Science*, pages 109–121.

Prior, A. N. 1967. *Past, Present and Future.* Clarendon Press, Oxford.

Przymusinski, T. 1989. An algorithm to compute circumscription. *Artificial Intelligence*, 38:49–73.

Putnam, H. 1970. Is semantics possible? In Kiefer, H. E. and Munitz, M. K., editors, *Language, Belief and Metaphysics*, pages 50–63. SUNY Press, Albany.

Quine, W. 1960. *Word and Object.* MIT Press, Cambridge, 2nd edition.

Quine, W. 1961. *From a Logical Point of View.* Harvard Univeristy Press, Cambridge, 2nd edition.

Quine, W. and Ullian, J. 1970. *The Web of Belief.* Random House, New York.

Rankin, T. L. 1988. When is reasoning nonmonotonic? In Fetzer, J. H., editor, *Aspects of Artificial Intelligence*, pages 309–322. Kluwer, Dordrecht.

Rasiowa, H. and Sikorski, R. 1963. *The Mathematics of Metamathematics.* Państowe Wydawnictwo Naukowe, Warszawa.

Reiter, R. 1978a. On closed world databases. In Gallaire, H. and Minker, J., editors, *Logic and Databases*, pages 55–76. Plenum, New York.

Reiter, R. 1978b. On reasoning by default. In Brachman, R. J. and Levesque, H. J., editors, *Readings in Knowledge Representation*, pages 401–410. Morgan Kaufmann, Los Altos. 1985.

Reiter, R. 1980. A logic for default reasoning. *Artificial Intelligence*, 13:81–132.

Reiter, R. 1987a. Nonmonotonic reasoning. *Annual Review of Computer Science*, 2:147–186.

Reiter, R. 1987b. A theory of diagnosis from first principles. *Artificial Intelligence*, 32:57–95.

Reiter, R. 1990. What should a database know? Technical Report KRR-TR-90-5, University of Toronto, Toronto.

Reiter, R. and Criscuolo, G. 1981. On interacting defaults. In *Proceedings of the Seventh International Joint Conference on Artificial Intelligence*, pages 270–276, Vancouver.

Reiter, R. and de Kleer, J. 1987. Foundations of assumption-based truth maintenance systems: Preliminary report. In *Proceedings of the Sixth National Conference on Artificial Intelligence*, pages 183–188, Seattle.

Rosch, E. 1978. Principles of categorization. In Rosch, E. and Lloyd, B., editors, *Cognition and Categorization*, pages 27–48. Lawrence Erlbaum Publishers, Hillsdale.

Rott, H. 1989. Conditionals and theory change: Revisions, expansions, and additions. *Synthese*, 81(1):91–113.

Rumelhart, D. E., McClelland, J. L., and PDP Research Group 1986. *Parallel Distributed Processing*, volume 1. MIT Press, Cambridge.

Sandewall, E. 1986. Nonmonotonic inference rules for multiple inheritance with exceptions. *Proceedings of the Institute of Electrical and Electronics Engineers*, 74(10):1345–1353.

Segerberg, K. 1970. Modal logics with linear alternative relations. *Theoria*, 36:310–322.

Segerberg, K. 1971. *An Essay in Classical Modal Logic.* Department of Philosophy, University of Uppsala, Uppsala.

Selman, B. 1990. *Tractable Default Reasoning.* PhD thesis, University of Toronto.

Selman, B. and Kautz, H. 1988. The complexity of model preference defaults. In *Proceedings of the Canadian Society for Computational Studies of Intelligence Conference*, pages 102–109, Edmonton.

Shepherdson, J. 1988. Negation in logic programming. In Minker, J., editor, *Foundations of Logic Programming and Deductive Databases*, pages 19–88. Morgan Kaufmann, Los Altos.

Shoham, Y. 1987. A semantical approach to nonmonotonic logics. In *Proceedings of the Symposium on Logic in Computer Science*, pages 275–279, Ithaca.

Shoham, Y. 1988. *Reasoning About Change: Time and Causation from the Standpoint of Artificial Intelligence.* MIT Press, Cambridge.

Smith, B. C. 1982. Prologue to "Reflections and Semantics in a Procedural Language". In Brachman, R. J. and Levesque, H. J., editors, *Readings in Knowledge Representation*, pages 32–40. Morgan Kaufmann, Los Altos. 1985.

Smullyan, A. 1948. Modality and description. *Journal of Symbolic Logic*, 13:31–37.

Stalnaker, R. C. 1968. A theory of conditionals. In Harper, W., Stalnaker, R., and Pearce, G., editors, *Ifs*, pages 41–55. D. Reidel, Dordrecht. 1981.

Stalnaker, R. C. 1980. A defense of conditional excluded middle. In Harper, W. L., Stalnaker, R., and Pearce, G., editors, *Ifs*, pages 87–104. D.Reidel, Dordecht. 1981.

Stalnaker, R. C. 1984. *Inquiry.* MIT Press, Cambridge.

Swain, M. 1981. *Reasons and Knowledge.* Cornell University Press, Ithaca.

Touretzky, D. S. 1986. *The Mathematics of Inheritance Systems.* Pitman, London.

Touretzky, D. S., Horty, J. F., and Thomason, R. H. 1987. A clash of intuitions: The current state of nonmonotonic multiple inheritance systems. In *Proceedings of the Tenth International Joint Conference on Artificial Intelligence*, pages 476–482, Milan.

Truszczyński, M. 1991. Modal interpretation of default logic. In *Proceedings of the Twelfth International Joint Conference on Artificial Intelligence*, pages 393–398, Sydney.

Tversky, A. and Kahneman, D. 1974. Judgment under uncertainty: Heuristics and biases. In Shafer, G. and Pearl, J., editors, *Readings in Uncertain Reasoning*, pages 32–39. Morgan Kaufmann, San Mateo. 1990.

van Fraassen, B. C. 1972. The logic of conditional obligation. *Journal of Philosophical Logic*, 1:417–438.

von Wright, G. H. 1964. A new system of deontic logic. In Hilpinen, R., editor, *Deontic Logic: Introductory and Systematic Readings*, pages 105–120. D.Reidel, Dordecht. 1981.

Winslett, M. 1988. Reasoning about action using a possible models approach. In *Proceedings of the Seventh National Conference on Artificial Intelligence*, pages 89–93, St. Paul.

Winslett, M. 1990. *Updating Logical Databases.* Cambridge University Press, Cambridge.

Index