In presenting this thesis in partial fulfilment of the requirements for an advanced
degree at the University of British Columbia, I agree that the Library shall make it
freely available for reference and study. I further agree that permission for extensive
copying of this thesis for scholarly purposes may be granted by the head of my
department or by his or her representatives. It is understood that copying or
publication of this thesis for financial gain shall not be allowed without my written
permission.

(Signature)
Abstract

We propose sometimes very plausible hypotheses as explanations for an observation, given what we know or believe. In light of new information, however, such explanations may become void. Under this view explanations are defeasible. In this thesis we present a general model of explanation in artificial intelligence based on a theory of belief revision, (or the process of expanding, correcting and contracting existing beliefs) which models the dynamics of belief in a given representation. We take seriously the notion that explanations can be constructed from default knowledge and should be determined relative to a background set of beliefs. Based on the idea that an explanation should not only render the observation plausible but be itself maximally plausible, a preference ordering on explanations naturally arises.

Our model of abduction (the process of inferring the preferred explanations) uses existing conditional logics for default reasoning and belief revision, based on standard modal systems. We end up with a semantics for explanation in terms of sets of possible worlds, which are simple modal structures also suitable for other forms of defeasible reasoning. The result of this thesis is an object-level account of abduction based on the revision of the epistemic state of a program or agent. Abductive frameworks like Theorist, and consistency based diagnosis (today's "canonical" frameworks for diagnosis) are shown to be captured in our work.
# Table of Contents

Abstract

List of Figures

Acknowledgement

1 Introduction

1.1 Explanations and AI .......................................... 2
1.2 Thesis Overview .................................................. 6

2 Prelude on Explanations, Belief Revision and Conditional Logic ............................................. 9

2.1 Logical Preliminaries .............................................. 9
2.2 Approaches to Explanation ........................................ 11
  2.2.1 Philosophical Notions of Explanation ...................... 11
  2.2.2 Can There Exist a Purely Semantic Account for Abduction? ........ 17
  2.2.3 Poole’s Architecture For Default and Abductive Reasoning ...... 19
  2.2.4 Brewka’s Preferred Subtheories .............................. 20
  2.2.5 Consistency Based Diagnosis ................................ 22
2.3 Belief Revision .................................................. 24
  2.3.1 The AGM theory of Revision ................................. 25
  2.3.2 Belief Revision and Default Reasoning ..................... 29
2.4 Modal and Conditional Logics ................................. 30
  2.4.1 Boutilier’s Logics as Calculus for Default Reasoning and Belief Revision ........................................ 31

3 Epistemic Explanations ............................................ 41
List of Figures

2.1 A CO-model (left) and a CT4O-model (right) ........................................ 34
2.2 K-revision model .................................................................................. 36
3.3 Total-order revision model for the rain-sprinkler example, where “rain” and “sprinkler” are rejected in $K$ ................................................................. 48
3.4 Revision model for the rain-sprinkler example, where “rain” is accepted in $K$ ........................................................................................................... 50
3.5 Revision model for the rain-sprinkler example, where “rain” is rejected in $K$ ........................................................................................................... 53
3.6 Preorder-revision model for the bleaching-T-shirt example ...................... 55
4.7 Theorist model for University Students example ..................................... 65
4.8 Theorist model for $D = \{a \supset b, a \supset -b\}$ .................................................... 68
4.9 Brewka model for University Students example ........................................ 72
5.10 Model reflecting expectation of correct behaviour, for two components .... 82
Acknowledgement

I am enormously grateful to my supervisor, Craig Boutilier, who has been a genuine maestro, and a constant source of direction and inspiration. I am also indebted to him for introducing me to modal logic, which has been an invaluable experience.

I want to express my deepest appreciation to David Poole, for being such a supportive and ever-present figure during my program. I also acknowledge his timely and productive review of this thesis.

For their helpful comments and aid with my perverse English I deeply acknowledge Andrew Csinger, Michael Horsch, Richard Dearden and Art Pope.

Unforeseeably, this work had its start at a seminar in belief revision at the University of Buenos Aires. My thanks to Raúl Carnota, Carlos Alchourrón and Adolfo Kvitca.

Above all, I want to thank my loving husband Alejandro, who has been an inexhaustible fount of encouragement, for letting me test my ideas with him, and for making this life-era so pleasant.

Certainly, I acknowledge the funding provided by my supervisor that made this research possible.
Chapter 1

Introduction

If we try to explain why someone who was expected at the meeting did not show up, we may say that he might have been trapped in heavy traffic, or that he simply forgot about the meeting. But we would not be likely to say that he suddenly died. Nor would we say that he was kidnapped by extraterrestrials.

Given an observation we want to determine if a sentence explains why the observation occurred. This thesis is about modeling explanations. It is clear that not any sentence will be considered explanatory. We will elaborate some conditions that an explanation should satisfy with respect to the observation and certain background knowledge or set of beliefs. But as we realize from the above example, not all explanations are equally plausible. While some explanations may seem pretty digestible, others may near absurdity. After all, absurdity is a matter of degree given what we believe. We will model a notion of plausibility of beliefs, and will be able to capture preferences in terms of plausibility among different explanations.

Going back to our example, suppose we adopt the hypothesis that he was delayed in the traffic. If later we learn that his child was being born at the time of the meeting, we will probably give up on the previously embraced explanation. This illustrates the defeasibility of explanations. Much of our knowledge may become void in light of new information; probably only tautologies can be granted indestructible standing. In light of proper counter-evidence, any explanation will fail. We take explanations to be inherently defeasible. We consider them as prima facie reasons for belief in the observation, given what we know. This stresses that explanations are relative to background knowledge. We propose a model of explanation based on a theory of belief revision, or the process.
of modeling expansions, corrections and contractions of beliefs. Another source of the
defeasibility (or nonmonotonicity) of explanations is that they may be framed based on
default knowledge, or expectations about the world as opposed to nomological universals.
The framework proposed in this work seriously contemplates that explanations can be
constructed based on defaults. For instance, defaults can have the form "usually, people
do not suddenly die".

This thesis takes as its starting-point the use of logic for knowledge representation
(KR) and reasoning in artificial intelligence (AI). Our development will be in terms
of conditional logic. We will use existing conditional logics for default reasoning and
belief revision, in particular, Boutilier's modal conditional logics. We inherit a standard
Kripkean possible worlds semantics suitable for other forms of defeasible reasoning like
subjunctive reasoning, belief revision, default reasoning and planning (Boutilier 1992a).
We are able to show formal correspondence between existing abductive frameworks and
our work. This thesis does not address computational methods for explanations.

1.1 Explanations and AI

As David Poole points out¹, when solving a problem in AI, we have to develop a program
that behaves as if it were intelligent. Such a program reasons not about the problem per
se, but from and about a representation of the problem. This is a crucial distinction.
Explanations in AI, then, are relative to the representation of the problem. This is true
for the treatment of explanations in this thesis, and any other AI framework. The notion
of what is an explanation has been widely studied in the philosophical community, and
different authors diverge in their interpretations. It would be nice to start with a generic
definition of an explanation and evaluate how AI captures the idea. But such a general
statement is a difficult enterprise in itself, and we will not pursue it in this thesis. Much
of the work in AI that relates to explanation is referred as abductive reasoning. The term
"abduction", as conceived by Charles Sanders Peirce, refers to the process of formulation

¹Personal communication.
of most plausible explanatory hypotheses. Unfortunately, this term has been specialized in the AI jargon, referring exclusively to some logic-based account. In this thesis, when referring to those logic based-accounts (like Theorist), we will call them “abductive” using the quotes. Otherwise, the term abduction will be used to refer to the inference to the best explanations.

Explanations in this thesis should not be regarded as a natural language construction, but as a logical abstraction, one of the two parts involved in an inferential relation: an hypothesis for some observation, or a cause for an effect, or a reason for some evidence, or a premiss for a conclusion, or more generally, an explanan for an explanandum. There are many AI applications that can be modeled as a form of abductive reasoning. Among them, diagnosis, image interpretation, natural language interpretation, learning. In diagnosis (Peng and Reggia 1990, Reiter 1987, de Kleer Mackworth Reiter 1990, Poole 1989, 1991), given an observation about the system’s behaviour, an explanation (some represented knowledge) of why such behaviour arose is provided by the program. In image interpretation (recognition) (Reiter and Mackworth 1989), given an image, the intelligent program should determine a scene that gave rise to such image. In natural language interpretation (Stickel 1990), (Hobbs 1990), given an utterance, a “story” behind such an expression has to be determined. In learning as a process of theory formation (Morris and O’Rorke 1990) given an observation that contradicts an existing theory, a new theory should be modeled that accounts for the observation. In argumentation, given a particular argument (or its negation), a set of reasons for it should be provided. The range of applications seems to be very large, we add to the above list court deliberation (what set of laws and regulations would lead to a particular sentence), detective reasoning (what set of hypotheses would lead to solving the crime), financial stock exchange (what market scenario may provoke a particular stock status), and economic analysis (how the variables should be combined to reach a particular economic situation).

Any account of explanation ought to resolve how to determine a set of preferred explanations. Some preference criteria is called for. At this point, there seems to be a lack
of consensus. As pointed out by David Poole\textsuperscript{2}, while in natural language understanding the most specific explanation would be preferred, in diagnostic tasks the least presumptive explanation would. In this work we argue for a measure of preference in terms of plausibility. Different applications should characterize what states of affairs are more plausible. Abductive frameworks in AI have not resolved satisfactorily the problem of preferences on explanations. Among the criteria used to capture preferences we found the following. The minimal cardinality criterion imposes that explanations consisting of fewer hypotheses should be preferred over those with a larger number of hypotheses. Peng and Reggia (1990) showed that this criterion may fail to reflect plausibility since an explanation of a symptom consisting of a rare disease will always be preferred over one consisting of two common diseases. The irredundancy criterion determines that supersets of explanations should not be preferred. Reiter's minimal diagnosis is based on this principle. Component failures are modeled as abnormalities. Given an observation that conflicts with the expected correct behaviour of the system, a set of failing components is conjectured to explain the discrepancy. Diagnoses consist of minimal (subset related) abnormalities. Although preferences in terms of irredundancy may reflect plausibility, it is hard to accommodate that the abnormality of one component be preferred over another one.

If numbers are available, probabilistic frameworks (de Kleer and Williams 1987, Poole 1992), may derive meaningful preferences. We can see our work capturing the spirit of probabilistic proposals but in a qualitative fashion.

Some authors like Kurt Konolige (1992a) start from a predefined set of potential explanations, and argue that capturing preferences is just a matter of mathematically expressing a partial order on the subsets of potential explanations. We argue against this approach because it ignores the context dependent nature of explanations with respect to the background knowledge. A fixed ordering of conjectures is problematic because

\textsuperscript{2}Personal communication.
Chapter 1. Introduction

it is implicitly established relative a particular state of affairs (say most normal situations, past experience, or whatever). Hence, if the context changes, the ordering becomes meaningless. But this dependency is not represented in those formalisms. Our approach overcomes these difficulties, and requires no explicit ordering of hypotheses. The preference ordering on explanations will originate from a preference ordering over states of affairs (possible worlds). We will discuss this problem of preferences further in the following chapters. Hector Levesque (1989) showed the impossibility of discriminating preferences among explanations based on any semantic measure when using classical material implication as the connective between explanation and observation. Based on this result, Levesque develops a syntactic measure of “simplicity” to account for preferences among explanations. We will elaborate on his results in Section 2.2.2.

Our account of explanation refrains from using material implication as the connective. In contrast, we will propose a conditional connective, which is not truth-functional. A natural account of preferences in terms of plausibility becomes possible. At another frontier, in most existing AI frameworks for abductive reasoning, if the observation is inconsistent with given facts or background knowledge, then it is unexplainable. Those frameworks do not contemplate the possibility of explaining counterfactual observations. For instance, consider the defaults “brunettes get a bronze tan” and “blonds do not get a bronze tan”, and suppose as a fact that “Fred is blond, and his brother is brunette”. The counterfactual observation of Fred having a bronze tan is not explainable, since the candidate explanation of Fred being brunette as his brother conflicts with the facts. We will propose an account of explanations where hypothetical situations be explained, by being able to reason counterfactually.

Some AI approaches to explanation/causation start from a causal theory, like Kolenlimage's default causal networks (1992b) and bayesian nets (Pearl 1988), and are intended and restricted to derivation of causal explanations. In order to appreciate the explanations arising from this thesis, we should mention that we take explanation as a broader notion than causation. Causation is associated with an asymmetrical relation between
cause and effect that can be detected by appealing to controllability of phenomena (manipulating the effect will not change the cause). In the concluding chapter of this thesis we discuss the idea of modeling causality as an extension of our framework.

1.2 Thesis Overview

In Chapter 2 we will introduce background material. The subsequent chapters will build on the concepts of this chapter. Motivating our work we will review some work from the philosophical literature. We will examine Peirce, Hempel, Quine and Ullian and Gärdenfors's work. Then we will turn to existing approaches to explanation from the AI perspective. In particular, we present the Theorist architecture for default and abductive reasoning, and consistency based diagnosis. We will briefly review the work of Alchourrón, Gärdenfors and Makinson on belief revision, on which we will base our analysis of explanation. We will then study Boutilier's modal conditional logics, and their role in default reasoning and belief revision, emphasizing the possible worlds semantics of these logics.

Chapter 3 will present our model of explanation based on the revision of the epistemic state of a program or ideal agent. We will consider a background epistemic state representing an agent's beliefs. Our main concern will be to develop a general account of abduction based on principled semantics and expressible in a logical calculus. We will model explanations relative to some background knowledge, respecting their inherent defeasibility. We will provide the capability to derive preferences in terms of plausibility, and explain factual and hypothetical observations.

What is to be explained (the explanandum) can either be accepted, rejected or indeterminate in the background epistemic state. We will discuss the appropriate epistemic status an explanation should have for each possible status of the explanandum. When the program (or ideal agent) receives an inquiry about an explanandum that is already believed (or accepted) by the program, we will argue that an explanation should also be believed.
Chapter 1. Introduction

We will analyze explanations via hypothetical changes in belief. Our strongest notion of explanation will impose that (hypothetical) belief in the explanation should result in full belief of the explanandum. We will name them *predictive explanations*. We will then explore alternative weaker notions of explanation. These will be characterized by the condition that acceptance of an explanation should make the explanandum possible. Differing in their explanatory power, will end up with a family of explanations referred as "epistemic explanations".

Via the Ramsey test for acceptance of conditionals we will be able to translate our characterization of explanations in terms of changes in belief, into subjunctive conditionals. Explanations will be expressed as subjunctive conditionals relative to the agent’s epistemic state.

Based on the idea that an explanation should not only make the explanandum plausible but the explanation itself be maximally plausible, a preference ordering on explanations will naturally arise. We will discuss some *pragmatic aspects* of explanation, showing that semantic criteria may fall short in order to shape a *simplest* or *most informative* sentence as an explanation for a given observation.

In Chapter 4 we will take Theorist as representative of "abductive" approaches to explanation. By interpreting the background epistemic state as a theory of expectations, we will show how Theorist can be recast in our framework. Theorist explanations will be modeled as conditional sentences. The correspondence with epistemic explanations will be drawn. Then we will explore Brewka’s extension of Theorist, that provides a prioritized setting. We will elaborate on the effect of priorities over defaults when used abductively. We will propose how to augment Theorist and also Brewka’s extension in order to provide predictive explanations, and also preferences among explanations.

Chapter 5 works out the correspondence between consistency based diagnosis and our model of explanation. By interpreting the expectation of correct behaviour the system as a default of component working normally, we will map the consistency based diagnosis account on a Theorist model. By mapping the consistency based and the “abductive”
Chapter 1. *Introduction*

Theorist framework on the same kind of model, we will be able to highlight connections between the two.

Finally, in Chapter 6 we will discuss the accomplishments of this thesis and propose several avenues for further research, indicating different ways in which our account of explanation can be generalized or extended.
Chapter 2

Prelude on Explanations, Belief Revision and Conditional Logic

In this chapter, we will lay the necessary foundations to later present our semantic account of explanations. We will review some existing approaches to explanations, mainly to motivate this work. We will first concentrate in some philosophic perspectives, reviewing the ideas of Peirce (1923), Hempel (1966), Gärdenfors (1988), Quine and Ullian (1978) and Lewis (1973).

Then we will examine some qualitative AI approaches to explanation, concentrating in Poole’s architecture for default and abductive reasoning (1989), and consistency based diagnosis (Reiter 1987; de Kleer, Mackworth, Reiter 1990). In Chapters 4 and 5 these two different accounts will be recast in our framework.

We will briefly review the work of Alchourrón, Gärdenfors and Makinson (AGM) on belief revision, upon which we will base our development of explanation. Then we will look over some modal and conditional logics, in particular Boutilier’s logics, emphasizing their ability to handle defeasible reasoning and belief revision. We will study how Boutilier’s conditional logics provide a logical calculus and a unified Kripkean semantics for the AGM theory of revision and default reasoning. This section will be particularly important as a preliminary to the chapters to follow.

2.1 Logical Preliminaries

We will refer with CPL to classical propositional logic. We take \( \mathbf{P} \) to denote a denumerable set of atomic variables and \( \mathbf{LCPL} \) to denote the propositional language over this set, based on classical propositional logic with the connectives \( \wedge, \vee, \neg, \supset \). Capital letters \( A, B, C \), and greek letters \( \alpha, \beta, \gamma \) will be used as variables over propositional formulae.
Often the letters $A, B, C$ will just denote propositional variables or atoms. Words such as *rain* will be used in examples and indicate atomic relation symbols. Lower-case letters $r, s, q$ will be used similarly. The symbols $\top, \bot$ denote truth and falsity respectively. Letters $S, T, X, Y$ will be used for sets. We will also consider a bimodal language $L_B$ based on $L_{CPL}$ augmented with two modal operators $\Box$ and $\Diamond$. The sentence $\Box \alpha$ indicates "$\alpha$ is true at all accessible worlds", while $\Diamond \alpha$ indicates "$\alpha$ is true at all inaccessible worlds". Boutilier's family of bimodal conditional logics are defined over $L_B$.

The symbol $\vdash$ is used to indicate derivability in different systems, using a subscript to specify the system; so $\vdash_{CPL}$ denotes derivability in CPL, $\vdash_{CT40}$ denotes derivability in CT40 (Boutilier 1992a). We will use $Cn$ to denote the logical consequence operation in the system we are working with.

We refer to propositional valuations as possible worlds or states of affairs. We will use the letters $w, v, u$ to designate possible worlds. For any sentence $\alpha$ in $L_{CPL}$, we will denote with $\mathcal{H}(\alpha)$ the proposition expressed by $\alpha$ (the set of possible worlds where the sentence $\alpha$ is true). Usually the letters $V, W$ will be used to indicate sets of possible worlds.

Semantic entailment will be denoted by the symbol $\models$, using subscripts to indicate entailment in different classes of models. Also $\models$ will be used to indicate satisfaction of formulae more generally. We will write $w \models \alpha$ if $w$ is a world that satisfies $\alpha$, meaning that $w$ assigns truth to $\alpha$. When $w \models \alpha$, we will say that $w$ is an $\alpha$-world. If $\alpha$ is a valid formulae, we will write $\models \alpha$. We will also write $w \models X$ for a set $X$, meaning that world $w$ satisfies each element of $X$.

Motivated by clarity concerns, we will write a conditional sentence $X \Rightarrow \beta$ when $X$ is a finite set, meaning that the conjunction of the elements of $X$ conditionally entail a sentence $\beta$. Ditto for a a conditional sentence based on the connective $\rightarrow$.

We will refer to the following properties of binary relations. Let $X$ be a set, and $R$ be a binary relation over elements of $X$.

- $R$ is reflexive in $X$ if and only if for all $x \in X : xRx$.
- $R$ is antisymmetric in $X$ if and only if for all $x, y \in X : xRy$ and $yRx$ only if $x = y$. 
Chapter 2. Prelude on Explanations, Belief Revision and Conditional Logic

R is transitive in $X$ if and only if for all $x, y, z \in X$: if $xRy$ and $yRz$ then $xRz$
R is connected in $X$ if and only if for all $x, y \in X$ if $x \neq y$, then $xRy$ or $yRx$
R is totally connected in $X$ if and only if for all $x, y \in X$: $xRy$ or $yRx$

The relation $R$ is a preorder on $X$ if and only if $R$ is reflexive and transitive. $R$ is a partial order on $X$ if and only if $R$ is reflexive, transitive and antisymmetric. $R$ is a total order on $X$ if an only if $R$ is reflexive, transitive and totally connected.

2.2 Approaches to Explanation

2.2.1 Philosophical Notions of Explanation

Peirce's notion of Abduction

In his “Chance, Love and Logic” (1923), Charles Sanders Peirce discusses the nature of deduction, induction and abduction\(^1\) as three different forms of reasoning. He starts with a syllogism to exemplify deductive inference. He explains that the major premise indicates a rule, as Rule.— *All men are mortal*, and the minor premise indicates a case under the rule, as Case.— *Enoch was a man*. The conclusion applies the rule to the case and states the result: Result.— *Enoch is mortal*. He asserts “all deduction is of this character; it is merely the application of general rules to particular cases.” Peirce shows how by inverting the deductive syllogism, two forms of *synthetic* inference are obtained, *hypothesis* and *induction*. He explains that making an hypothesis is the inference of a case from a rule and result; and describes induction as the inference of the rule from the case and result. Peirce exemplifies the three kinds of inference as follows (page 134).

**DEDUCTION**
Rule.— All the beans from this bag are white.
Case.— These beans are from this bag.
\(\therefore\) Result.— These beans are white.

**INDUCTION**
Case.— These beans are from this bag.

---
\(^1\)In these essays Peirce refers to abduction with the term “hypothesis”; years later he renamed this form of reasoning as “abduction”.

Result.— These beans are white.
\[\therefore\ \text{Rule.} \quad \text{All the beans from this bag are white.}\]

\textbf{HYPOTHESIS}

Rule.— All the beans from this bag are white.
Result.— These beans are white.
\[\therefore\ \text{Case.} \quad \text{These beans are from this bag.}\]

He considers hypotheses as explanatory suppositions (page 135):

"Hypothesis is where we find some very curious circumstance, which would be explained by the supposition that it was a case of a certain general rule . . . Numberless documents and monuments refer to a conqueror named Napoléon Bonaparte. Though we have not seen the man, yet we can not explain what we have seen, namely all these documents and monuments, without supposing he really existed."

Peirce calls \textit{abduction} to the process of formulation and selection of hypotheses in scientific inquiry. As Nicholas Rescher (1978) puts it, abduction accounts for the process of "conjectural proliferation of explanatory hypothesis that merit closer scrutiny". Peirce describes abduction as the suggestion of a theory that would render necessary (a strong sense of explanation) some surprising phenomena. But, as Rescher explains, "conjectural fancy is limitless"; then, some guidance for effective theorizing is needed. This guidance is given by Peirce’s concept of plausibility. These quotes are extracted from (Rescher 1978, page 44).

"Every inquiry whatsoever takes its rise in the observation . . . of some surprising phenomenon, some experience which either disappoints an expectation, or breaks in upon some habit of expectation . . . At length a conjecture arises that furnishes a possible Explanation . . . On account of this Explanation, the inquirer is led to regard his conjecture or hypothesis, with favor. As I phrase it, he provisionally holds it to be ‘Plausible’; this acceptance ranges in different cases -and reasonably so- from a mere expression of it in the interrogative mood, as a question meriting attention and reply, up through appraisals of Plausibility, to uncontrollable inclination to believe. (CP, 6.469 [1908])

. . . By plausibility, I mean the degree to which a theory ought to recommend itself to our belief . . ."(CP, 8.223 [c. 1910])
Other Philosophical Work on Explanations

The problem of explanations has been deeply studied by philosophers. Hempel (1966) defines two basic requirements for scientific explanations: the requirement of explanatory relevance and the requirement of testability. The former guarantees that the explanation be related with the essential of the question. The latter says that an explanation should be a "risky" or "non trivial" (in the sense of (Popper, 1969)) assertion of why a given observation occurred. An explanation consists of an *explanandum* sentence, the sentence describing the phenomenon that needs to be explained (observation), and *explanan* sentences, which specify explanatory information that account for the observation. Hempel defines his *deductive nomological explanations* as formed by explanan sentences consisting of general laws and assertions about particular facts. Thus, in deductive-nomological explanations, explanans deductively imply the explanandum, satisfying the two requirements for scientific explanation in the strongest sense.

Gärdenfors (1988, Chapters 8 and 9, analysis of explanation and causation) requires that explanations be evaluated relative to a background theory or epistemic circumstances; that is, the connection between explanan and explanandum should be given relative to an epistemic state. Gärdenfors's view of including the role epistemic circumstances contrasts with Hempel's deductive nomological explanations.

The central idea in Gärdenfors' analysis is that explanations should decrease the degree of surprise of the explanandum in a non-trivial way. He requires the explanandum sentence to be accepted in the epistemic state where the inquiry arises, while explanations should not. The power of an explanation is given by its ability to decrease the degree of surprise of the explanandum. Accepting an explanation should make the explanandum less surprising relative to the (possibly hypothetical) epistemic state where the explanandum is indeterminate. The belief value of the explanandum in this state is compared to the belief value of the explanandum in the hypothetical state resulting after learning an explanation.

Let's see an example. If we wonder "why Victoria is tanned although it is winter
here”, the answer that “she has recently spent a week in the Canary Islands” is an explanation satisfying Gärdenfors’ conditions. Suppose we didn’t know that Victoria is tanned, then, being winter here, Victoria’s tan becomes less surprising if we learn that she spent a week in the Canary Islands.

Gärdenfors illustrates how his conditions on explanations account for “spurious” explanations. Basing his model on the dynamics of epistemic states, what has been an explanation relative to an epistemic state may cease to be an explanation in a state containing more knowledge. We regard spurious explanations as characterizing the “non-monotonicity” or defeasibility of explanations (page 184).

“...For example, I may be surprised by the fact that Victoria is tanned in the month of February. If I ask for an explanation and get the answer that Victoria has recently spent a week on the Canary Islands, I accept this as an explanation because I know that most people who have recently been on the Canary Islands are tanned. However, if I later learn that it had been raining on the Canary Islands the week Victoria was there, the old explanation is no longer satisfactory, and I once more lack information of why she is tanned.”

Our work (to be presented in the next chapter) will strongly follow Gärdenfors’ view in that we will consider explanations relative to an epistemic state. However, having different ideas of explanation in mind, our account will differ from his. Gärdenfors’s analysis is formulated in a probabilistic model of epistemic states, while ours will be based on belief sets. More important is that we have a different perspective with respect to what are necessary conditions for explanation. His view is suitable for scientific inquiry, while ours may not be. We will carefully discuss this issue in Chapter 3.

Quine and Ullian’s (1978) conceive explanations as plausible hypotheses. Our model of explanation will allude to their view.

(page 66) “...What we try to do in framing hypotheses is to explain some otherwise unexplained happenings by inventing a plausible story, a plausible description or history of relevant portions of the world.”

They discuss five “virtues” that an hypothesis may enjoy in varying degrees. The first
three virtues account for the plausibility of hypotheses by appealing to their “believability” given our background beliefs. Virtue I is *conservatism* or *compatibility with previous beliefs*.

(page 66) “Acceptance of a hypothesis is of course like acceptance of any belief in that it demands rejection of whatever conflicts with it. The less rejection of prior beliefs required, the more plausible the hypothesis —other things being equal.

...our friend the amateur magician tells us what card we have drawn. How did he do it? Perhaps by luck, one chance in fifty-two; but this conflicts with our reasonable belief, if all unstated, that he would not have volunteered a performance that depended on that kind of luck. Perhaps the card were marked; but this conflicts with our belief that he had no access to them, they being ours. Perhaps he peeked or pushed, with a help of a sleight-of-hand; but this conflicts with our belief in our perceptiveness. Perhaps he resorted to telepathy or clairvoyance; but this would wreck havoc with our whole web of belief. The counsel of conservatism is the sleight-of-hand”

Virtue II is *modesty*, stating one hypothesis more modest than another when logically weaker. The more modest hypothesis will be implied by the other without implying it. Another (rather different) characterization of a modest hypothesis is as assuming events of a more usual and familiar sort, hence more to be expected. *Simplicity* makes virtue III, which strives for avoiding gratuitous complications in hypotheses. Virtue IV, *generality*, represents the quality of having a wide range of application. In other words, generality fights against the *ad hoc*. The fifth virtue is *refutability*, and defends hypotheses from being insusceptible of confirmation and useless for prediction.

(page 79) “...some imaginable event, recognizable if it occurs, must suffice to refute the hypothesis. Otherwise, the hypothesis predicts nothing, is confirmed by nothing, and confers upon us no earthly good beyond perhaps a mistaken piece of mind.”

Having analyzed the five virtues of an hypothesis, Quine and Ullian present the idea of explanatory hypotheses as “predictive explanations”.

(page 108) “The immediate utility of a good hypothesis is as an aid to prediction. ...The relation between an explanatory hypothesis and what it explains seems
somewhat like implication; at any rate it transmits plausibility. That is, if someone believed the hypothesis to begin with, he should thereby be inclined to believe also in what it purports to explain.”

The spirit of our work is very much related to Lewis’ (1973, 1977) counterfactual analysis of causation. From a philosophical perspective, causation and explanation are very related but two different problems. Causal explanations are a special case of explanations. Lewis (1973) argues that regularity analysis of causation suffers from confusing causation with other causal relations such as epiphenomena, secondary effects and preempted causes. He proposes a counterfactual analysis of causation that attacks these problems, based on the following counterfactual statement: Had the cause been absent, its effects -some of them at least and usually all- would have been absent as well. Lewis defines causal dependence among single events as two counterfactual dependences: If the cause (C) had not been, then the effect (E) never had existed, (taking > as a “generic” counterfactual connective, we shall write \( \neg C > \neg E \)). And, Had the cause occurred, the effect would have occurred as well (C > E). Then, causation is defined as “causal dependence among actual events”: C is a cause for E if and only if both are part of the actual state of affairs, and moreover, \( \neg C > \neg E \).

Lewis’s work has been criticized for not handling cases of multiple-causation. While one cause could have been absent, another cause could have occurred, thus neglecting the counterfactual condition: Had the cause been absent, the effect had been absent as well. We conclude that this counterfactual dependence is definitely inadequate for explanations, since we may expect multiple explanations for an inquiry. However, our analysis of predictive explanations will share with Lewis’ work the other counterfactual dependence, which we phrase as: Were an explanation believed, so too would the observation. Jaegwon Kim (1973) points out a number of weaknesses in Lewis’ treatment of causation, arguing that not only is causal dependence captured by counterfactual dependence, but also analytical or logical dependence and other kinds of determination. Here are two of his examples. The counterfactual assertion “If yesterday had not been Monday, today would not be Tuesday” manifests a logical or analytical dependence, but not causal. “If
my sister had not given birth at time t, I would not have become an uncle at time t.” manifests determination but not causal dependence. We conclude that logical or analytical dependencies can be acceptable explanations. According to Lewis, causation is a transitive relation that occurs among actual events. But explanations are not necessarily transitive. Just consider any of our famous examples in AI. Adults are usually employed, University students are usually adults, but University students are usually unemployed. Being a university student explains Fred being an adult. Being an adult is an explanation for being employed; however, being a university student is not a reasonable explanation for being employed. In our concluding chapter we will briefly discuss the extension of our framework to account for causality.

2.2.2 Can There Exist a Purely Semantic Account for Abduction?

From an AI perspective this is a fundamental question. Hector Levesque (1989), defining abduction in terms of a model of belief, demonstrates the impossibility of sorting out preferred (or say, most plausible) explanations, and discusses the need of a syntactic measure of simplicity to do it. Levesque sees abduction as inference from what we are trying to explain to the simplest (or most preferred or plausible) explanations, relative to a particular model of belief (so that different models of belief give rise to different forms of abductive reasoning). However, he commits to a truth functional connective (material implication) to model the logical relation between explanation and observation.

Levesque assumes a logical language $\mathcal{L}^*$ with a modal operator for belief $B_\lambda$, and an objective sublanguage $\mathcal{L}$. Atomic sentences of $\mathcal{L}^*$ are of the form $B_\lambda \alpha$, where $\alpha$ is a sentence of $\mathcal{L}$. $B_\lambda \alpha$ says that $\alpha$ is a belief of type $\lambda$, $e \models B_\lambda \alpha$ says that $B_\lambda \alpha$ is true at epistemic state $e$. Levesque defines an explanation for a sentence $\beta$ with respect to an epistemic state $e$, for a type of belief $\lambda$ as follows:

\[ e \models [B_\lambda (\alpha \supset \beta) \land \neg B_\lambda \neg \alpha] \]

Explaining $\beta$ is accomplished by finding all those “simplest” $\alpha$ (given a syntactic measure of simplicity) satisfying the above definition. Levesque presents the following example.
Chapter 2. Prelude on Explanations, Belief Revision and Conditional Logic

"Suppose we know that male and (hepatitis ⊃ jaundice) are both true. If we observe jaundice in a patient, we might be interested in determining what might explain it, based on what we know. In this case the answer is clearly hepatitis, but it is not obvious how to characterize in general the answers we are looking for.

...in terms of propositions there will be propositions that are logically-too-strong and others that are logically-too-weak. For instance, (hepatitis ∧ migraines) accounts for the jaundice in that it is consistent with what is known and if it were true, then jaundice would be too. Similarly, (hepatitis ∨ ¬male) accounts for jaundice since it too is consistent with what is known, and if it were true, then jaundice would be also, since male is known to be true. Yet (hepatitis ∧ migraines) implies hepatitis which implies (hepatitis ∨ ¬male)."

Then he shows that the proposition denoting hepatitis can not be distinguished from these other propositions.

"To see this, suppose the contrary that there were a function $F$ that given a proposition expressed by $(α ⊃ β)$ and the one expressed by $β$ would always return the one expressed by $α$. That is, suppose that for every $α$ and $β$, $F(||(α ⊃ β)||, ||β||) = ||α||$. Then, we would have $F(||(q ⊃ q)||, ||q||) = ||q||$, and $F(||(⊥ ⊃ q)||, ||q||) = ||⊥||$. However, $||q ⊃ q|| = ||(⊥ ⊃ q)||$ since the two sentences are logically equivalent. But this implies that $||⊥|| = ||q||$, which is incorrect. Such a function $F$ cannot exist, and we are forced to go beyond the logic of the sentences (that is beyond the propositions expressed) to differentiate hepatitis from other potential explanations."

We take Levesque's demonstration of the non-existence of such a function $F$ as an indication of the inappropriateness of material implication to represent explanations. This actually motivates our account for explanations which will not be based on a truth functional connective, allowing for preferences in terms of plausibility, and capturing them purely logically. Still, "pragmatic" concerns with respect to the usefulness of explanations may call for some syntactic measure of simplicity. We will elaborate on this in Chapter 3.

In Levesque's account, explanations can not be disbelieved in the background epistemic state. This is necessary since the link between explanation and explanandum is material implication. As a consequence, a sentence that is rejected in the background epistemic state is not explainable. We aim for a more general approach to explanation that permits the explanation and explanandum be hypothetical beliefs.
2.2.3 Poole’s Architecture For Default and Abductive Reasoning

Poole (1988, 1989, 1990), in his pioneer Theorist framework for default and abductive reasoning, considers two activities, explaining observations and predicting what is expected to be true. Explanations are modeled as arising from a process of theory formation; given some observations, the system constructs a theory which would explain those observations. Poole assumes a standard first order language. An instance of a formula is a substitution of terms in this language for free variables in the formula. Theorist framework is elegantly defined in terms of two sets of formulae. The set of facts \( \mathcal{F} \), which are taken to be true of the domain, and the set of “possible hypotheses” \( \mathcal{H} \). The pool of hypotheses \( \mathcal{H} \) consists of two sets, defaults \( \mathcal{D} \) and conjectures \( \mathcal{C} \). Defaults in \( \mathcal{D} \) are considered “normality assumptions” that can be assumed to be true given no evidence to the contrary. Defaults can be used in prediction and explanation. \( \mathcal{C} \) is the set of conjectures, which are considered “abnormality assumptions”. Conjectures are possible hypotheses which can be used only for explaining observations.

We will refer with the pair \( \langle \mathcal{F}, \mathcal{H} \rangle \) to a default theory. We will slightly modify Poole’s notation in the next definitions. A scenario is defined as a set of hypotheses that could be true based on the given facts.

**Definition 2.1 (Poole 1989)** A scenario of a default theory of \( \langle \mathcal{F}, \mathcal{H} \rangle \) is a set \( H \) of ground instances of elements of \( \mathcal{H} \) such that \( H \cup \mathcal{F} \) is consistent.

An extension of a default theory is a maximal scenario.

**Definition 2.2 (Poole 1989)** An extension of \( \langle \mathcal{F}, \mathcal{D} \rangle \) is the set of logical consequences of \( \mathcal{F} \) together with a maximal (with respect to set inclusion) scenario of \( \langle \mathcal{F}, \mathcal{D} \rangle \).

Predictions are skeptical conclusions from a default theory that are expected to be true given the facts, no matter which defaults are assumed. Predictions are based on the normality assumptions or expectations about the world, that are consistent with the facts. Conjectures do not participate in forming predictions.
Definition 2.3 (Poole 1989) \( \gamma \) is predicted iff \( \gamma \) is in all extensions of \( \langle F, D \rangle \).

Poole defines an explanation of an observation as a scenario of the default theory that gives rise to the observation.

Definition 2.4 (Poole 1989) If \( \beta \) is a closed formula, an explanation of \( \beta \) is a scenario that together with \( F \) implies \( \beta \).

So, an explanation for an observation \( \beta \) from \( \langle F, H \rangle \) is formed by sets \( C \) and \( D \) instances of elements of \( C \) and \( D \) respectively, such that

\[
F \cup C \cup D \models \beta, \text{ and } F \cup C \cup D \text{ is consistent}
\]

then, \( C \cup D \) is an explanation of \( \beta \). Notice that depending on the default theory, even contradictory explanations can be drawn for a given observation. For example, let \( \langle F, H \rangle \) such that \( F = \{ \} \) and \( H = D \cup C \), \( D = \{ (A \wedge Q) \supset B, (\neg A \wedge R) \supset B \} \), \( C = \{ A, \neg A, Q, R \} \). \( A \wedge Q \) is an explanation for \( B \) and so is \( \neg A \wedge R \). Notice also that for some default theories, the same explanation can explain contradictory conclusions. For example, let \( \langle F, H \rangle \) such that \( H = D \cup C \), where \( F = \{ \} \), \( D = \{ A \supset B, A \supset \neg B \} \), and \( C = \{ A \} \). \( A \) explains \( B \) and and also \( A \) explains \( \neg A \).

The following theorem demonstrates that whenever an observation is explainable, there is an explanation consisting of a maximal set of defaults.

Theorem 2.1 (Poole) There is an explanation of \( \beta \) iff \( \beta \) is in some extension.

2.2.4 Brewka’s Preferred Subtheories

Gerhard Brewka (1989) extends Poole’s Theorist framework by introducing priorities among defaults. To preserve an homogeneous notation we will adapt Brewka’s notation to make it consistent with our exposition of Theorist. Brewka presents his framework for default reasoning considering a set of defaults with different degrees of reliability, allowing to express that one default may have priority over another one. That is, defaults have
the ability to block other defaults. Brewka does not isolate facts from defaults, arguing that facts can always be expressed in the rank of the most reliable formulae. Without loss of generality, we will introduce a set $\mathcal{F}$, possibly empty, of sentences of true facts, that we consider non-defeasible. We then define a *Brewka theory* as follows.

**Definition 2.5** A *Brewka theory* is a pair $\langle \mathcal{F}, \mathcal{B} \rangle$ such that $\mathcal{F}$ is a set of true facts, and $\mathcal{B} = (B_1, B_2, \ldots B_n)$ is a set of defaults where the rank $B_i$ is more reliable than $B_j$ iff $i < j$, and no rank of $\mathcal{B}$ needs to be consistent.

A Theorist default theory (with no conjectures) $\langle \mathcal{F}, \mathcal{D} \rangle$ corresponds to a Brewka theory $\langle \mathcal{F}, \mathcal{B} \rangle$ when $\mathcal{B}$ possesses a unique level of reliability, i.e. $\mathcal{B} = (\mathcal{D})$. Brewka defines a preferred subtheory as the counterpart of an extension. We will adapt Brewka's definition of a preferred subtheory to allow for the facts $\mathcal{F}$ calling it an extension.

**Definition 2.6** Let a Brewka theory $\langle \mathcal{F}, \mathcal{B} \rangle$, such that $\mathcal{B} = (B_1, \ldots, B_n)$. Then $S = \mathcal{F} \cup S_1 \cup \ldots S_n$ is an *extension* of $\langle \mathcal{F}, \mathcal{B} \rangle$ iff for all $k$ such that $1 < k < n$, $S = \mathcal{F} \cup S_1 \cup \ldots \cup S_k$ is a maximal consistent subset of $\mathcal{F} \cup B_1 \cup \ldots \cup B_k$.

In case of conflicting evidence, a Brewka theory will generate multiple extensions in the same way Theorist would. Brewka defines two notions of provability from a default theory. Strong provability corresponds to containment in all extensions (preferred subtheories), while weak provability corresponds to containment in some extensions. Boutilier (1992f) studied a notion of entailment for Brewka theories, corresponding to strong provability, that he called $B$-entailment. We will rename $B$-entailment as *Strong $B$-entailment*, and define *Weak $B$-entailment* as the counterpart for weak provability. We should specify how premises interact with a Brewka theory. We take a premiss to be a consistent sentence that should not be violated. Now extensions should contain a base set formed by the facts $\mathcal{F}$ together with some premiss $\alpha$.

**Definition 2.7** Let $\mathcal{B} = (B_1, \ldots, B_n)$. $S = (\mathcal{F} \cup \alpha) \cup S_1 \cup \ldots \cup S_n$ is an *extension* of $\langle (\mathcal{F} \cup \alpha), \mathcal{B} \rangle$ iff for all $k$ such that $1 \leq k \leq n$, $S = (\mathcal{F} \cup \alpha) \cup S_1 \cup \ldots \cup S_k$ is a maximal consistent subset of $(\mathcal{F} \cup \alpha) \cup B_1 \cup \ldots \cup B_k$. 
Chapter 2. Prelude on Explanations, Belief Revision and Conditional Logic

Definition 2.8 \( \beta \) is Brewka-Strongly entailed by \( \alpha \) with respect to \( (F, B) \) (written \( F \cup \alpha \vdash_{BS} \beta \)) iff \( \beta \) is entailed by all extensions of \( ((F \cup \alpha), B) \).

Brewka-Strong entailment corresponds to the notion of predictions.

Definition 2.9 \( \beta \) is Brewka-Weakly entailed by \( \alpha \) with respect to \( (F, B) \) (written \( F \cup \alpha \vdash_{BW} \beta \)) iff \( \beta \) is entailed by some extension of \( ((F \cup \alpha), B) \).

### 2.2.5 Consistency Based Diagnosis

The diagnostic task is to determine why a correctly designed system is not functioning as it was intended. Raymond Reiter's (1987) theory of diagnosis from first principles relies on the so called “Principle of Parsimony”: If we don't know of any deviations from normal behaviour, we will assume that every component is working normally. If we know of deviations from normal behaviour, we should presume as few faults as possible.

Reiter starts with a description of the system (SD) that specifies how the system normally behaves on the assumption that all components (COMPS) are working correctly. If an observation (OBS) is logically inconsistent with the system description, then not all components can be working correctly. A diagnosis is called for. A set of components assumed to be functioning abnormally should explain the discrepancy. But according to the parsimony principle, such set of abnormal components is assumed to be minimal.

The formalism assumes a predicate \( ab() \) that applies to components to denote that they are functioning abnormally.

**Definition 2.10 (Reiter)** A *diagnosis* for \( (SD, COMPS, OBS) \) is a minimal set \( \Delta \subseteq COMPS \) s.t. \( SD \cup OBS \cup \{ab(c) | c \in \Delta \} \cup \{\neg ab(c) | c \in COMPS - \Delta \} \) is consistent.

The following are consequences of the definition.

**Proposition 2.2 (Reiter)** A diagnosis exists iff \( SD \cup OBS \) is consistent.

As a result, if the observation is inconsistent with the system description then the observation is not explainable.
Proposition 2.3 (Reiter) $\Delta = \emptyset$ is a diagnosis (and the only diagnosis) for $(SD, COMPS, OBS)$ iff $SD \cup OBS \cup \{\neg ab(c) | c \in COMPS\}$ is consistent.

Then, if the observation does not conflict with normal behaviour of the system, we conclude that all components are behaving correctly.

Proposition 2.4 (Reiter) $\Delta$ is a diagnosis iff for each $c_i \in \Delta$, $SD \cup OBS \cup \{\neg ab(c) | c \in COMPS - \Delta\} \models ab(c_i)$

The faulty components are determined by the normal components in $COMPS - \Delta$, given the system description and the observation.

Reiter gives the following correspondence of minimal diagnosis in default logic. The predicted behaviour of the system given a diagnosis is denoted by $\Pi$. A set of defaults expresses that each component should be acting normally, unless it is inconsistent to assume so.

Theorem 2.5 (Reiter) $E$ is an extension for the default theory

$(\{ \frac{\top}{\neg ab(c)} | c \in COMPS \}, SD \cup OBS)$ iff for some diagnosis $\Delta$ for $(SD, COMPS, OBS)$, such that $SD \cup OBS \cup \{ab(c) | c \in \Delta\} \cup \{\neg ab(c) | c \in COMPS - \Delta\} \models \Pi$, $E \models \Pi$.

De Kleer, Mackworth Reiter (dKMR) (1990) showed that when the system description models exclusively the correct behaviour (that is, it includes no fault models nor exoneration axioms), assuming a superset of the abnormal components in $\Delta$ leads also to a diagnosis, although not a minimal one. So, specifying a minimal diagnosis, all diagnoses were automatically characterized. However, if fault models and/or exoneration axioms are included in $SD$, then a superset of the abnormal components in a diagnosis may not lead to a diagnosis. De Kleer, Mackworth and Reiter (1990) proposed a more general definition of a consistency based diagnosis that contemplates that the system description may include fault models and/or exoneration axioms. A diagnosis is denoted by a conjunction of $ab$-literals, which explicitly indicates whether each component is normal.
or abnormal. The set of components COMPS is partitioned in two sets \( C_p \) and \( C_n \). 
\( D(C_p, C_n) \) is defined as the conjunction:

\[
\left( \bigwedge_{c \in C_p} ab(c) \right) \land \left( \bigwedge_{c \in C_n} \neg ab(c) \right)
\]

**Definition 2.11 (dKMR)** Let \( \Delta \subseteq \text{COMPS} \). A diagnosis for \((SD, \text{COMPS}, \text{OBS})\) is \( D(\Delta, \text{COMPS} - \Delta) \) such that \( SD \cup \text{OBS} \cup \{ D(\Delta, \text{COMPS} - \Delta) \} \) is satisfiable.

This definition characterizes *all* diagnosis for a given observation. dKMR define a *minimal* diagnosis for an observation as a diagnosis that assumes a minimal (subset related) set of abnormal components. This definition corresponds precisely to Reiter's definition of diagnosis (Definition 2.10).

**Definition 2.12 (dKMR)** A diagnosis \( D(\Delta, \text{COMPS} - \Delta) \) is a *minimal diagnosis* iff for no proper subset \( \Delta' \) of \( \Delta \) is \( D(\Delta', \text{COMPS} - \Delta') \) a diagnosis.

They show that whenever there is a diagnosis \( D(\Delta, \text{COMPS} - \Delta) \), there is a minimal diagnosis, \( D(\Delta', \text{COMPS} - \Delta') \), such that \( \Delta' \) is minimal \( \Delta' \subseteq \Delta \).

### 2.3 Belief Revision

When solving problems in AI we start from a representation of a problem. But such a representation may only be applicable if we can understand and model how to update the representation in light of new information. A theory governing the *dynamics* of the given representation is required. Theories of *belief revision* are relevant to AI addressing this issue.

We aim for a theory of explanation that is relative to a background knowledge or epistemic state of a program. Such a notion of explanation is defeasible in the sense that as the background epistemic state changes, certain explanations may no longer hold. The epistemic state of a program or agent is expected to be in constant change, reflecting the diverse inputs from the world. Then an account of explanation respecting
its defeasible nature shall be based on a theory of belief revision. Our cornerstone on explanations will be given by the following notion: "were the explanation believed, the explanandum should become plausible". To evaluate this conditional relative to the background epistemic state, some hypothetical change in belief must operate. For these reasons we will base our account of explanation on a theory of belief revision. We will talk about a knowledge base or a belief set indistinguishably. No differences should be ascribed to our use of the terms of knowledge and belief.

Different approaches have been proposed to capture the process of revising beliefs. We will concentrate in the work of Alchourrón, Gärdenfors and Makinson (AGM), whose account of belief revision has been widely accepted. The AGM theory models how an ideal agent or program corrects its beliefs about the world when it finds them to be mistaken. It results suitable to model hypothetical changes in belief. We will adopt it as the underlying theory of revision in our account of explanation.

Other approaches to revision have been presented in the AI literature. For example, contrasting with the AGM belief revision, which can be described as modeling changes in belief in a static world, Katsuno and Mendelzon's (1991a,b) notion of belief update captures changes in belief about a changing world. In Chapter 6, we will briefly discuss a counterpart notion of explanation based on update semantics as an avenue of investigation. Among others, Doyle's Truth Maintenance System (TMS) (1979) was pioneer in AI embedding the foundations theory of revision which keeps track of all the reasons for belief. This approach contrasts to the coherence tradition with which the AGM theory identifies.

2.3.1 The AGM theory of Revision

The AGM theory (AGM 1985, Gärdenfors 1988) models the dynamics of belief as governed by the logic of theory change. We will follow the AGM formalization of belief revision in terms of belief sets, which assumes an ideal reasoner, whose beliefs are totally consistent and closed under logical consequence. Belief sets represent a particular
Chapter 2. Prelude on Explanations, Belief Revision and Conditional Logic

belief state by a deductively closed set of sentences (we will assume the underlying logic of beliefs is CPL, but nothing crucial depends on this). An arbitrary belief set will be denoted with $K$. If $K$ is a consistent belief set, then for any sentence $A$ only three different epistemic attitudes concerning $A$ can be expressed. If $A \in K$, $A$ is accepted or believed; if $\neg A \in K$, $A$ is rejected, or disbelieved; and, if $A \not\in K$ and $\neg A \not\in K$, $A$ is indeterminate. Belief sets admit three kinds of epistemic changes: expansions, revisions and contractions. Expansions consist in adding a new belief $A$ (and its consequences) into $K$. Expansions are defined as $K_A^+ = Cn(K \cup \{A\})$; then, to obtain a new consistent belief set $K_1$ it is required that $A$ be consistent with $K$.

Revisions model the process of introducing in $K$ a new belief $A$, that may contradict beliefs already in $K$. Certain beliefs from $K$ should be given up in order to make place for $A$, and maintain consistency. Following the principle of informational economy, the idea is that “as few” beliefs as possible and the “least entrenched” beliefs should be given up in order to preserve consistency. By “few” it is meant that the change in informational content in theory $K$ be minimal. The AGM theory proposes the following eight postulates that constrain the revised belief set.

(K*1) For any sentence $A$ and any belief set $K$, $K_A^*$ is a belief set

(K*2) $A \in K_A^*$

(K*3) $K_A^* \subseteq K_A^+$

(K*4) If $\neg A \not\in K$, then $K_A^+ = K_A^*$

(K*5) $K_A^* = \bot$ if and only if $\vdash \neg A$

(K*6) If $\vdash A \equiv B$, then $K_A^* = K_B^*$

(K*7) $K_{A \& B}^* \subseteq (K_A^*)_B^+$

(K*8) If $\neg B \not\in K_A^*$, then $(K_A^*)_B^+ \subseteq K_{A \& B}^*$

Contractions are changes in a belief state that involve giving up some beliefs without incorporating new ones. Contractions model the resulting belief state after the ideal agent “forgets” some information. Given a belief set $K$, the contraction of $K$ by $A$
is notated as $K^-_A$. In this contracted belief state the agent should not hold belief in $A$. When retracting a belief $A$ from $K$, there may be other beliefs in $K$ that entail $A$ (or other beliefs that jointly entail $A$ without separately doing so). In order to keep $K^-_A$ closed under logical consequences, it is necessary to give up $A$ and other beliefs as well. The problem is to determine which beliefs should be given up and which should be retained. The AGM theory provides eight postulates ($K^-1$) to ($K^-8$) that characterize contractions.

($K^-1$) For any sentence $A$ and any belief set $K$, $K^-_A$ is a belief set

($K^-2$) $K^-_A \subseteq K$

($K^-3$) If $A \notin K$, then $K^-_A = K$

($K^-4$) If not $\vdash A$, then $A \notin K^-_A$

($K^-5$) If $A \in K$, then $K \subseteq (K^-_A)^+_A$

($K^-6$) If $\vdash A \equiv B$, then $K^-_A = K^-_B$

($K^-7$) $K^-_A \cap K^-_B \subseteq K^-_{A \& B}$

($K^-8$) If $A \notin K^-_{A \& B}$, then $K^-_{A \& B} \subseteq K^-_A$

By the Levi identity revisions can be defined in terms of contractions and expansions:

$$K^*_A = (K^-_A)^+_A$$

The Levi identity defines revisions as first pruning away all potential inconsistencies, and then adding the new belief. The Harper identity provides a definition of contractions in terms of revisions:

$$K^-_A = (K \cap K^*_A)$$

The state where we become ignorant about $A$ is captured as what $K$ and $K^*_A$ have in common. To abandon belief in $A$ is just to make both $A$ and $\neg A$ epistemically possible. Gärdenfors et al. propose two different ways to construct a contraction function satisfying ($K^-1$) to ($K^-8$). By means of the Levi identity, a revision function that satisfies ($K^*1$) to
(K*8), is automatically determined. The first approach to define a contraction function to get $K_A^-$ form $K$, and a sentence $A$, is based on maximal subsets of $K$ that do not entail $A$. $K \perp A$ denotes the set of all these maximal subsets. It is assumed that there is some ordering of maximal subsets in $K \perp A$. By means of a selection function $S$ the most relevant subsets from $K \perp A$ are picked. They define a Partial Meet contraction function as follows:

**Definition 2.13 (Gärdenfors)** $K_A^- = \cap S(K \perp A)$

The second approach to construct a contraction function to obtain $K_A^-$ is based on some ordering of the sentences in $K$. The ordering is associated with the epistemic entrenchment of sentences. For sentences $A$, $B$, $A \leq_{EE} B$ means that “$B$ is at least as epistemologically entrenched as $A$ relative to epistemic state $K$”. Gärdenfors indicates that the epistemic entrenchment of a sentence is tied to its overall informational value within the belief set. For example, lawlike sentences generally have greater epistemic entrenchment than accidental generalizations. When forming contractions, the sentences that are retracted are those with the lowest epistemic entrenchment. Tautologies are the most entrenched beliefs; hence they are never given up. He proposes the following postulates that constrain an ordering of entrenchment.

**(EE1)** If $A \leq_{EE} B$ and $B \leq_{EE} C$, then $A \leq_{EE} C$ (Transitivity)

**(EE2)** If $A \models B$ then $A \leq_{EE} B$ (Dominance)

**(EE3)** For any $A$ and $B$, $A \leq_{EE} A \land B$ or $B \leq_{EE} A \land B$ (Conjunctiveness)

**(EE4)** When $K \neq K_L$, $A \not\in K$ iff $A \leq_{EE} B$ for all $B$ (Minimality)

**(EE5)** If $B \leq_{EE} A$ for all $B$, then $\models A$

**(EE1)-(EE3)** imply connectivity, namely, either $A \leq_{EE} B$ or $B \leq_{EE} A$ (the epistemic entrenchment ordering will cover all the sentences).

Adam Grove (1988) has given a concrete modeling for theory change in terms of his “system of spheres”, furnishing the AGM theory with a possible worlds semantics. Grove’s system of spheres (similar to Lewis’) consists in sets of possible worlds ordered
concentrically. The system of spheres is ordered under inclusion. The set of worlds in the center of the system forms the inner sphere that represents a theory of interest $K$. Distance to the centered sphere means departure from theory $K$.

2.3.2 Belief Revision and Default Reasoning

Default reasoning is the process of jumping to conclusions in the presence of incomplete knowledge. It has been commonly agreed and re-agreed that the process is non-monotonic since in light of new information, old conclusions may be no longer valid. Makinson and Gärdenfors (1990, 1991), revealed the relation between skeptical default reasoning and belief revision. The idea is that the underlying theory being object of revision be interpreted as a theory of expectations (or defaults) about the world. Then, conclusions in default reasoning correspond to the revision of a theory of expectations by the facts. Gärdenfors and Makinson developed a non-monotonic inference relation, proving a unified treatment of nonmonotonic logics and belief revision. Their nonmonotonic inference ($\vdash$) is based on a notion of “expectations”. Now $K$ represents a theory of expectations about the world. Expectations include not only our firm beliefs, but also the propositions that are regarded as plausible enough to be used as a basis for inference as long as they don’t give rise to inconsistency. Gärdenfors and Makinson’s thesis is:

$$\beta \in K^*_\alpha$$ is translated into $\alpha \vdash_K \beta$, and conversely.

In terms of this inference relation they establish the connection with Poole’s Theorist framework with respect to the process of default prediction or intersection of all extensions of a default theory.

Boutilier (1992a)(1992c) pushed the connection even further, showing that the logics governing belief revision and conditional default reasoning are indeed identical. He developed a family of modal conditional logics (see below) suitable for both types of reasoning, providing a uniform semantic framework based on standard (Kripkean) models. He drew the connection between default reasoning and belief revision by defining normative and subjunctive conditionals (see below) in his bimodal conditional logic $CO^*$. The
A subjunctive conditional, via the Ramsey test, characterizes AGM revisions, while default rules of the form “A normally implies B” are modeled as normative conditionals $A \Rightarrow B$. Boutilier showed the two conditionals to be the same, and how default reasoning can be viewed as a special case of belief revision.

2.4 Modal and Conditional Logics

Conditionals have the form “If $A$ were the case, then $B$ would be the case” or “If $A$ is the case, then $B$ is (will be) the case”, where $A$ may or may not contradict our knowledge or beliefs. Conditional logics were initially developed for modeling “if ... then” statements in natural language. Robert Stalnaker (1968) gives a possible worlds semantics for his conditional logic. The subjunctive conditional $A > B$, is read as “if $A$ were true $B$ would be true”. Stalnaker’s conditional is non-transitive (from $A > B$ and $B > C$ one can not infer $A > C$), hence, it does not satisfy the strengthening rule (from $A > B$ one can not infer $A \land C > B$). And it does not satisfy the contraposition rule (from $A > B$ one can not infer $\neg B > \neg A$). These qualities make the conditional connective suitable for subjunctive reasoning. For instance, we accept the conditional “If this match were struck, it would light”, while we deny that “If this match were wet and struck, it would light”. Stalnaker gives the following “recipe” based on the Ramsey test to evaluate a conditional in a given state of belief:

“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 then true.” (Stalnaker 1968, page 44)

The connection between belief changes and conditional sentences is given by the Ramsey test. A conditional $A > B$ is accepted in an epistemic state $K$ iff $B$ is accepted in the revision of $K$ by $A$. This connection is foundational in our conditional theory of explanation.

David Lewis (1973) proposes a conditional theory of counterfactuals (differing in some aspects to Stalnaker’s), giving a possible worlds semantics based on a “system of spheres”.
Any particular sphere around a world \( w \) contains the worlds that resemble to \( w \) at least to a certain degree, and, of course, ties are permitted. The system is nested, in that for any two spheres around \( w \), one is included in the other. Lewis provides the following truth condition for his "would" counterfactual conditional relative to the system of spheres around a world \( w \). \( A > B \) is true at \( w \) if and only if either (a) no \( A \)-world belongs to any sphere, or (b) some sphere \( S \) contains at least one \( A \)-world, and \( A \supset B \) holds at every world in \( S \). Then, \( A > B \) is non-vacuously true at \( w \) if some world accessible to \( w \) in which \( A \) and \( B \) hold resembles to \( w \) more than any other world satisfying \( A \) and \( \neg B \).

Lewis also defines a "might" counterfactual in terms of his "would" counterfactual, as \( A \sim C \equiv \neg(A > \neg C) \). The truth condition for the might-conditional, relative to the system of spheres around a world \( w \), is as follows. \( A \sim C \) is true at \( w \) if and only if some \( A \)-world belong to some sphere, and every sphere \( S \) that contains at least one \( A \)-world, contains at least one world where \( A \wedge B \) holds. If the "would" counterfactual \( A > B \) is non-vacuously true, then the "might" counterfactual \( A \sim B \) is also true. If \( A > B \) and \( A > \neg B \) are both false, then \( A \sim B \) and \( A \sim \neg B \) are both true. However, when \( A > B \) is false, but \( A > \neg B \) is true, then \( B \) holds at none of the closest \( A \)-worlds, hence \( A \sim B \) is false. Finally, if there are no \( A \)-worlds, the conditional \( A \sim B \) is also false.

Matthew Ginsberg (1986) discussed the role of counterfactuals in AI, describing how applications like planning and automated diagnosis can be modeled via counterfactual reasoning.

2.4.1 Boutilier's Logics as Calculus for Default Reasoning and Belief Revision

We will focus on two of Boutilier's logics, CT4O and CO. These logics, suitable for revision and default reasoning, possess a possible worlds semantics based on standard models (Kripke models). CT4O and CO are based on a bimodal language \( L_B \), with the modal operators \( \square \) for truth at accessible worlds, and \( \Box \) for truth at inaccessible worlds. The following connectives are defined in terms of \( \square \) and \( \Box \): \( \diamond \alpha \equiv \neg \square \neg \alpha \); \( \Diamond \equiv \square \square \).
Chapter 2. Prelude on Explanations, Belief Revision and Conditional Logic

\[ -\square \neg \alpha; \square \equiv df \quad \square \alpha \land \square \alpha; \Diamond \equiv df \quad \Diamond \alpha \lor \Diamond \alpha. \]  
\[ \Diamond \alpha \text{ is true at a world iff } \alpha \text{ holds at some accessible world; } \Diamond \alpha \text{ is true at a world iff } \alpha \text{ holds at some inaccessible world; } \square \alpha \text{ holds iff } \alpha \text{ holds at all worlds. } \Diamond \alpha \text{ holds iff } \alpha \text{ holds at some worlds. Notice that the connectives } \Diamond \text{ and } \square \text{ express global statements in a model.} \]

We will use extensively the term cluster to indicate a maximal set of mutually accessible worlds. The following is a general definition of a cluster that applies to any reflexive and transitive model.

**Definition 2.14** Given any model \( M = (W, R, \varphi) \), for which \( R \) is reflexive and transitive a cluster is defined to be a 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 has this property.

The logic CT4O characterizes the class of models where \( R \) induces a partially ordered set of clusters of mutually accessible worlds (S4 structures).

**Definition 2.15 (Boutilier)** A CT4O-model is a triple \( M = (W, R, \varphi) \) where \( W \) is a set of worlds with valuation function \( \varphi \) and \( R \) is a reflexive and transitive binary accessibility relation over \( W \). Satisfaction is defined in the usual way, with the truth of a modal formula at a world defined as: \( \models_w \square \alpha \text{ iff for each } v \text{ such that } wRv, \models_v \alpha \) and \( M \models_w \square \alpha \text{ iff for each } v \text{ such that not } wRv, \models_v \alpha. \)

**Definition 2.16 (Boutilier)** The conditional logic CT4O is the smallest \( S \subset L_B \) such that \( S \) contains CPL (and its substitution instances) and the following axiom schemata, and is closed under the following rules of inference:

\[\begin{align*}
K & \quad \square (A \supset B) \supset (\square A \supset \square B) \\
K' & \quad \square (A \supset B) \supset (\square A \supset \square B) \\
T & \quad \square A \supset A \\
4 & \quad \square A \supset \square \square A \\
H & \quad \Diamond (\square A \land \square B) \supset \square (A \lor B) \\
Nes & \quad \text{From } A \text{ infer } \square A. \\
MP & \quad \text{From } A \supset B \text{ and } A \text{ infer } B. 
\end{align*}\]
Provability and derivability are defined in the standard fashion, in terms of theoremhood.

**Theorem 2.6 (Boutilier)** \( \vdash_{CT40} \alpha \iff \models_{CT40} \alpha. \)

In many circumstances we want to ensure that all logically possible worlds are taken into consideration. The logic CT40\(^*\) characterizes such full models.

**Definition 2.17 (Boutilier)** CT40\(^*\) is the smallest extension of CT40 containing instances of the following axiom: LP \( \lozenge \alpha \) for all satisfiable propositional \( \alpha \).

**Theorem 2.7 (Boutilier)** \( \vdash_{CT40^*} \alpha \iff \models_{CT40^*} \alpha. \)

The other logic that we are interested in is CO which characterizes structures consisting of a totally ordered set of clusters of mutually accessible worlds (S4.3 structures).

**Definition 2.18 (Boutilier)** A CO-model is a triple \( M = (W, R, \varphi) \) where \( W \) is a set of worlds with valuation function \( \varphi \) and \( R \) is a transitive and connected (transitivity and connectivity imply reflexivity) accessibility relation over \( W \).

The conditional logic CO is the smallest \( S \subseteq L_B \) such that \( S \) contains CPL (and its substitution instances) and the following axiom schemata K, K', T, 4, H and

\[
S \quad A \supset \lozenge \lozenge A
\]

and is closed under the rules of inference Nes and MP. Provability and derivability are defined in the standard fashion, in terms of theoremhood.

**Theorem 2.8 (Boutilier)** \( \vdash_{CO} \alpha \iff \models_{CO} \alpha. \)

The logic CO\(^*\) is based on full CO-models. CO\(^*\) is the smallest extension of CO containing instances of the axiom LP. CO\(^*\) is also sound and complete.

We should notice that CO models are a special case of CT40 models. The difference between the two lies in that the accessibility relation in CT40 models is transitive and reflexive, while in CO it also requires to be connected. A CT40-model is formed by a
Chapter 2. Prelude on Explanations, Belief Revision and Conditional Logic

Figure 2.1: A CO-model (left) and a CT4O-model (right).

lattice of clusters, a CO-model by a unique “chain” of clusters. Figure 2.1 shows a CO and a CT4O model.

Revision Models and Models of Normality

Boutilier (1992a)(1992c) demonstrated that the formal processes of belief revision and default reasoning are governed by the same logic. But he showed that practical considerations distinguish the two types of reasoning. He captured the difference between the two with a different interpretation on the accessibility relation \( R \) among possible worlds. In models for default reasoning, worlds or states of affairs have a fixed ordering according to a measure of normality, or typicality. The accessibility relation \( R \) is interpreted as follows: world \( v \) is accessible to world \( w \) if and only if \( v \) is at least as normal a \( w \).

In models for revision, worlds are ordered respecting a plausibility measure relative to a belief set \( K \) that will be the object of revision. Boutilier defines models for revision by imposing an interpretation on accessibility relation \( R \) over possible worlds as a measure of closeness to theory \( K \), or plausibility with respect to theory \( K \). Plausibility reflects the degree to which one would accept \( w \) as a possible state of affairs given that belief in
Chapter 2. Prelude on Explanations, Belief Revision and Conditional Logic

$K$ may have to be given up. $R$ is defined as follows: $wRv$ iff $v$ is as plausible as $w$ given theory $K$. $v$ is more plausible than $w$ iff $wRv$ but not $vRw$. $R$ is required to be reflexive, since every world is as plausible as itself, and transitive, since if $w$ is as plausible as $v$, and $v$ is as plausible as $u$, then $w$ is as plausible as $u$. In some cases $R$ is required to be connected, namely, all worlds be comparable in terms of plausibility (either $wRv$ or $wRv$ for all $w, v$). CT4O-models seem appropriate for revision; however, in order to use them, the theory $K$ should be adequately represented. Boutilier achieves this by insisting that those worlds consistent with $K$ should be exactly those minimal in $R$. That is,

$$vRw \text{ for all } v \in W \text{ if and only if } M \models_w K$$

This condition ensures that no world is more plausible than any other world consistent with $K$, and that all $K$-worlds are equally plausible. Models that satisfy this constraint are called $K$-revision models. If we let $\|K\|$ to denote the set of worlds satisfying each formula in $K$, a model $M$ is a $K$-revision model if and only if $\|K\|$ forms a minimal cluster at the bottom of the model. Boutilier states this constraint in the bimodal language $L_B$, in the case of CO-models, for any $K$ that is finitely expressible as $KB$ as:

$$O(KB) \equiv (KB \supset (\Box KB \land \Box \neg KB))$$

This ensures that any $KB$-world sees every other $KB$-world ($\Box \neg KB$), and that it sees only $KB$-worlds ($\Box KB$). The sentence $O(KB)$ is intended to mean we “only know” $KB$ (Levesque).

Let $K = Cn(KB)$, such that $KB = \{ \neg A, B \}$. Figure 2.2 shows a $K$-revision model. The minimal cluster satisfies $\neg A, B$; hence, $A$ is rejected in $K$, $B$ is accepted in $K$, while $C$ is just considered epistemically possible.

We will work with two versions of $K$-revision models, differing in the properties of the accessibility relation. If the accessibility relation is just reflexive and transitive, Boutilier denotes them preorder revision-models. In contrast, if in addition the accessibility relation is totally connected, the models are called total order revision-models. The difference between the two is simply that in total-order revision models every pair
of worlds are comparable with respect to plausibility or closeness to theory $K$, while in preorder revision-models, some worlds may be incomparable. Full models (all propositional valuations are represented) should be considered in order to allow every consistent sentence be capable of generating a consistent revision (this relates to the AGM postulate $K^{-5}$, see below). Full total-order revision models are CO* models, and full preorder revision-models are CT4O* models. Boutilier defines a modality for belief. The modality $B$ is defined equivalently in CT4O and CO (although in CO it can have a simpler definition):

$$ B\alpha \equiv_{df} \Box \Diamond \Box \alpha $$

$B\alpha$ is read as $\alpha$ is believed (or accepted) in $K$. The definition says that every world in the model should regard possible that $\alpha$ is true at all accessible worlds. This is the case when $\alpha$ holds at every world in the minimal clusters.

Boutilier defines the *plausibility* of a proposition relative to a revision-model $M$. A proposition $A$ is at least as plausible as $B$ when for every $B$-world $w$ there is some $A$ world that is at least as plausible as $w$. 
Definition 2.19 (Boutilier) Let $M$ a CT4O-model. $A$ is at least as plausible as $B$ in context of $M$ (written $A \leq B$) if and only if $M \models \Box (B \supset \Diamond A)$.

Proposition $A$ is more plausible than $B$ just when $A \leq B$ and not $B \leq A$. When interpreting the accessibility relation $R$ as an ordering of normality among possible worlds, $A \leq B$ may reflect a comparative measure of normality. The notion of plausibility normality will be used extensively in our account of explanations.

We are ready to introduce the conditional connective $\Rightarrow$. In models for revision this connective will be used to represent and reason about subjunctives and revision policies. The connective $\Rightarrow$ is defined equivalently in the two logics, CT4O and CO.

Definition 2.20 (Boutilier) $A \Rightarrow B \equiv_d (\Box \neg A \lor \Diamond (A \land \Box (A \supset B)))$

The conditional $A \Rightarrow B$ (in a model $M$) says that for every world $w$, either $w$ has no access to any $A$-world, or there is some $A$-world $v$ accessible to $w$ such that for every world $u$ accessible to $v$, $\models_u A \supset B$ holds. Notice that $\Rightarrow$ has a "global" nature, or in other words, the truth of the conditional is not defined relative to a particular world, but relative to the whole model.

Referring to figure 2.1, let's call $M_1$ to the CO-model (on the left) and $M_2$ to the CT4O-model (on the right). In model $M_1$ there is a unique $R$-minimal cluster of $A$-worlds. In contrast, in model $M_2$ there are two $R$-minimal clusters of $A$-worlds. In both models satisfaction of the conditional $A \Rightarrow B$ occurs since $B$ holds at each of the most plausible ($R$-minimal) $A$-worlds.

Now we can see how Boutilier captures AGM revision. Let a $K$ be a $K$-revision model. When revising $K$ by $A$, we adopt as representative of the revised belief set $K_A^*$ the set $Pt(A)$, of most plausible $A$-worlds in $M$. Using the Ramsey test for acceptance of conditionals $B \in K_A^*$ is equated with $M \models A \Rightarrow B$, both capturing that $B$ is true at each of the most plausible $A$-worlds. Given the Ramsey test, $\Rightarrow$ is nothing more than a subjunctive conditional, interpreted as "If $K$ were revised by $A$, then $B$ would be believed". For any propositional $A \in L_{CPL}$, the belief set resulting from revision of $K$ by $A$ is:
In the model of figure 2.2 the conditional $M \models A \Rightarrow \neg B$ holds, meaning that $\neg B$ would be accepted as a result of revising $K$ by $A$. The belief set $K_A^*$ is characterized by the two clusters in the top of the model. Notice that $M \models A \not\iff C$, which says that if we revise $K$ by $A$, $C$ will not become accepted.

Boutilier demonstrates that the revision function determined by a full total-order revision-model satisfies the eight AGM postulates for revision ($K^*1 - K^*8$), while the revision function determined by a preorder revision-model satisfies $K^*1 - K^*7$.

Let's see the semantic correspondence of AGM contraction in $K$-revision models. Let a belief set $K$ such that $A$ is rejected (namely, $\neg A$ is accepted). To contract belief in $\neg A$ implies to give $A$ a possibility; that is, to move into an epistemic state in which $A$ is epistemically possible. The belief set resulting for contracting $K$ by $\neg A$ is determined by the set of worlds $\|K\| \cup Pl(A)$. In the model of figure 2.2 the belief set $K_{\neg A}$ is characterized by three clusters, the minimal cluster together with the two clusters in the top of the model. Notice that the proposition $(B \supset C)$ is accepted in $K_{\neg A}$.

We will present now another of Boutilier’s conditionals, that we will notate as $\rightarrow$. This conditional has also a global nature but expresses a weaker notion than $\Rightarrow$. $A \rightarrow C$ captures the idea that “at some set of equally most plausible states of affairs where $A$ holds, $C$ holds as well”. The conditional $\rightarrow$ is defined as follows, in both logics, CO and CT4O.

**Definition 2.21 (Boutilier)** $A \rightarrow C \equiv_{df} \Box \neg A \lor \lozenge (A \land \Box (A \supset C))$

The sentence $A \rightarrow C$ says that either $\neg A$ holds universally, or there is some $A$-world $w$ such that for every world $v$ accessible to $w$, $v, \models A \supset C$ holds. Then, the conditional $A \rightarrow C$ holds in a CT4O-model whenever $C$ holds at least at one of the $R$-minimal clusters of $A$-worlds. However, in a CO-model, since clusters of worlds are totally ordered, there can only be a unique $R$-minimal cluster of $A$-worlds. Then $A \rightarrow C$ will be satisfied in a CO-model, whenever $C$ holds at the $R$-minimal cluster of $A$-worlds. But
this is precisely the same requirement for satisfaction of the stronger conditional \( A \Rightarrow C \). Consequently the two connectives, \( \Rightarrow \) and \( \rightarrow \), become equivalent in CO-models, as the following proposition shows.

**Proposition 2.9 (Boutilier)** \( \vdash_{CO} A \rightarrow C \equiv A \Rightarrow C \)

Boutilier (1990)(1992a) observed that this weak conditional connective is *paraconsistent* in CT4O-models. Since it is possible to assert \( A \rightarrow C \) and \( A \rightarrow \neg C \) while \( A \not\rightarrow (C \land \neg C) \); that is, without \( A \) entailing an inconsistency. The CT4O-model in figure 2.1 satisfies both \( A \rightarrow C \) and \( A \rightarrow \neg C \), while the CO-model satisfies neither. In a CO-models the conditional \( \rightarrow \) does not behave paraconsistently since it just becomes equivalent to the stronger connective \( \Rightarrow \), hence loosing its intended meaning.

According to the interpretation of the accessibility relation in terms of normality, the reading of the conditional \( A \Rightarrow B \) is stated as “in the most normal situations in which \( A \) holds, \( B \) holds as well”, or “\( A \) normally implies \( B \)”. Default rules like “birds fly” and “penguins don’t fly” are naturally expressed as \( \text{bird} \Rightarrow \text{fly} \) and \( \text{penguin} \Rightarrow \neg \text{fly} \). The nonmonotonicity of default reasoning is nicely captured, since these conditional default rules are exception allowing. Boutilier shows how skeptical default reasoning can be viewed as the revision of a theory of expectations (defaults) by the facts.

In the next chapters we will only be concerned with CO and CT4O models that satisfy the so-called Limit Assumption. This assumption states that in each model there exists some minimal \( A \)-world for any proposition \( A \). Pictorially, we can say that models that satisfy the Limit Assumption have are “bottomed”; that is, they have a set of *minimal* clusters.\(^2\) The following propositions will be of use in many of the proofs of the next chapters. The propositions describe truth conditions for conditionals \( A \Rightarrow B \) and \( A \rightarrow B \) relative to CT4O-models that satisfy the Limit Assumption.

**Proposition 2.10** Let \( M = \langle W, R, \varphi \rangle \) a CT4O well founded model (\( R \) is transitive and reflexive, and satisfies the Limit Assumption). \( M \models A \Rightarrow B \) iff every minimal \( A \)-cluster in \( M \) satisfies \( B \).

\(^2\)The Limit Assumption is discussed in detail in Lewis (1973) and Boutilier (1992a).
Proposition 2.11 Let $M = \langle W, R, \varphi \rangle$ a CT4O well founded model ($R$ is transitive and reflexive, and satisfies the Limit Assumption). $M \models A \rightarrow B$ iff either there are no minimal $A$-clusters, or there is a minimal $A$-cluster in $M$ that satisfies $B$.

The two conditional connectives $\Rightarrow$ and $\rightarrow$ will be used as connectives for explanation in our model of abduction.
Chapter 3

Epistemic Explanations

Why would Fred contract AIDS? Compare the following two explanatory answers.\(^1\) By Fred's practicing unsafe sex. By Fred's being exposed to the HIV virus. The second answer has a predictive nature. If we believed that Fred was exposed to the virus, we would predict that Fred contracted AIDS. The first explanation is not predictive, since even though it may be somewhat probable that Fred contracted AIDS, practicing unsafe sex is not enough of a reason for us to believe that Fred actually contracted AIDS. It could have perfectly been the case that the persons that Fred had met were not infected with the virus. We will just say that Fred might have contracted AIDS. Based on this distinction we will differentiate predictive explanations from non-predictive ones.

Our modeling of explanation assumes an ideal agent or program, which is logically omniscient and perfectly consistent, inconsistency being an a intolerable state for it. Let \( K \) be a deductively closed set of objective beliefs representing knowledge or beliefs\(^2\) about the world held by the agent. We also expect some conditional beliefs (and maybe some other preferences too) that constrain the manner in which the agent is willing to revise its beliefs. These are revision policies, subjunctive conditionals, of the form “If the agent were to believe in \( A \), then it would believe in \( B \).”

We assume the agent will be confronted with explanation-seeking—“why” questions, for which it would give explanations relative to its epistemic state \( K \). Namely, we are modeling external inquiries that “consult” the agent's credences. We will attempt to explain two kinds of objective sentences, beliefs and non-beliefs.\(^3\) We will generically use

\(^1\)Thanks to David Poole and Craig Boutilier for proposing and analyzing this example.
\(^2\)For the purpose of this thesis we will consider belief and knowledge as interchangeable concepts.
\(^3\)As future work we contemplate explaining conditional statements. For example ("The Little Prince" page 53), “Why would an explorer who told lies bring disaster on the books of the geographer?” We
the word *explanandum* and sometimes *observation* to refer to the object of inquiry. We will use the greek letter $\beta$ to denote a generic explanandum or observation, and $\alpha$ will refer to an explanation for $\beta$. Through this chapter will assume $\alpha, \beta \in \mathbb{L}_{CPL}$.

Along with the idealizations of our modeling, we will assume that did the agent learn a new piece of information, it would automatically incorporate some explanation for it (at worst, a disjunction of reasons, or the trivial explanation). Such an explanation would arise from the closure of its beliefs together with the newly incorporated knowledge. It may not be a "single" reason, but a disjunction of possible reasons. The request to explain $\beta$ is interpreted by the agent as "Why do/would $\beta$ be true, based on what I know?" Then, if $\beta$ was already accepted in the epistemic state, whatever reason for $\beta$ already incorporated in the epistemic state should be an explanation for $\beta$. Then, we argue that when the agent receives an inquiry about an explanandum that it believes true, an explanation should also be believed true. Namely,

"Were the object of inquiry already believed, an explanation for it would be believed too."

The following condition reflects our commitment.

\[(\text{Commit}) \quad \text{if } \beta \in K \text{ then } \alpha \in K\]

As Boutilier has pointed out\(^4\), this view is not tenable for scientific inquiry, where observations are believed, but explanations are yet unknown. The explanations to be studied in this chapter are based on this commitment. A notion of scientific explanation may be modeled by denying it.

We will study with two notions of explanation: predictive and non-predictive.

### 3.1 Predictive Explanations

Predictive explanations render necessary the explanandum. This is a strong notion of an explanation: belief in the explanation will induce belief in the explanandum. The

---

\(^4\)Personal communication.
slogan will be *Were the explanation believed, so too would the explanandum.* A number of conditions follow, defining when \( \alpha \) explains \( \beta \), relative to a background epistemic state \( K \), where the inquiry arises. The explanandum \( \beta \) may either be rejected, indeterminate or accepted in \( K \). According to *Commit*, if \( \beta \) is an observation that is already accepted in \( K \), an explanation \( \alpha \) should also be accepted in \( K \). The agent must have incorporated some reason along with the observation; if modeling the physical world, an actual cause may account for the observation.

Imagine now \( \beta \) rejected in \( K \), that is \( \neg \beta \in K \). We seek an explanation \( \alpha \), such that \( \alpha \) predictively accounts for \( \beta \). We argue that \( \alpha \) should necessarily be rejected in \( K \) as well. Otherwise, were \( \alpha \) accepted or indeterminate, any \( \alpha \)-world in \( K \) would still satisfy \( \neg \beta \) (since \( \beta \) is rejected). Then, \( \alpha \) would hardly make sense as a predictive explanation for \( \beta \).

Finally, suppose \( \beta \) indeterminate in \( K \). Were \( \alpha \) rejected in \( K \), then in every epistemically possible world where \( \beta \) holds, \( \alpha \) would be absent, hence \( \alpha \) would not be accounting the possibility of \( \beta \). If \( \alpha \) were accepted while \( \beta \) is not (\( \beta \) is actually indeterminate), how can \( \alpha \) “predict” \( \beta \)? We conclude that \( \alpha \) should be indeterminate in \( K \). Hence, we propose:

> “The object of inquiry and the explanation should have the same epistemic status in \( K \”).

We translate the above statement into the following condition:

\[
(\text{EpStatus}) \quad \alpha, \beta \in K; \text{ or } \neg \alpha, \neg \beta \in K; \text{ or } \alpha, \neg \alpha, \beta, \neg \beta \notin K
\]

But obviously \( \text{EpStatus} \) is not enough. Even though we may disbelieve both, that the grass is wet and that Fred is a drug-addict, one does not explain the other. We need to refer to some “connection” between \( \alpha \) and \( \beta \). Reflecting our predictive desideratum, we propose:

> “Were the explanation believed, the explanandum would be believed too”.

Chapter 3. *Epistemic Explanations* 43
The idea is that at all most plausible $\alpha$-worlds, $\beta$ should also hold; for otherwise, $\alpha$ would not predictively account for $\beta$.

\[(\text{Predict})\quad \beta \in K^*_\alpha\]

However, once we have adopted EpStatus, if $\beta$ is already accepted in $K$, Predict dilutes or becomes automatically satisfied by considering any arbitrary belief. To evaluate Predict non-trivially, we should hypothetically suspend belief in the explanandum, moving into an epistemic state where $\beta$ is not accepted. This hypothetical state is $K^-\beta$. From the AGM contraction postulates, we know that if $\beta \notin K$, then $K^-\beta = K$, then we can re-express Predict as follows,

\[(\text{Predict}')\quad \beta \in (K^-\beta)^*_\alpha\]

i.e., we hypothetically suspend belief in $\beta$ and ask “What would “cause” its belief?” If $\alpha$ fails to satisfy Predict', then clearly, $\alpha$ can not be a predictive explanation for $\beta$.

Consider now the following situation. Suppose we believed Fred was infected with the HIV virus. This would predictively explain Fred having AIDS. Had we believed that Fred did \textit{not} have AIDS, we would have positively believed that Fred was \textit{not} exposed to the HIV virus. We propose a third condition on predictive explanations, that we phrase as follows:

“Had the observation been absent, the explanation would be absent too”.

Candidate hypotheses violating this condition clearly fail to be explanations. For instance, the fact the the grass is wet does not explain that Fred is in love. Since even if the grass were not wet, our belief about Fred being enamored would remain. We argue, then, that the wet grass does not explain Fred’s being in love. We can express our proposed condition as:

\[(\text{Absent})\quad -\alpha \in K^-\beta\]
Absent requires an explanation to be rejected in those “closest” states of affairs where the observation is rejected. Otherwise, if \( \alpha \) were accepted in some states of affairs where \( \neg \beta \) holds, such an explanation would also be accounting for the negation of the observation, so it would be deprived of its predictive explanatory power. Unfortunately, if the object inquiry is disbelieved (i.e. \( \neg \beta \in K \)), Condition Absent becomes automatically satisfied, once we accept EpStatus. The truth of Absent can be non-trivially tested in an epistemic state where we suspend belief in the explanation. That is, by hypothetically moving to the “closest” states of affairs where the potential explanation is indeterminate (epistemically possible). This state is \( K^-_\alpha \). From \( K^-_\alpha \), we should evaluate whether disbelief in the observation induces disbelief in the explanation. Again, from the AGM postulates for contractions, we know that if \( \neg \alpha \notin K \) then \( K^-_\alpha = K \). So we can re-express Absent as:

\[(\text{Absent'}) \quad \neg \alpha \in (K^-_\alpha)^*-\beta\]

So far we have proposed EpStatus, Predict', and Absent' as three necessary conditions to determine when a sentence \( \alpha \) explains an explanandum \( \beta \).

We will now present logical counterparts of the conditions described above using Boutilier’s logic CT4O. We will assume a plausibility ordering over possible worlds, induced by the agent’s objective belief set \( K \), together with its revision policies and possibly other preferences too. We will assume the existence of a \( K \)-revision model capturing the ordering. As we discussed in Section 2.4.1, via the Ramsey test, revision policies are just subjunctive conditionals of the form \( A \Rightarrow B \) and plausibility comparisons can have the form \( C \leq D \). We will work with two versions of the plausibility ordering, one determining a preorder revision model, and the other a total order revision model (see Section 2.4.1). Unless specified, we will consider indifferent whether the underlying model is a preorder or total order revision-model.

We should notice that, in general, there are multiple admissible orderings among possible worlds. As shown by Boutilier, a fixed ordering worlds reflecting an ordering of normality identifies the epistemic state \( K \) with a theory of expectations. Models based on this fixed ordering of normality are suitable for default reasoning. The counterpart
explanations arising from these models will be the object of study in the next two chapters.

Let's assume a K-revision model \( M \). Condition Commit is easily formulated using Boutilier's modality for belief \( B \),

\[
\text{(3.1)} \quad M \models B\beta \supset B\alpha
\]

and condition EpStatus as,

\[
\text{(3.2)} \quad M \models (B\alpha \equiv B\beta) \land (B\neg\alpha \equiv B\neg\beta)
\]

Predict' sanctions that \( \beta \) should be accepted in the belief state resulting from the revision of \( (K\beta) \) by \( \alpha \). This complicates matters since in order to capture revisions over this epistemic state, the corresponding model \( M_\beta \) suitable for revision of \( K_\beta \) has to be defined. While the semantic model of AGM contraction constrains the ordering of possible worlds in the contracted state, it fails to provide a new ordering. Let's leave \( M_\beta \) underspecified, taking it as any model such that the worlds in \( K_\beta \) form the minimal cluster.\(^5\) Following the Ramsey test, Predict' is translated as

\[
\text{(3.3)} \quad M_\beta \models \alpha \Rightarrow \beta
\]

Let's recall that we needed to evaluate Predict with respect to \( K_\beta \) in order to avoid automatic satisfaction when \( \beta \) was already accepted in \( K \). Let's start from \( \beta \in K \), and assume EpStatus is satisfied; hence, \( \alpha \in K \). The belief set \( K_\beta \) contains all \( K \)-worlds together with all most plausible \( \neg\beta \)-worlds in \( M \). Then, since each \( K \)-world was an \( \alpha \)-world, \( \neg\alpha \notin K \). Then, revising \( K_\beta \) by \( \alpha \) is just a consistent revision (or an expansion), selecting the \( \alpha \)-worlds from \( K_\beta \). \( \beta \) belongs to \( (K_\beta)_\alpha \) if and only if none of the \( \alpha \)-worlds is a \( \neg\beta \)-world. This will only be the case if and only if each of the "closest" \( \neg\beta \)-worlds in \( M \) is \( \neg\alpha \)-world. But this is exactly what Condition Absent tests for, namely, \( \neg\alpha \in K_{\neg\beta}^\alpha \)

We have just proved:

\(^5\)Boutilier's (1992e) work on "natural revision models" extends the AGM theory of revision. Boutilier's natural revision function allows for iterated revisions, preserving a maximum conditional information. The models \( M^\alpha_\beta \) and \( M_{\neg\beta} \) for \( K_\beta \), \( K_{\neg\beta} \) respectively, are fully defined and subsequent revisions and contractions become possible to calculate.
Proposition 3.1 Let $\alpha, \beta \in K$. Then, $\beta \in (K_{\beta})^*_{\alpha}$ iff $\neg \alpha \in K_{\beta}$.

This proposition has a crucial consequence. It says that whenever the explanandum is already believed, non trivial testing of Predict is equivalent to testing Condition Absent. As a result, we can discard equation 3.3. We will keep a formulation of Predict

\[
(3.4) \quad M \models \alpha \Rightarrow \beta
\]

A similar situation holds for Absent. In order to test it non-trivially it should be evaluated with respect to $K_{-\alpha}$; that is, we should test whether $\neg \alpha \in (K_{-\alpha})^*_{\beta}$. Again we confront the problem that in order to calculate revisions of $K_{-\alpha}$, the model $M_{-\alpha}$ is underdefined. Following the Ramsey test, Absent' is translated as

\[
(3.5) \quad M_{-\alpha} \models \neg \beta \Rightarrow \neg \alpha
\]

As with condition Predict, we can show that we can get away without defining $M_{-\alpha}$. Assume $\neg \beta \in K$. By EpStatus $\neg \alpha \in K$ as well. The set $K_{-\alpha}$ consists of all $K$-worlds together with the most plausible $\alpha$-worlds in $M$. As a result, it is guaranteed that $\beta \notin K_{-\alpha}$. Now, revising $K_{-\alpha}$ by $\neg \beta$ is just a consistent revision (or an expansion); $(K_{-\alpha})^*_{\neg \beta}$ is formed by the $\neg \beta$ worlds in $K_{-\alpha}$. Evaluating Absent is just to test whether such $\neg \beta$-worlds are $\neg \alpha$ worlds. This would be the case if and only if none of the most plausible $\alpha$-worlds in $M$ are $\neg \beta$ worlds; in other words, each of the most plausible $\alpha$-worlds should be a $\beta$-world. But this is precisely what Predict states. Thus, we have proved the following proposition:

Proposition 3.2 Let $\neg \alpha, \neg \beta \in K$. Then, $\alpha \in (K_{-\alpha})^*_{\neg \beta}$ iff $M \models \alpha \Rightarrow \beta$.

Then, we just express Absent:

\[
(3.6) \quad M \models \neg \beta \Rightarrow \neg \alpha
\]

Given propositions 3.1 and 3.2, if Conditions EpStatus, Predict and Absent are satisfied, so are Predict' and Absent'. This allows for a pretty simple definition of predictive explanation, just relative to the $K$-revision model $M$. 
Definition 3.1 Let $\alpha, \beta \in L_{CPL}$. $\alpha$ is a predictive explanation for $\beta$ relative to $K$ iff (i) $M \models (B\alpha \equiv B\beta) \land (B\neg\alpha \equiv B\neg\beta)$; (ii) $M \models \alpha \Rightarrow \beta$; and (iii) $M \models \neg\beta \Rightarrow \neg\alpha$.

Consider the following example. Suppose we know that “if it rained the grass would be wet. So would be the case if the sprinkler had been on”. In addition we know that the “sprinkler” and “rain” are mutually exclusive conditions, since we live in an area where people do not water in the rain. Suppose we know that presently “the grass is wet”, and being in winter Vancouver, “rain” is more plausible than “sprinkler”. Let’s denote “sprinkler”, “rain”, and “wet-grass” with the lower-case letters $s$, $r$, $w$, respectively. We represent the above knowledge as $KB = \{-w, -r, -s\}$ and the conditionals $s \Rightarrow w$, $r \Rightarrow w$, $r \Rightarrow \neg s$ and $s \Rightarrow \neg r$ and the plausibility statement $r < s$. Figure 3.3 shows a corresponding $K$-revision model. Even though the grass is not wet, we may question,
“Why would the grass get wet?” “Rain” is a predictive explanation for “wet-grass”, and so is “sprinkler”. Notice that in the model r, s and w are all rejected in K (hence, they have the same epistemic status) and r \Rightarrow w, s \Rightarrow w, and trivially \neg w \Rightarrow \neg s, \neg w \Rightarrow \neg r. Also, r \land s explains w, and r \lor s explains w.

According to Propositions 3.1 and 3.2, for the cases \beta \in K and \neg \beta \in K, we just need to test only one of conditions Predict or Absent, the other being trivial. We can show that when \beta is indeterminate in K the same simplification holds.

**Proposition 3.3** Let \alpha be a predictive explanation for \beta such that \alpha is indeterminate in K. Then, Predict and Absent are equivalent.

We will now examine some extra conditions on explanations that may be of interest. Once we have identified a predictive explanation, we may want to determine whether it covers all the most plausible situations where the explanandum arises. We propose:

“Had the explanation been absent, the observation would have been absent as well”.

This condition can be expressed as:

\[(Cover)\quad \neg \beta \in K_{\alpha}^{*}\]

Cover can be seen as the requirement that the explanation cover all possible hypotheses that could account for the explanandum. As it stands, Cover trivially holds when the explanation and explanandum are rejected in K. Therefore, non-trivial evaluation of it requires us to move into an hypothetical epistemic state where the explanandum is indeterminate, K_{\neg \beta}. Again, we know that K_{\neg \beta} = K when \neg \beta \notin K, and Cover can be re-expressed as:

\[(Cover')\quad \neg \beta \in (K_{\neg \beta})_{\alpha}^{*}\]

Via the Ramsey test we formulate Cover' as

\[(3.7)\quad M_{\neg \beta} \models \neg \alpha \Rightarrow \neg \beta\]
It is interesting to notice what it means for an explanation $\alpha$ to fail to obey to $\textit{Cover}$. Such a situation arises if at some of the most plausible states of affairs where $\beta$ holds, $\alpha$ does not. This may tell us that either there is another explanation accounting for $\beta$ when $\alpha$ fails to do so.

Let's see it with one example. Assume again the situation of our previous rain-sprinkler example, but suppose that it is presently raining. We have $KB = \{w, r, \neg s\}$, the conditionals $s \Rightarrow w$, $r \Rightarrow w$, $r \Rightarrow \neg s$ and $s \Rightarrow \neg r$, and the plausibility statement $r < s$. Figure 3.4 shows a corresponding $K$-revision model. Then, “rain” is a predictive explanation for “wet-grass” ($r, w$ are accepted in $K$, $r \Rightarrow w$, and $\neg w \Rightarrow \neg r$). However, “rain” fails to cover all possible reasons for “wet-grass”, for instance “sprinkler” could...
also be a candidate. Therefore, "rain" is not a covering (predictive) explanation for "wet-grass", since even if it didn’t rain, the grass could have been wet anyway. This is precisely indicated by \( \neg r \neq \neg w \). A predictive explanation \( \alpha \) may fail the covering condition if \( \alpha \) is too specific an instance of an actual explanation that would fully account for the explanandum. Imagine an explanation conjoined with any irrelevant detail, that by chance happens to be true. For example, "hepatitis would fully account for jaundice", while "hepatitis_and_wet_grass" is an irrelevant coincidence. At the most plausible situations where the conjunction "hepatitis_and_wet_grass" does not hold we expect "hepatitis" still be true. Then, "jaundice" is perfectly consistent with the situations where the conjunction "hepatitis_and_wet_grass" does not hold. We return to this in Section 3.4.2.

Shortly, we will see that Cover is very closely related to the following condition, which we name Correl that checks for the correlation between the explanation and explanandum. Let’s recall that Predict demands that whenever an explanation \( \alpha \) holds, \( \beta \) should hold. \( \alpha \) and \( \beta \) are correlated, if, in addition, whenever \( \beta \) holds \( \alpha \) holds as well. We propose,

"Were the explanandum believed, so too would the explanation".

This can be expressed as:

\[(Correl) \quad \alpha \in K^*_\beta\]

As can be expected, Correl is trivial for inquiries that are already in \( K \). Non-trivial evaluation of it should be performed from the epistemic state where the explanation is indeterminate, i.e. \( K^- \alpha \). When \( \alpha \not\in K, K^-_\alpha = K \); then, we reformulate Correl as

\[(Correl') \quad \alpha \in (K^-_\alpha)_\beta\]

We express Correl' as

\[(3.8) \quad M^-_\alpha \models \beta \Rightarrow \alpha\]

With similar reasoning used to relate Predict' and Absent', we can relate Cover' and Correl'.
**Proposition 3.4** Let $\alpha$ be a predictive explanation for $\beta$, such that $\neg \beta \in K$. Then, $\neg \beta \in (K_{\neg \beta})^*_{\alpha}$ iff $M \models \beta \Rightarrow \alpha$.

**Proposition 3.5** Let $\alpha$ be a predictive explanation for $\beta$, such that $\beta \in K$. Then, $\alpha \in (K_{\beta})^*_{\alpha}$ iff $M \models \neg \alpha \Rightarrow \neg \beta$.

**Proposition 3.6** Let $\alpha$ be a predictive explanation for $\beta$, such that $\beta$ is indeterminate in $K$. Then, Correl and Cover are equivalent.

We can give the following definition of a “covering” explanation.

**Definition 3.2** $\alpha$ is a covering explanation for $\beta$ iff (i) $\alpha$ is a predictive explanation for $\beta$; (ii) $M \models \beta \Rightarrow \alpha$; and (iii) $M \models \neg \alpha \Rightarrow \neg \beta$.

Covering explanations are really strong. They give all possible reasons that predict the explanandum.

By weakening Condition $\text{EpStatus}$ into $\text{Commit}$, we can give an alternative definition of an explanation, that we will call quasi-predictive. As we can expect, quasi-predictive explanations are intimately related to their predictive cousins. But we will see later, that the quasi-predictive notion will accommodate explanation that will not be preferred. Quasi-predictive explanations satisfy $\text{Commit}$, $\text{Predict}$ and $\text{Absent}$ (that is, expressions 3.1, 3.4 and 3.6).

**Definition 3.3** $\alpha$ is a quasi-predictive explanation for $\beta$ relative to $K$ iff (i) $M \models B\beta \supset B\alpha$; (ii) $M \models \alpha \Rightarrow \beta$; and (iii) $M \models \neg \beta \Rightarrow \neg \alpha$.

Let’s use our rain-sprinkler example to illustrate a quasi-predictive explanation. Suppose now that we positively know that it did not rain. But we don’t know whether the grass is wet, nor whether the sprinkler is on. Then $KB = \{\neg r\}$, and $s \Rightarrow w$, $r \Rightarrow w$, $r \Rightarrow \neg s$ and $s \Rightarrow \neg r$. Notice that we have to give up the plausibility statement $r < s$.

Figure 3.5 shows the corresponding $K$-revision model. “Rain” is a quasi-predictive explanation for wet grass. But not a predictive one, since “rain” is considered epistemically impossible while “wet grass” is not. In contrast, “sprinkler” is a predictive explanation for “wet grass” (“sprinkler” and “wet grass” have the same epistemic status).
3.2 Non-Predictive Explanations

In this section we will concentrate in non-predictive explanations, capturing the notion that an explanation might yield the explanandum. We will call them “might” explanations. In defining them, we will keep our commitment to Condition Commit which assures that if the object of inquiry is already believed, an explanation should also be believed. “Might” explanations are not predictive. Here is an example. As we illustrated before, if we ask: “Why might Fred contract AIDS?” The answer: “Fred practicing unsafe sex” is just a “might” explanation. The notion asserts that an explanation is consistent with the explanandum; that is, it makes it possible. “Might”-explanations can be associated with the idea of excuses: believing in the excuse may not be sufficient to induce belief in the explanandum. Excuses just render the explanandum possible. This idea will turn out to be related with consistency based diagnosis, where some components assumed to be abnormal make the unexpected observation possible.
Capturing the notion of “might” explanations we propose,

“Were an explanation believed, the explanandum might be believed”

This says the explanandum should be regarded possible once we accept an explanation. The following condition reflects this consistency notion of “might” explanations.

\[(\text{Might}) \quad \neg \beta \not\in K^*_{\alpha}\]

As expected, if \(\beta\) is already epistemically possible (i.e., accepted or indeterminate), any arbitrary belief \(\alpha\) satisfies \(\text{Might}\). However, it is not clear how to avoid this triviality, or whether it makes any sense to ask for a “might” explanation of something we already consider possible. This also points the weak character of “might” explanations. \(\text{Might}\) is expressed as:

\[(3.9) \quad M \models \alpha \not\models \neg \beta\]

We give the following definition of “might” explanations.

**Definition 3.4** \(\alpha\) is a *might explanation* for \(\beta\) iff (i) \(M \models B\beta \supset B\alpha\); and (ii) \(M \models \alpha \not\models \neg \beta\).

Let’s contrast a “might”-explanation with a non-explanation. Believing in a “might”-explanation should make the explanandum at least somewhat possible; then, a non-explanation should make the explanandum impossible. That is, \(\alpha\) is a non-explanation if \(\alpha \models \neg \beta\).

Finally, we introduce another kind of non-predictive explanation. Consider this example.

Fred wants to discolor his T-shirt. He bought a new bleaching product that only works in hot water. Fred is about to do his laundry, and he recalls that the plumbing has just been under repair, so the water may not be hot enough. Were Fred to wash his T-shirt with the product, it might not discolor. However, it is equally plausible to say, that were Fred to wash his T-shirt with the new product, it might discolor.

These are a special kind of “might” explanations. They capture the idea that in some set of equally plausible situations, the explanation accounts for the explanandum.
the product explains in this sense a discoloured T-shirt, since in all of the most normal situations where Fred uses the bleaching liquid and the water is not too cool, the T-shirt gets bleached. However, using the product also explains a non-bleached T-shirt, since in all of the most normal situations where the bleaching liquid is in cold water, we expect no bleaching power. We will call these explanations "weak" explanations. In order to shape them in a logical calculus, we are required to think in terms of preorder-revision models (CT4O models). These models allow us to reason about incomparable scenarios modeled as different sets of plausible states of affairs. The weak conditional connective $\rightarrow$ in CT4O can be used to map weak-explanations.

**Definition 3.5** $\alpha$ is a weak explanation for $\beta$ iff (i) $M \models B\beta \supset B\alpha$; (ii) $M \models B\neg\alpha \supset B\neg\beta$; and, (iii) $M \models \alpha \rightarrow \beta$.

Figure 3.6: Preorder-revision model for the bleaching-T-shirt example.

Let's analyze Fred's bleaching-problem. We know that "if the bleaching product is used in hot water, the T-shirt gets bleached" and "if the bleaching product is not used
in hot water, the T-shirt does not get bleached". Let's denote "new-product", "hot-water", and "bleached-T-shirt" with the lower-case letters $n$, $p$ and $b$ respectively. We will represent the fact that we don't know whether the water might be hot enough, as incomparable scenarios in a CT4O-model. We represent the circumstance that Fred has not yet used the bleaching-product and that the T-shirt is not bleached as $KB = \{ \neg n, \neg b \}$, and the conditional knowledge as: $n \land h \Rightarrow b$, $n \land \neg h \Rightarrow \neg b$. Figure 3.6 shows the corresponding preorder revision-model. "New-product" is a weak-explanation for "bleached-T-shirt" since the model satisfies $n \rightarrow b$, and $n$ and $b$ are rejected in $K$ (then, $B \neg n \supset B \neg b$, and trivially $B b \supset B n$). Notice also that the model satisfies the conditional $n \rightarrow \neg b$ which indicates that under some set of equally most plausible situations where Fred uses the new product, the T-shirt does not get bleached (namely in cool water). This is why "new product" is a weak-explanation instead of a predictive one.

As we studied in Section 2.4.1, in total-order revision models (CO-models) $\rightarrow$ becomes equivalent to its stronger cousin $\Rightarrow$. Then, if the model $M$ is a total-order-revision model, then a weak explanation is just a quasi-predictive explanation. It is clear that predictive explanations are the strongest explanations we presented, and automatically satisfy the conditions of the non-predictive; that is, a predictive explanation is always a weak explanation and a might explanation. Similarly, a weak explanation is automatically a might explanation.

### 3.3 Preferences in terms of Plausibility

We have studied a number of conditions that relate explanation and explanandum. However, they provide us no indication of preference among the many possible explanations. A preferred explanation should not only make the explanandum less surprising but be itself as unsurprising as possible. We will base our notion of preference on the plausibility ordering of states of affairs (which is just induced by the objective belief set and the conditionals held by the agent). We propose a definition of comparative preference among explanations based on Boutilier's definition of plausibility. In this section we will
refer generically to "an explanation" without qualification, to mean that it can be of any type, i.e., predictive, might, etc.

**Definition 3.6** Let $\alpha$ and $\alpha'$ be explanations for $\beta$. $\alpha$ is at least as preferred as $\alpha'$, iff $\alpha \preceq \alpha'$ (namely, $M \models \Box(\alpha' \supset \Diamond \alpha)$).

$\alpha$ is considered at least as preferred as $\alpha'$ just when $\alpha$ is at least as plausible as $\alpha'$. $\alpha$ and $\alpha'$ are equally preferred whenever $\alpha \leq \alpha'$, and $\alpha' \leq \alpha$. In preorder revision-models not all propositions are comparable; then, we can have explanations $\alpha, \alpha'$ that are incomparable.

We will define an explanation as a preferred one if it is in the cut of the most preferred explanations for $\beta$.

**Definition 3.7** Let $\alpha$ be an explanation for $\beta$. $\alpha$ is a preferred explanation for $\beta$ iff there is no other explanation $\alpha'$ such that $\alpha' < \alpha$.

In our rain-sprinkler example of Figure 3.3, we have "rain" and "sprinkler" predictive explanations for "wet-grass", where "rain" is a preferred explanation, while "sprinkler" is not. Let's notice that the trivial explanation is always a preferred predictive explanation for itself (it is trivially quasi-predictive as well).

**Proposition 3.7** $\beta$ is always a preferred (quasi)-predictive explanation for $\beta$.

And $\beta$ is always at least as preferred as any other (quasi)-predictive explanation.

**Proposition 3.8** For any predictive or quasi-predictive explanation $\alpha$ for $\beta$, $\beta \preceq \alpha$.

In the case of predictive and quasi-predictive explanations we can provide a test that determines whether $\alpha$ is a preferred explanation for $\beta$. We should simply ask for an explanation $\alpha$ to be consistent with all the most normal scenarios where the explanandum $\beta$ holds.

**Proposition 3.9** Let $\alpha$ a (quasi)-predictive explanation for $\beta$. $\alpha \preceq \beta$ iff $M \models \beta \not\rightarrow \neg \alpha$. 
Equivalently, an explanation $\alpha$ is not preferred iff $M \models \beta \rightarrow \neg \alpha$.

The following proposition shows that preferred quasi-predictive explanations and preferred predictive ones coincide. This means that the difference between the predictive and quasi-predictive notions arises among non-preferred explanations.

**Proposition 3.10** $\alpha$ is a preferred quasi-predictive explanation for $\beta$ iff $\alpha$ is a preferred predictive explanation for $\beta$.

A criterion of preference in terms of plausibility sanctions that epistemically possible explanations be preferred over the epistemically impossible. It also indicates preferences among the different epistemically impossible explanations, suggesting the least divergent from theory $K$. However, such a criterion provides absolutely no indication of preferences among explanations that are already epistemically possible. If $\alpha$ is accepted or indeterminate, it already enjoys the highest plausibility, since $\alpha$-worlds are among the most plausible worlds. Consequently, we have the next proposition.

**Proposition 3.11** Let $\neg \beta \not\in K$, (i.e, $\beta$ is accepted or indeterminate), and $\alpha, \alpha'$ be any predictive explanations for $\beta$. Then, $\alpha \leq \alpha'$ and $\alpha' \leq \alpha$.

At first view, we could argue that one way to compare explanations that are accepted or indeterminate in $K$, is to perform the comparison with respect to a hypothetical epistemic state where the explanandum $\beta$ is rejected, namely $K_{\neg \beta}$. Then, preference should arise from the ordering of worlds in the revised model. Once again, we confront the problem that $M_{\neg \beta}^*$ is left underspecified by the AGM semantics. By assuming Boutilier's natural revision function, which defines it, we can show that the idea of comparing predictive explanations in such a model is rapidly defeated.

**Proposition 3.12** Let $\neg \beta \not\in K$. If $\alpha, \alpha'$ are predictive explanations for $\beta$, then $\alpha, \alpha'$ are equally plausible in $M_{\neg \beta}^*$.

We conclude that there can be no grounds for preferences for explanations of epistemically possible explanandums.
Chapter 3. Epistemic Explanations

3.4 Pragmatics of Explanation

In any system for explanations ultimately a sentence has to be returned. Not all the sentences sanctioned by the semantic conditions we proposed look exactly as we may expect. That is, several explanatory sentences may denote the same proposition that satisfies our desiderata. In this section we will briefly discuss trivial explanations, explanations including irrelevant information, and some problem with disjunctions in explanations. We think it might be useful to rule these problems out as a matter of pragmatics of explanation, much like Gricean maxims. We regard them as pragmatic issues since they relate to how explanations are used in different contexts or applications.

3.4.1 Trivial Explanations

In many applications, when non-trivial explanations are possible, trivial explanations are not desirable. As expected, trivial explanations "trivially" satisfy the required conditions for predictive explanations, hence, for the weaker counterparts as well. EpStatus is automatically satisfied, since \( \beta \) has the same epistemic status as \( \beta \). So are Predict and Absent, since \( M \models \beta \Rightarrow \beta \) (theorem ID is derivable in CO and CT4O). Moreover, a trivial explanation satisfies Correl and Cover.

**Proposition 3.13** A trivial explanation for \( \beta \) is a predictive covering explanation for \( \beta \).

In order to address the issue of trivial explanations from a semantic perspective, we may explore the following question. Given an epistemic state \( K \), *When does there exist a non-trivial explanation for \( \beta \)?* Or, conversely, *When does \( \beta \) have only a trivial explanation?* Levesque (1989) suggested that if \( K \) is completely ignorant, nothing is predictively explainable but trivially. In the same line of reasoning, if \( \beta \in K \) and \( \beta \) is all we know in \( K \), \( \beta \) is only trivially explainable. And, if \( \neg \beta \) is all we know in \( K \), \( \beta \) is only trivially predictively explainable. There are possibly many cases other than the ones just described. Identifying them may illuminate how to attack semantically the problem of
trivial explanations. In general trivial explanations are undesirable for their uninformative nature; however, when non-trivial explanations are non-existent, trivial explanations may be desirable. Besides, in certain applications the trivial explanation may be a true candidate.

We speculate that explanations that subsume the explanandum, and those that are subsumed by the explanandum may also lead to a notion of un informativeness. (Namely, given an explanandum $\beta$, $\alpha$ lies in this class if either the proposition denoted by $\alpha$ is included in the one denoted by $\beta$, or the other way around: $\|\alpha\| \subseteq \|\beta\|$ or $\|\beta\| \subseteq \|\alpha\|$.)

### 3.4.2 Irrelevant Factors and Disjunctions in Explanations

Explanation sentences that have incorporated irrelevant factors (say, they are a conjunction of a proposition that accounts for the observation and a completely irrelevant proposition that just happens to be consistent) sanction commitment on issues that should be irrelevant, hence not part of an explanation. Say, if rain is an explanation for wet_grass, so could be $\text{rain} \land \text{my\_car\_is\_red}$, if at the set of most plausible states of affairs where rain holds, $\text{rain} \land \text{my\_car\_is\_red}$ also holds; then at the most plausible scenarios, the two denote the same proposition.

A related problem is that explanations may include background beliefs. With background belief we refer to those accepted and non-contingent credences; an example can our belief “the day has 24 hours”. According to the conditions we imposed on explanations, if $\gamma$ is a background belief, $\alpha \land \gamma$ is an explanation for $\beta$ whenever $\alpha$ is. So, $\text{rain} \land \text{the\_day\_has\_24\_hs}$ is as good an explanation as $\text{rain}$. The sibling case is given by the disjunction of an explanation with the negation of a background belief; namely $\alpha \lor \neg \gamma$ is as good an explanation for $\beta$ as $\alpha$ is. Then, $\text{rain} \lor \neg (\text{the\_day\_has\_24\_hs})$ is as good as $\text{rain}$. Background beliefs may be captured by the schema $M \models \Box (\gamma \supset \Box \gamma)$, (indicating that if $\gamma$ holds at some world $w$ then it holds at every world at least as plausible as $w$).

Our semantic analysis points to a preferred set of propositional valuations in the revision model where explanation and explanandum hold. We can see such valuations
indicating a "plain" explanation sentence, that would include background beliefs. From this plain sentence some "filtering" may be applied. A "simplest" sentence may be formulated, for example, using the comparison of sentence-simplicity as defined by Levesque (1989). Interestingly, not all applications based on abductive reasoning seek the same kind of explanations. For example, in image interpretation, an explanation (i.e. an interpretation of an image) conjoined with background beliefs about what the scene is likely to be, may be desirable.
In this chapter we take Theorist as a representative of the "abductive" approaches to explanation and recast it in our framework. We will achieve this by interpreting the background epistemic state of the agent (or program) as a theory of expectations ((Gärdenfors 1990), (Makinson and Gärdenfors 1990) (Gärdenfors and Makinson 1991), (Boutilier 1992c), (Boutilier 1992f)). Instead of attributing a subjunctive interpretation to conditionals, they will denote defaults or expectations. We will show why Theorist explanations are paraconsistent and non-predictive, and how they can be made predictive. Then we will map in our model Brewka’s extension of Theorist, which provides a prioritized default setting. We will define a weak and a predictive notion of explanation in Brewka setting. We will elaborate on the effect of priorities over defaults when used abductively. We will propose how to augment Theorist and also Brewka’s extension to provide preferences among explanations. Capturing the Theorist framework and Brewka’s extension will evidence the expressive power and generality of our modeling of explanation.

4.1 Capturing Theorist

Our goal is to express Theorist explanations as conditional sentences. We will first define a Kripke model for the Theorist framework that we will call $M_D$. Then we will identify a class of models to which $M_D$ belongs to, a modal system that characterizes such class and allows us to express global conditional statements. As Theorist defines a framework for both default and abductive reasoning, we should construct a model that satisfies the
conditionals corresponding to both skeptical default inferences as well as abductive inferences. Our modeling should contemplate that Theorist’s sets of defaults and conjectures need not be consistent, and are not closed under (classical) logical consequence. Theorist framework is based on maxiconsistent constructions from the set of defaults. This makes it syntax-sensitive with respect to the defaults (since depending on how the defaults are expressed, different maximal subsets of defaults may be derived). Our mapping of Theorist will not consider constraints.

As usual, we take \( P \) a denumerable set of atomic variables, and \( L_{CPL} \) denotes the propositional language over this set. Sets \( \mathcal{F}, \mathcal{D}, \mathcal{C} \) will be taken as sets of propositional sentences, consisting of ground instances of the formulae sanctioned by the Theorist facts, defaults and conjectures. We should assume a fixed set of defaults \( 1 \).

We will define a Kripke model \( M_D \), based on Boutilier’s (1992f) key the idea of counting default violations. This model will be just relative to the set of defaults \( D \); then, it will be suitable for any set of facts \( \mathcal{F} \). Following Theorist syntax-sensitive spirit, we define for each world its default-violation set.

**Definition 4.1** For any valuation \( w \in W \), the set of defaults violated by \( w \) is denoted \( V(w) = \{ d : w \text{ falsifies default formulae } d \in D \} \).

Now, the ordering of worlds is induced by the set inclusion ordering among each world’s default-violation-set. If we interpret defaults as normality assumptions, the accessibility relation orders worlds according to their degree of normality.

**Definition 4.2** The Theorist model for \( D \) is denoted by \( M_D = \langle W, R, \varphi \rangle \), where \( W \) is the set of possible worlds such that all propositional valuations are present\( ^2 \); \( \varphi \) maps \( P \) into \( 2^W \) (\( \varphi(A) \) is the set of worlds where \( A \) holds); and \( R \) is defined as follows: for each \( v, w \in W, wRv \text{ iff } V(v) \subseteq V(w) \).

---

\(^1\)This requirement could be relaxed if we defined a “revision” of the Theorist model \( M_D \) by a new default sentence \( d' \), hence, obtaining a new model \( M'_D = (M_D)_{d'}^* \). We could define the contraction of \( M_D \) by an existing default sentence, obtaining a new model \( M''_D = (M_D)_{d}^{d'} \), where the old default \( d \) has been “forgotten”. However, forgetting defaults is not contemplated in Theorist.

\(^2\)Modeling Theorist with constraints may require to exclude some propositional valuations from \( W \).
Chapter 4. Capturing “Abductive” Approaches

It should be clear that \( wRv \) and \( vRw \) if and only if \( V(v) = V(w) \). The situation where \( wRv \), but not \( vRw \) arises only when \( V(v) \subset V(w) \). The definition gives rise to a model forming a lattice. \( R \) divides the set of worlds into clusters. Those worlds violating exactly the same set of defaults are in the same cluster. A world \( w \) can “see” another world \( v \) if and only if the set of defaults violated by \( w \) contains the set of defaults violated by \( v \). A world \( w \) is in a minimal cluster of the lattice if there are no other worlds that violate a proper subset of defaults violated by \( w \). Not all worlds are “comparable” with respect to \( R \). For any two worlds \( w, v \), if the default-violation set of world \( v \) is not included in the default-violation set of \( w \), nor the other way around, then \( w \) can not see \( v \), and \( v \) can not see \( w \). Therefore, the accessibility relation \( R \) is not connected on \( W \). \( M_D \) lies within the class of reflexive and transitive models. Our definition of a model \( M_D \) corrects Boutilier’s (1992f) modeling of Theorist, which assumed total connectedness a (CO-model).

**Proposition 4.1** \( M_D \) is a \( CT4O^* \) model.

Theorist models are well-founded models, which means that they satisfy the so-called Limit Assumption. This assumption states that in each model there exists some minimal \( A \)-world for any proposition \( A \). That is, Theorist models are “bottomed”.

**Proposition 4.2** Let \( M_D = \langle W, R, \varphi \rangle \) be the Theorist model for \( D \). Let each single default formula \( d \in D \) be satisfiable. \( M_D \) possesses a single \( R \)-minimum cluster iff \( D \) is consistent.

As the set of defaults \( D \) is finite, the corresponding model \( M_D \) will have a finite number of \( R \)-minimal clusters. Let’s analyze the famous University Students example. Let \( u \) denote “university student”, \( a \) denote “adult”, and \( e \) denote “employed”. Figure 4.7 shows the \( CT4O^* \) model for the set of defaults \( D = \{ u \supseteq a, u \supseteq \neg e, a \supset e \} \). For convenience we name the defaults as follows: \( h1 : u \supseteq a, h2 : u \supseteq \neg e, h3 : a \supset e \). Let’s write for each world its default-violation set.
The key concept is to characterize extensions of \( (\mathcal{F}, \mathcal{D}) \) in \( M_D \). According to Poole's definition, an extension (see definition 2.2) is the set of logical consequences of \( \mathcal{F} \) together with a maximal consistent subset of defaults of \( \mathcal{D} \). Then, each extension results in minimal (subset related) default violations. We will introduce some terminology. Let's assume a \( X \) to be any consistent set of propositional formulae.

**Definition 4.3** Let \( M = \langle W, R, \varphi \rangle \) any model for which \( R \) is transitive (and we will always assume reflexive). A world \( w \in W \) is \( R \)-minimal \( X \)-world iff \( M \models_w X \), and for every world \( v \in W \), such that \( wRv \) but not \( vRw \), \( M \models_v \neg X \).

A minimal \( X \)-cluster is a maximal set of mutually accessible \( R \)-minimal \( X \)-worlds. We
say that a set of worlds $U$ characterizes a set $Y$ such that $Y = Cn(Y)$ if an only if $U = \|Y\|$. 

Lemma 4.3 $S$ is an extension of $(X, D)$ iff there is a minimal $X$-cluster that characterizes $S$.

Theorist predictions correspond to what all extensions have in common. As shown by Boutilier, if we see a default theory as a theory of expectations, skeptical default conclusions can be understood as the revision of the default theory by the facts, where we want to keep as many expectations as possible without forcing inconsistency. Under this reading, predictions correspond to the AGM revision of the default theory by the facts. The Ramsey test equates $B \in K_\delta$ with $A \Rightarrow B$. The next theorems, lemmas and propositions assume a Theorist pair $(\mathcal{F}, D)$, $M_D$ the model for $D$, and denote with $F$ the conjunction of $\mathcal{F}$. The next theorem embodies Boutilier's (1992f) result for Theorist predictions, in our CT4O-model.

Theorem 4.4 $\gamma$ is predicted in Theorist sense from $(\mathcal{F}, D)$ iff $M_D \models F \Rightarrow \gamma$, and $F$ is satisfiable.

Let's illustrate predictions with the university-student example. If we assume as fact that Fred is an adult, that is $\mathcal{F} = \{a\}$, we predict Fred is employed and that he’s not a university student. This is reflected in the model $M_D$ of Figure 4.7 since $M_D \models a \Rightarrow e$ and $M_D \models a \Rightarrow \neg u$. In contrast, if we take as fact that Fred is a university student, that is $\mathcal{F} = \{u\}$, we can not predict him being an adult, nor being employed. We have $M_D \not\models u \Rightarrow a$, and $M_D \not\models u \Rightarrow e$. Notice also that $M_D \not\models u \Rightarrow \neg a$, and $M_D \not\models u \Rightarrow \neg e$.

We should now explore the correspondence of Theorist explanations in $M_D$. A Theorist explanation for an observation $\beta$ (see definition 2.4) is a subset $D$ of defaults and a subset of conjectures $C$ that together with the facts entail the observation, provided that the facts together with the defaults and conjectures are consistent ($\mathcal{F} \cup C \cup D \models \beta$ and $\mathcal{F} \cup C \cup D$ is consistent). Let’s recall that our definition of a Theorist model is just
relative to the set of defaults $\mathcal{D}$; hence, facts $\mathcal{F}$ and conjectures $\mathcal{C}$ play no role in the definition of $M_D$. We would like to express Theorist explanations in CT4O, without making explicit which particular defaults are considered. Poole (see theorem 2.1) states that $\gamma$ is explainable iff $\gamma$ is in some extension of the default theory. Consequently, $\beta$ is in some extension containing $\mathcal{C}$. Obviously we can not use the same connective that we have used for predictions, since a statement $(\mathcal{F} \cup \mathcal{C}) \Rightarrow \beta$ would involve every extension containing the facts $\mathcal{F}$ and conjectures $\mathcal{C}$. Theorist explanations allow just some particular defaults to be assumed. The connective $\longrightarrow$ gives the desired behaviour in CT4O.

**Theorem 4.5** Let $D \subseteq \mathcal{D}$ and $C \subseteq \mathcal{C}$. $D \cup C$ is a Theorist explanation for $\beta$. iff $M_D \models (\mathcal{F} \cup C) \longrightarrow \beta$ and $(\mathcal{F} \cup C)$ is consistent.$^3$

Theorist sanctions that observations inconsistent with the facts are not explainable. This follows from theorem 4.5.

**Corollary 4.6** If $\mathcal{F} \models_{CPL} \neg \beta$ then there is no $C$ such that $M_D \models (\mathcal{F} \cup C) \longrightarrow \beta$ and $(\mathcal{F} \cup C)$ is consistent.

As we noticed when studying Boutilier's logics (Section 2.4.1) the conditional $\longrightarrow$ behaves paraconsistently in CT4O. The next example illustrates that Theorist explanations are in some sense paraconsistent, hence, nicely captured by $\longrightarrow$. Let $\mathcal{F} = Cn(\emptyset)$, $\mathcal{D} = \{a \supset b, a \supset \neg b\}$, and $\mathcal{C} = \{a\}$. Figure 4.8 shows the $M_D$ model for $\mathcal{D}$. The following conditionals are satisfied: $M_D \models a \rightarrow b$, and $M_D \models a \rightarrow \neg b$. So, $a$ is an explanation for $b$ and simultaneously $a$ is an explanation for $\neg b$. On the Theorist side we have the two counterpart explanations. For $D = \{a \supset b\}$, and $C = \{a\}, C \cup D$ explains $b$. While for $D = \{a \supset \neg b\}$, and $C = \{a\}, C \cup D$ explains $\neg b$. This means Theorist explanations are weak explanations, and motivates our development of a stronger notion. Theorist explanations are the counterpart of our "weak"-epistemic explanations when the background epistemic theory is interpreted as a theory of expectations.

---

$^3$Since $(\mathcal{F} \cup C)$ is a set of formulae, the conditional $(\mathcal{F} \cup C) \longrightarrow \beta$ should be understood as the conjunction of the elements in $\mathcal{F} \cup C$ entailing $\beta$. 
Before turning our attention to predictive explanations we illustrate the aptness of our modeling of Theorist explanation. In Theorist, if $\gamma$ is predicted from $(\mathcal{F}, \mathcal{D})$ then $\gamma$ is explainable with no conjectures, and $\neg \gamma$ is not explainable with no conjectures. The following proposition shows that our mapping obeys this behaviour.

**Proposition 4.7** If $M \models F \Rightarrow \beta$ and $F$ is satisfiable then (a) $\beta$ is explainable with no conjectures, and (b) $\neg \beta$ is not explainable with no conjectures.

Notice that the converse of this proposition does not necessarily hold. Assume in our university-student example, that Fred is not employed and he’s a university student or an adult, that is, $\mathcal{F} = \{\neg e, (u \lor a)\}$. We can weakly explain that Fred is university student (by assuming the contrapositive of the default $a \triangleright e$, we conclude $\neg a$ and $u$).

The model $M_D$ of Figure 4.7 sanctions the conditional $\mathcal{F} \rightarrow u$, or in other words, $u$ is in some extension of the default theory. However, we can not explain $\neg u$ since $\neg u$ is in no extension of the default theory (since $\mathcal{F}$ is consistent with $u \triangleright e$ and $u \triangleright \neg e$ and $u \triangleright a$, we can not conclude $\neg u$). This is reflected in the Theorist model since $M_D \nvdash \mathcal{F} \rightarrow \neg u$ (notice that the pair of worlds labeled $w2$ and $w6$ in Figure 4.7 denote an extension).
4.2 Extending Theorist: Predictive Explanations

In this section we will explore the notion of "predictive" explanations. Predictive explanations arise as the "natural" abductive side of default predictions. If some conjectures predictively explain an observation, such conjectures shall not explain its negation. At first view, we could propose \( C \subseteq C \) as a predictive explanation for an observation \( \beta \), iff for any \( D \subseteq D, F \cup D \cup C \models_{\text{CPL}} \beta \), while \( F \cup D \cup C \) is consistent. However, \( \emptyset \subseteq D \), so the notion of predictive explanation would be trivial, since it would just reduce to classical entailment (i.e., \( F \cup C \models_{\text{CPL}} \beta \)). A better statement of predictive explanations in the Theorist setting is given by the following definition.

**Definition 4.4** Let the Theorist triple \((F, D, C)\), and an observation \( \beta \). \( C \) is a predictive explanation for \( \beta \) iff \( \beta \) belongs to all extensions of \((F \cup C, D)\).

An observation is predictively explained by some conjectures if the observation is entailed by the facts together with *every maximal* set of defaults consistent with the facts and conjectures. The next theorem proves that the conditional connective \( \Rightarrow \) captures precisely such notion of a predictive explanation in the Theorist setting.

**Theorem 4.8** \( C \) is a predictive explanation for \( \beta \) iff \( M_D \models (F \cup C) \Rightarrow \beta \), and \((F \cup C)\) is consistent.

This theorem demonstrates that our notion of a predictive explanation naturally lead to our extension of Theorist of Definition 4.4.

Let’s give one example using the university-student model of Figure 4.7. Let’s assume a set of conjectures \( C = \{u, \neg u, e, \neg e, a, \neg a\} \). Suppose as fact that Fred is an adult, \( F = \{a\} \), then we predict that Fred is employed and he’s not a university student \((M_D \models a \Rightarrow e \text{ and } M_D \models a \Rightarrow \neg u)\). If we observe that Fred is not employed then we can explain it (weakly) by conjecturing that he is a university student, since \( M_D \models a \land u \rightarrow e \). But also by conjecturing that he’s a university student we can weakly explain that he is employed \((M_D \models a \land u \rightarrow e)\), contrary to our observation. Then, we can not predictively
explain that he is employed, which is reflected by $M_D \not\models a \land u \Rightarrow e$.

Assume now as a fact that Fred is a university student, so $\mathcal{F} = \{u\}$. By conjecturing that Fred is employed we can predictively explain that he is an adult, as $M_D \models u \land e \Rightarrow a$.

Clearly, we can also weakly explain his adulthood, since $M_D \models u \land e \rightarrow a$. The next proposition shows that predictive explanations are a subset of Theorist (non-predictive) explanations. It also stresses the non-paraconsistent nature of predictive explanations.

We say that a Theorist explanation for $\beta$ assumes conjectures $C$ if and only if there is $D \subseteq D$ such that $\mathcal{F} \cup C \cup D \models \beta$ and $\mathcal{F} \cup D \cup C$ is consistent.

**Proposition 4.9** Let $C \subseteq C$. If $C$ is a predictive explanation for $\beta$ then (a) there is a Theorist explanation for $\beta$ assuming $C$; and (b) $\neg \beta$ is not explainable assuming $C$.

Notice that the converse of this proposition does not necessarily hold, for the same reason that the converse of Proposition 4.4 does not necessarily hold.

### 4.3 Capturing Brewka’s Preferred Subtheories

In this section we will concentrate in Brewka’s extension of Theorist (presented in Section 2.2.4, which adds priorities to defaults. A Brewka theory is a pair $(\mathcal{F}, \mathcal{B})$ such that $\mathcal{F}$ is a set of true facts, and $\mathcal{B} = (B_1, B_2, \ldots, B_n)$ is a set of defaults where the rank $B_i$ is more reliable than $B_j$ iff $i < j$, and no rank of $\mathcal{B}$ needs to be consistent. As we have done with Theorist, we will first define a Kripke model for Brewka’s framework that we will call $M_B$. Boutilier (1992f) stated how Brewka’s framework can be modeled counting default rule violations. We propose a model $M_B$ very related to Boutilier’s; however, his assumption that the model was connected was not right. Each world possesses, for each rank of the default theory $\mathcal{B}$, a hypotheses-violation set.

**Definition 4.5** For any valuation $w \in W$, the set of defaults of rank $i$ violated by $w$ is denoted $V^i_w = \{d \in B_i : w \text{ falsifies default formula } d \}$.

Given a default theory $\mathcal{B} = (B_1, \ldots, B_n)$, each world will have associated $n$ hypothesis-violation sets, one per rank, and we will consider (for technical convenience) an empty
violation set for (non-existent) rank \( n + 1 \), i.e. \( V_w^{n+1} = \{\} \).

We shall induce a normality ordering among worlds, from the reliability ordering of the hypotheses in \( B \). Given a pair of worlds \( v, w \) we will determine the rank \( i \) such that the set of hypotheses violated by \( w \) strictly contains the set of hypotheses violated by \( v \) at rank \( i \), and for all the ranks more reliable than \( i \), the violation sets for \( w \) and \( v \) coincide. We will identify such rank \( i \) with \( \text{min}(v, w) \). Let's recall that the hypotheses with the highest reliability are in \( B_1 \), the next to the highest priority are in \( B_2 \), so on.

**Definition 4.6** Let a default theory \( B = (B_1, \ldots, B_n) \). For worlds \( w, v \), let

\[
\text{min}(v, w) = \min \{ i : (V_w^i \subseteq V_v^i \text{ or } V_v^i \subseteq V_w^i) \text{ and } \forall j \text{ such that } 1 \leq j < i, V_v^j = V_w^j \}
\]

If the above set is empty, let \( \text{min}(v, w) = n + 1 \).

If a world \( w \) violates a hypothesis from a more reliable rank than any of the hypotheses violated by world \( v \), then \( v \) should be considered a more normal world than \( w \). However, if \( v \) and \( w \) violate the same hypotheses up to rank \( i - 1 \), world \( v \) should be considered more normal than \( w \) only if the violation set of \( v \) for rank \( i \) is strictly included in the violation set of \( w \), independently of other hypothesis violations at any less reliable rank.

**Definition 4.7** The Brewka model for \( B = (B_1, \ldots, B_n) \) is denoted by the triple \( M_B = (W, R, \varphi) \), where \( W \) is the set of possible worlds such that all propositional valuations are present; \( \varphi \) maps \( P \) into \( 2^W \) (\( \varphi(A) \) is the set of worlds where \( A \) holds); and \( R \) is defined as follows: for each \( v, w \in W \), \( wRv \) iff \( V_v^i \subseteq V_w^i \), \( \forall i \) s.t. \( 1 \leq i \leq \text{min}(v, w) \).

The definition gives rise to a CT4O-model. Those worlds violating exactly the same hypotheses are in the same cluster. A world \( w \) can “see” another world \( v \) if and only if the set of hypotheses violated by \( w \) at rank \( \text{min}(v, w) \) strictly contains the set of hypotheses violated by \( v \) that rank, and for all more reliable ranks the violation sets for \( w \) and \( v \) coincide. A world \( w \) is in a minimal cluster of the lattice if there are no other worlds that violate a proper subset of hypotheses violated by \( w \).

**Proposition 4.10** \( M_B \) is a CT4O* model.
As expected $M_B$ models satisfies the Limit Assumption (in each model there is some minimal $A$-world for any proposition $A$). Next example shows a $M_B$ model for the University Students example, assuming $B = (B_1, B_2)$, $B_1 = \{h1 : u \supset a, h2 : u \supset \neg e\}$, $B_2 = \{h3 : a \supset e\}$. Let's write for each world its hypotheses violation sets. The correspondent $M_B$ model is shown in Figure 4.9.
Let's characterize extensions in \( M_B \). An extension \( S \) is formed by maximal subsets of hypotheses preferring the most reliable ranks; then it seems natural that the characterization of \( S \) in \( M_B \) be given by most normal worlds. Similar results to those obtained for Theorist hold for Brewka models. Let \( M_B \) the model for a default theory \( B = (B_1, \ldots, B_n) \). Let's assume a \( X \) to be any consistent set of propositional formulae.

**Lemma 4.11** \( S \) is an extension of \((X, B)\) iff there a is minimal \( X \)-cluster characterizing \( S \).

For the following propositions, lemmas and theorems we shall assume a Brewka theory \((\mathcal{F}, T)\), such that \( B = (B_1, \ldots, B_n) \), \( M_B = (W, R, \varphi) \) the Brewka model for \( B \) and \( F \) denoting the conjunction of \( \mathcal{F} \).

**Proposition 4.12** Let each single default formula \( d \in B \) be satisfiable. \( M_B \) possesses a single \( R \)-minimum cluster iff \( B \) is consistent.

Predictions (skeptical default conclusions) are specified by what all extensions have in common, and correspond to Strong B-entailment given the facts. The next theorem shows the correspondence in \( M_B \).

**Theorem 4.13** \( \mathcal{F} \models_{BS} \gamma \) iff \( M_B \models F \Rightarrow \gamma \), and \( F \) satisfiable.

Assume the default theory of the university-students example and \( \mathcal{F} = \{u\} \). Then \( u \models_{BS} \neg e \), since \( \neg e \) is entailed by all extensions of \((\mathcal{F}, B)\). Notice that there is only one extension of \((\mathcal{F}, B)\) and it corresponds to the shaded area in Figure 4.9, the most normal world where \( u \) holds. Hence \( M_B \models u \Rightarrow \neg e \).
Chapter 4. Capturing "Abductive" Approaches

Given a theory \((\mathcal{F}, \mathcal{B})\), and an observation \(\beta\) we can ask What's a premiss \(\alpha\) that together with the facts Strongly-B-entails \(\beta\)? Equivalently, What's a premiss \(\alpha\) such that all extensions of \(\langle (\mathcal{F} \cup \alpha), \mathcal{B} \rangle\) entail \(\beta\)? In a belief revision style, this idea results in the question: By what premiss (together with the facts) should the expectation-theory be revised, such that the observations become believed? The Ramsey test equates \(\beta \in \mathcal{B}^{\mathcal{F} \cup \alpha}_{\mathcal{P}}\) with \((F \land \alpha) \Rightarrow \beta\). We will elaborate on the two notions of explainability from Brewka-theories, that arise from an abductive reading of the two kinds of entailment. Explanations arising from Strong \(B\)-entailment will be referred to as predictive explanations. As a consequence of Theorem 4.13, the following holds:

\[
\mathcal{F} \cup \alpha \vdash_{BS} \gamma \text{ iff } M_{\mathcal{B}} \models (F \land \alpha) \Rightarrow \gamma \text{ and } (F \land \alpha) \text{ satisfiable.}
\]

We say \(\alpha\) is a predictive explanation for \(\gamma\).

The explainability notion arising from Weak \(B\)-entailment is non-predictive. For an observation \(\beta\) we can ask What's a premiss \(\alpha\) such that there is at least one extension of \(\langle (\mathcal{F} \cup \alpha), \mathcal{B} \rangle\) that entails \(\beta\)? This notion corresponds to Theorist explanations but in a prioritized default theory. The conditional \((\mathcal{F} \cup \alpha) \rightarrow \beta\) expresses that there is some extension of \(\langle (\mathcal{F} \cup \alpha), \mathcal{B} \rangle\) in which \(\beta\) holds. A non-predictive explanation accounts for the observation in some of the maximally plausible scenarios.

**Theorem 4.14** \(\mathcal{F} \cup \alpha \vdash_{BW} \beta \text{ iff } M_{\mathcal{B}} \models (F \land \alpha) \rightarrow \beta \text{ and } (F \land \alpha) \text{ satisfiable.}\)

Due to the close correspondence between Theorist and Brewka's frameworks, Theorems 4.13, and 4.14 are not surprising. The same insight about the paraconsistency of \(\rightarrow\) in CT4O applies to Weak \(B\)-entailment.

Let's discuss how a prioritized default theory affects the abductive activity. Since Brewka's framework is a prioritized version of Theorist, and we have provided CT4O models for both, we are able to make a legitimate comparison of the explanations that arise in each. Let's compare the University Students example. The following conditionals are sanctioned in the Theorist model \(M_{\mathcal{D}}\) (see Figure 4.7): \((u \land a) \rightarrow e, (u \land a) \rightarrow \neg e, u \rightarrow e, u \rightarrow \neg e, u \rightarrow a, u \rightarrow \neg a, T \Rightarrow \neg u, (u \land \neg a) \Rightarrow \neg e.\)
The Brewka model $M_B$ sanctions the following conditionals. $(u \land a) \rightarrow \neg e$; moreover, $(u \land a) \Rightarrow \neg e, a \Rightarrow \neg u, (u \land \neg a) \Rightarrow \neg e$; however, $M_B \models (u \land a) \not\rightarrow e$ and $M_B \models (u \land a) \not\Rightarrow e$.

In the model of a prioritized default theory, at least as many constraints on the ordering of worlds are imposed as in the model of the flat theory. Even though a Brewka model isn’t necessarily connected (i.e. CT4O-model), the rate of connectedness may be higher than the counterpart Theorist model. For the next propositions let’s assume a Brewka theory $\langle \mathcal{F}, B \rangle$, where $B = (B_1, \ldots, B_n)$, and a Theorist pair $\langle \mathcal{F}, D \rangle$, such that $D = \bigcup_{i=1}^{n} B_i$, $M_D = \langle W, R_D, \varphi \rangle$ denotes the Theorist model for a $D$, and $M_B = \langle W, R_B, \varphi \rangle$ the Brewka model for $B$. The following proposition shows that $M_B$ is at least as connected as $M_D$.

**Proposition 4.15** If $wR_D v$ then $wR_B v$.

**Proposition 4.16** $M_D$ and $M_B$ have the same clusters.

**Proposition 4.17** Every $R_B$-minimal cluster in $M_B$ is a $R_D$-minimal cluster in $M_D$.

**Proposition 4.18** If $D$ is consistent, then $M_D$ and $M_B$ have the same unique minimal cluster.

We can now show that predictive explanations in Theorist setting are predictive explanations in Brewka’s framework.

**Proposition 4.19** If $M_D \models \alpha \Rightarrow \beta$ then $M_B \models \alpha \Rightarrow \beta$.

We can now show that non-predictive explanations in Brewka’s framework are non-predictive explanations Theorist setting. Thus Brewka model adds (possibly) predictive explanations. The prioritized default setting together with the notion of preferences (see below) can be viewed as “restricting’ what counts as a good explanation.

**Proposition 4.20** If $M_B \models \alpha \rightarrow \beta$ then $M_D \models \alpha \rightarrow \beta$. 
4.4 Preferences: Maximizing Normality

None of the two notions of explanation (predictive and non-predictive) give any indication of preferences among the many different possible choices of what can be hypothesized in order to account for an observation; and as Rescher (1978) pointed out "conjectural fancy is limitless". In this section we introduce a natural notion of preference among explanations, that arises from the ordering of worlds in the model. Roughly, a world is "lower" in the ordering if it violates "fewer" defaults. Now, if defaults represent expectations, normality assumptions or statements of high probability then a world that violates fewer hypotheses is to be considered more "normal". It seems natural that preferred explanations be sanctioned by those most normal worlds where the observation holds. We will define a notion of comparative preference for predictive explanations in Theorist/Brewka setting, where conjectures that are consistent with more (under set inclusion) defaults are preferred. This notion naturally arises from the preferences in terms of plausibility on epistemic explanations that we studied in Chapter 3. Neither Theorist nor Brewka's framework provide a notion of preference; so our proposal really augments their capabilities. Let $F$ denote the conjunction of $\mathcal{F}$, and $\alpha, \alpha'$ the conjunction of elements in $C, C' \subseteq \mathcal{C}$ (or just premisses in $\mathcal{L}_{CPL}$).

**Definition 4.8** Let $\alpha, \alpha'$ be predictive Theorist explanations for $\beta$ given a theory $(\mathcal{F}, \mathcal{D})$. $\alpha$ is a *at least as preferred as* $\alpha'$ (written $\alpha \preceq_{\mathcal{F}} \alpha'$) iff each maximal subset of defaults $D'$ consistent with $\alpha' \cup \mathcal{F}$ is contained in some in some maximal subset of defaults $D$ consistent with $\alpha \cup \mathcal{F}$.

In certain cases two predictive explanations may result *incomparable*; that is, the maximal subset of defaults consistent with one may not be contained in any of the maximal subsets of the other one, nor the other way around. Then we may want to say that both are preferred. So we define a notion of a *preferred* predictive explanation.

**Definition 4.9** Let $\alpha$ be a *predictive* Theorist explanation for $\beta$ given a theory $(\mathcal{F}, \mathcal{D})$. $\alpha$ is a *preferred* explanation for $\beta$ iff there is no other predictive explanation $\alpha'$ such that
each maximal subset of defaults $D$ consistent with $\alpha \cup \mathcal{F}$ is strictly contained in some maximal subset of defaults $D'$ consistent with $\alpha' \cup \mathcal{F}$.

Identical counterpart definitions can be given in Brewka setting. For the theorems and propositions to follow, let’s assume a CT4O model $M$ that indistinctively may reflect a Brewka or a Theorist model. We can show that the definition of comparative preference corresponds to comparative normality in model $M$ (i.e. the corresponding $M_D/M_B$). Since the propositions and lemmas that follow hold for both Theorist and Brewka modeling we will refer generically to “predictive explanations” and “non-predictive explanations” without making explicit under which system.

**Theorem 4.21** Let $\alpha, \alpha'$ be predictive explanations for $\beta$. $\alpha \leq \beta \alpha'$ iff $M \models \Box((\alpha' \wedge F) \supset \Diamond(\alpha \wedge F))$.

This theorem crystallizes the connection with our notion of preferences on epistemic explanations. In Chapter 3 we defined preferences on predictive explanations by comparing the plausibility of the explanations. We based our notion on the idea that an explanation should not only make the observation plausible but be itself maximally plausible. If defaults reflect assumptions of normality, then a preferred explanation in a default setting should be maximally normal; that is, be consistent with as many defaults as possible. We can now show the counterpart results from the ones obtained in Chapter 3 for this preference in terms of normality.

**Corollary 4.22** The preference relation $\leq$ is reflexive and transitive.

An observation $\beta$ is at least as preferred as any predictive explanation for it. A predictive explanation $\alpha$ for an observation $\beta$ sanctions that $\beta$ should belong to all the extensions consistent with $\alpha$ and the facts. Then it is clear that the observation must be consistent with at least as many defaults as the explanation $\alpha$ is. This is shown in the following proposition.

**Proposition 4.23** Let $\alpha$ a predictive explanation for $\beta$, then $\beta \leq \beta \alpha$. 
We may be also interested in preferences among non-predictive explanations. We can apply the idea of maximizing normality. Conjectures $\alpha$ that are consistent with more (under set inclusion) defaults should be preferred. But as non-predictive explanations are paraconsistent (in the sense that they can explain both an observation and its negation), then there can be some extensions of satisfying $\beta$ and others satisfying $\neg\beta$. So when comparing the "normality" of two non-predictive explanations we may want to restrict the comparison to the respective maximal sets of defaults consistent with the observation.

When modeling a prioritized set of defaults (as for example obtaining the priorities by Pearl's (1990) $Z$ ranking of default rules), the ordering of worlds induced by the priorities determines a meaningful preference ordering of worlds in the structure. One can speculate that in a prioritized framework, the inverse of the priority ordering among defaults must play a significant role when drawing explanations. That is, when explaining an observation, a least "exceptional" explanation may be preferred. It seems natural to draw it from the least exceptional rank of hypotheses. This makes sense, as long as we are not violating any of the most reliable hypotheses. This is precisely captured by giving preferring explanations that hold at most normal worlds, since such worlds violate the fewest hypotheses. We have seen that the degree of comparability of states of affairs in a model of a prioritized default theory may be richer than in the correspondent model with no priorities. Predictive and non-predictive explanations that are incomparable in a model of a flat default theory, may be comparable in the model for the correspondent prioritized theory. We also conclude that the priorities in a default framework act as a pruning mechanism, restricting the non-predictive explanations and increasing the predictive ones. A prioritized default setting should give rise to more meaningful preferred predictive explanations.
Chapter 5

Capturing Consistency Based Diagnosis

Model-based diagnosis is an application of abductive reasoning, that brought the attention of many researchers in AI. Different models of diagnosis have been proposed in the literature. "Abductive" diagnosis (as formalized in the Theorist framework) and consistency based diagnosis ((Reiter 1987b),) are today’s canonical qualitative approaches. Poole (1989b), (1990) and Konolige (1992b),(1992a) have investigated in detail the connections between the two frameworks. In this chapter we will recast the consistency based diagnosis framework (presented in Section 2.2.5) in CT4O. We will achieve this by mapping the consistency based model onto a Theorist model, obtaining a unified semantic account for the two approaches to diagnosis. We will demonstrate that a consistency based diagnosis correspond to the idea of an excuse for the conflictive observation, and draw the connection with our (epistemic) “might” explanations of Chapter 3. We show formal correspondence between our proposed notion of preferences on explanations and minimal diagnosis.

5.1 Recasting Consistency Based Diagnosis in CT4O

When doing diagnosis, we want to find out what is wrong with some system. The goal is to explain an observation that conflicts with our expectation of normal behaviour of the system. de Kleer Mackworth Reiter (dKMR) in their consistency based framework assume a system with set of components COMPS, represented as a set of constants; a system description, axiomatized as a set of first order sentences; and a set of observation OBS, represented as a set of first order sentences. The system description axiomatizes the correct behaviour of the system, under the explicit assumption that each component
is working correctly. Knowledge about how problems (symptoms, diseases) manifest themselves can also be included as part of the system description. That is, fault models and/or exoneration axioms can be involved in the description of the system. The set of observations OBS has to be explained. If the observations do not conflict with the expected (correct) behaviour of the system, then every component is assumed to be normal and it is concluded that there is nothing wrong with the system. If, however the observations do conflict with the correct behaviour, then some components must be malfunctioning. Guided by the “principle of parsimony” with respect to component failures, the set of faulty components is considered minimal. This minimal set of abnormal components renders OBS consistent.

Let a toy system of a lamp, formed by a plug and a bulb, such that if both components are working correctly we expect light to be produced. This is encoded in the following system description:

\[-\text{ab.bulb} \land \text{ab.plug} \supset \text{light}\]

Add the constraint that dark and light are mutually exclusive. If we observe OBS = light, this is consistent with our expectation of components being normal; then, OBS is a consistent observation. In contrast, OBS = dark is an inconsistent observation. Evidently, some component must be wrong.

We should define a suitable Kripke model for the consistency based diagnosis framework. We will consider SD a propositional sentence denoting the conjunction of the ground instances of the first order axiomatization of the system description. We denote with \( \beta \) the conjunction of the (ground instances of the) sentences in OBS. dKMR use the literal \( ab(c) \) to indicate that component \( c \) is functioning abnormally, and \( \neg ab(c) \) to indicate that component \( c \) is working correctly. We will just refer to the propositional variable \( \text{ab}_c \) and to its negation \( \neg \text{ab}_c \). dKMR denote a diagnosis for an observation with a sentence \( D(C_p, C_n) \):

\[
\left[ \bigwedge_{c \in \Delta} ab(c) \right] \land \left[ \bigwedge_{c \in \text{COMPS} - \Delta} \neg ab(c) \right]
\]
where the components in $\Delta \subseteq \text{COMPS}$ are assumed to be malfunctioning while the rest, i.e. $\text{COMPS} - \Delta$, are assumed to be working correctly, such that

$$\text{SDUD}(C_p, C_n) \cup \text{OBS} \text{ is consistent}$$

Reiter (1987) showed formal correspondence between his theory of diagnosis from first principles and default logic. He assumed a set of defaults stating that each component is not malfunctioning. Based on his connection, we will assume a set $\mathcal{D}$ containing as many defaults as components, indicating that each component is not abnormal. We express the defaults in propositional language, so $\mathcal{D} = \{\neg \text{ab}_1, \ldots, \neg \text{ab}_n\}$. The principle of parsimony with respect to component failure shall induce a plausibility ordering over possible worlds. Worlds reflecting fewer (under set inclusion) failing components should be regarded as more "normal" than those with more failures. This ordering is precisely captured by a Theorist model for $\mathcal{D}$. We assume a model $M_\mathcal{D}$ for $\mathcal{D} = \{\neg \text{ab}_1, \ldots, \neg \text{ab}_n\}$, as defined in Chapter 4 (Definitions 4.1 and 4.2). We obtained a model that is defined just relative to the set of expectations of normal component behaviour. The model possess a minimal cluster at the bottom denoting all the worlds where all components are assumed normal. Figure 5.10 illustrates a model $M_\mathcal{D}$ for $\mathcal{D} = \{\neg \text{ab}_1, \neg \text{ab}_2\}$, indicating the expectation that each of the two components, $c_1, c_2$ are working correctly. As we studied, $M_\mathcal{D}$ is a CT4O*-model. Worlds violating the same normality assumptions are in the same cluster. Worlds with violating more (set inclusion) normalities have access to worlds violating fewer normalities. Since the model $M_\mathcal{D}$ is defined solely with respect to the defaults $\mathcal{D}$, all worlds satisfying the system description $SD$ are scattered in each cluster. If there are SD-worlds in the minimal cluster of $M_\mathcal{D}$, then those worlds correspond to models of the system with no failures. Obviously such SD-worlds in the minimal cluster (which involve no abnormalities) exist if and only if $\text{SDU} \neg \text{ab}(c)$ for all components $c \in \text{COMPS}$ is consistent. In order to derive diagnosis we will only be concerned in worlds satisfying the system description $1$ SD; that is, worlds violating SD (i.e.

---

1Allowing $\neg$SD-worlds be eligible to explain conflicting observations suggests a way of modeling what people in hardware-verification do. Given an observation that conflicts with the system description may
Chapter 5. Capturing Consistency Based Diagnosis

Figure 5.10: Model reflecting expectation of correct behaviour, for two components.

$\neg$-SD-worlds) will not be of interest.\(^2\) We can see that the system description SD plays a similar role than the set of facts $\mathcal{F}$ in Theorist setting.

For the purpose of our presentation, we will denote with $\alpha \in \mathcal{L}_{CPL}$ the conjunction of abnormal-components, and with $\gamma \in \mathcal{L}_{CPL}$ the conjunction of not-abnormal-components. Namely, $\alpha = \bigwedge_i \neg ab \cdot c_i$ for the components $c_i \in \Delta$, and $\gamma = \bigwedge_j \neg \neg ab \cdot c_j$ for the components $c_j \in \text{COMPS} - \Delta$. The conjunction $\alpha \land \gamma$ is always satisfiable, and denotes a unique possible “state” of the system, such that each component is known to be normal or abnormal.

Let’s assume that all components are working normally. If an observation conflicts with the expected behavior of the system, we conclude that the assumption of correct behaviour for all components can not hold. We define an _excuse_\(^3\) for the observation, as be explained by some components being abnormal or by suggesting that the system could be badly designed. An explanation that involves no abnormalities and contradicts the current system description proposes a new system design.

\(^2\)This suggests a simplification of the model $M_D$ for consistency based diagnosis. A model $M_D = (W, R, \varphi)$, such that the set of possible worlds $W$ includes exclusively just $SD$-worlds ($M_D \models \Box \text{SD}$), obtaining a model that is not _full_.

\(^3\)The term “excuse” is borrowed from Konolige (1992b).
an hypothesis about some components being faulty, which makes the observation possible.

**Definition 5.1** \( \alpha \) is an *excuse* for \( \beta \) iff \( M_D \models SD \land \alpha \neq \neg \beta \), such that \( \alpha = \bigwedge_i ab \cdot c_i \) for components \( c_i \in \Delta \subseteq \text{COMPS} \).

If we interpret \( D \) as a theory of expectations, excuses for an observation \( \beta \) relative to a model \( M_D \) are equated via the Ramsey test with:

\[
\neg \beta \notin D_{SD\land \alpha}^*
\]

\( \alpha \) is an excuse for \( \beta \), when by revising our expectations of normal behaviour accepting \( \alpha \) (together with the system description), we can not regard the observation as impossible.

Excuses in \( M_D \) have their correspondence to “might” explanations in revision models (studied in Chapter 3). We defined \( A \) as a “might” explanation for \( B \) relative to a belief set \( K \) when \( \neg B \notin K^*_A \). Let’s see one example of an excuse for our lamp system. Figure 5.10 shows the \( M_D \) model for the set of defaults \( D = \{ \neg ab\text{-plug}, \neg ab\text{-bulb} \} \), by representing \( ab\text{-plug} \) with \( ab_1 \), \( ab\text{-bulb} \) with \( ab_2 \), and \( \text{light} \) with \( l \). Worlds satisfying the system description \( SD \) are shaded. Let’s assume \( SD = \neg ab\text{-bulb} \land \neg ab\text{-plug} \supset \text{light} \). Let \( dark = \neg \text{light} \). Given our definition of an excuse, \( ab\text{-bulb} \) is an excuse for \( dark \), so is \( ab\text{-plug} \), and so is the conjunction \( ab\text{-plug} \land ab\text{-bulb} \), since \( M_D \models SD \land ab\text{-bulb} \neq \neg dark \), \( M_D \models SD \land ab\text{-plug} \neq \neg dark \) and \( M_D \models SD \land ab\text{-bulb} \land ab\text{-bulb} \neq \neg dark \). The next theorem demonstrates that the concept of an excuse captures precisely the notion of a diagnosis in the consistency based framework.

**Theorem 5.1** \( \alpha \) is an excuse for \( \beta \) iff \( \alpha \land \gamma \) is a diagnosis for \( \beta \), where \( \alpha = \bigwedge_i ab \cdot c_i \) for \( c_i \in \Delta \), and \( \gamma = \bigwedge_j \neg ab \cdot c_j \) for \( c_j \in \text{COMPS} - \Delta \).

dKMR show that there exists a diagnosis for an observation if and only if the observation is consistent with the system description. The same result holds for excuses in \( M_D \).

**Proposition 5.2** There exists an excuse for \( \beta \) iff \( SD \land \beta \) is satisfiable.
Based on Boutilier's notion comparative normality we can derive preferred excuses relative to model $M_D$. Preferred excuses will be those maximally normal; that is, preferred excuses are determined by the most normal worlds consistent with the observation. These are the worlds where as few components as possible are considered abnormal, and render the observation possible.

**Definition 5.2** Let $\alpha$ an excuse for $\beta$. $\alpha$ is a preferred excuse for $\beta$ iff there is no excuse $\alpha'$ such that $\alpha' < \alpha$ (namely, there is no $\alpha'$ such that $M \models \bar{\Box}(\alpha' \supset \Diamond \alpha') \land \bar{\Diamond}(\alpha' \land \Box \neg \alpha)$).

dKMR define a minimal diagnosis as a diagnosis consisting of minimal set $\Delta \subseteq \text{COMPS}$ of components assumed to be working abnormally, given the system description and the observation. We can show formal correspondence between preferred excuses and minimal diagnoses.

**Theorem 5.3** $\alpha$ is a preferred excuse for $\beta$ iff $\alpha \land \gamma$ is a minimal diagnosis, where $\alpha = \bigwedge_i \text{ab}.c_i$ for $c_i \in \Delta$, and $\gamma = \bigwedge_j \neg \text{ab}.c_j$ for $c_j \in \text{COMPS} - \Delta$.

We have the following obvious corollary, that asserts that whenever there is an excuse for $\beta$, there is always some preferred excuse assuming a minimal set of abnormal components.

**Corollary 5.4** If there is some excuse for $\beta$ then, there is $\alpha'$ a preferred excuse for $\beta$.

The empty diagnosis is the only minimal diagnosis when the observation does not conflict with what the system should do if all its components were behaving correctly. That is, if there is nothing wrong there is no reason to conjecture a faulty component.

**Proposition 5.5** $\alpha = T$ is a preferred excuse for $\beta$ iff $SD \models SD \cup \neg \text{ab}(c)$ is consistent for all components $c \in \text{COMPS}$.

de Kleer Mackworth Reiter observed the following when the system description allows fault models and/or exoneration axioms. Not every superset of the faulty components involved in a minimal diagnosis need provide a diagnosis. Let's modify the previous bulb example. Let's assume two bulbs, and the following system description (which includes a description of the system behaviour when components are faulty):
Chapter 5. Capturing Consistency Based Diagnosis

\[ \neg ab\_bulb1 \land \neg ab\_bulb2 \Rightarrow \text{bright\_light} \quad ab\_bulb1 \land ab\_bulb2 \Rightarrow \text{dark} \]

Let’s add that \textit{tenuous\_light}, \textit{bright\_light} and \textit{dark} are mutually inconsistent. \textit{ab\_bulb1} \land \neg \textit{ab\_bulb2} is a minimal diagnosis for \textit{tenuous\_light} and so is \textit{(ab\_bulb2} \land \neg \textit{ab\_bulb1}); however, \textit{ab\_bulb1} \land \textit{ab\_bulb2} is \textit{not} a diagnosis for \textit{tenuous\_light}.

Consider the set of defaults \( D = \{ \neg ab\_bulb1, \neg ab\_bulb2 \} \), and the corresponding model \( M_D \). Then, \( M_D \models SD \land ab\_bulb1 \Rightarrow \neg \textit{tenuous\_light}, M_D \models SD \land ab\_bulb2 \Rightarrow \neg \textit{tenuous\_light} \); however, \( M_D \models SD \land ab\_bulb1 \land ab\_bulb2 \Rightarrow \neg \textit{tenuous\_light} \). The example demonstrates the following proposition.

**Proposition 5.6** If \( \alpha \) is an excuse for \( \beta \), \((\alpha \land \alpha')\) need not be an excuse, where \( \alpha = \bigwedge_i ab\_c_i \) for \( c_i \in \Delta \), and \( \alpha' = \bigwedge_j ab\_c_j \) for \( c_j \in \Delta' \), for any \( \Delta' \subseteq \text{COMPS} \).

We can characterize minimal diagnosis as preferred excuses in \( M_D \). While an ordinary excuse should render the observation consistent, a preferred excuse should be as normal as possible. However there could be several scenarios consistent with the observation that assume a minimal number of failing components. The next theorem uses the weak conditional connective \( \rightarrow \), that allows us to reason with incomparable scenarios. Let’s recall that the conditional \( B \rightarrow C \) is read as “at some maximal set of most normal worlds where \( B \) is true, \( C \) is also true”. The assertion \( B \rightarrow C \) is perfectly consistent with the existence of other most normal situations where \( B \) holds, but \( C \) does not. In our semantic model, we can identify preferred excuses exactly when they are entailed by some maximal scenario where the observation holds.

**Theorem 5.7** \( M_D \models SD \land \beta \rightarrow \alpha \) iff \( \alpha \land \gamma \) is a minimal diagnosis for \( \beta \).

where \( \alpha = \bigwedge_i ab\_c_i \) for \( c_i \in \Delta \) and \( \gamma = \bigwedge_j \neg ab\_c_j \) for \( c_j \in \text{COMPS} - \Delta \).

The set of all preferred excuses gets naturally identified.

**Corollary 5.8** Let \( \Psi = \{ \alpha \text{ s.t. } M_D \models SD \land \beta \rightarrow \alpha, \text{ where } \alpha = \bigwedge_i ab\_c_i \text{ for } c_i \in \Delta \} \).

\( M_D \models SD \land \beta \Rightarrow \forall \Psi \) iff for each \( \alpha \in \Psi \), \( \alpha \land \gamma \) is a minimal diagnosis for \( \beta \), s.t. \( \gamma = \bigwedge_j \neg ab\_c_j \) for \( c_j \in \text{COMPS} - \Delta \).
Chapter 5. Capturing Consistency Based Diagnosis

Our characterization of preferred excuses seems intimately related with dKMR’s characterization of minimal diagnosis in terms of minimal conflicts. Given SDUOBS, they define a conflict as a disjunction of ab-literals entailed by SDUOBS. When SD and OBS are propositional, a conflict is an ab-clause which is an implicate of SDU OBS. A minimal conflict is any ab-clause which is a prime implicate of SDU OBS. A minimal conflict is positive if its ab-literals are all positive. Their principal theorem characterizing minimal diagnosis states that \( \alpha \land \gamma \) is a minimal diagnosis iff \( \alpha \) is a prime implicant of the set of positive minimal conflicts, where \( \alpha = \Lambda_i ab.c_i \) for \( c_i \in \Delta \) and \( \gamma = \Lambda_j \neg ab.c_j \) for \( c_j \in COMPS - \Delta \). Since preferred excuses have absolute correspondence with minimal diagnosis, then, we dKMR’s characterization of all minimal diagnosis should hold for all preferred excuses. Let \( \Pi \) be the set of all positive conflicts of SDUOBS. For each element (clause) \( \pi \in \Pi \) we have \( (SDUOBS \models \pi) \). The following relation seems to hold for each conflict \( \pi \) and any excuse \( \alpha \):

\[
M_D \models \Box \Box (\pi \land \Box (\pi \supset \alpha)).
\]

It seems obvious that the R-minimal worlds in \( M_D \) where all \( \pi \in \Pi \) hold characterizes the set of preferred excuses. In other words, the R-minimal worlds where the conjunction of all minimal conflicts hold \( (\land \Pi) \) characterizes precisely the disjunction of all minimal diagnoses \( (\lor \Psi, \text{for } \Psi \text{ the set of all minimal diagnoses}) \).

5.2 On the Relation between Consistency Based and “Abductive” Diagnosis

We have mapped the consistency based diagnosis framework on a Theorist model \( M_D \), for a set of defaults \( D = \{\neg ab.c_1, \ldots, \neg ab.c_n \} \). By assuming \( F = SD \), and a set of conjectures \( C = \{ab.c_1, \ldots, ab.c_n\} \), we are able to draw legitimate comparisons between the diagnoses arising from the two frameworks. We can show that if \( \alpha \) is a Theorist explanation for \( \beta \) in \( M_D \), then \( \alpha \) is an excuse for \( \beta \).

**Proposition 5.9** If \( M_D \models SD \land \alpha \rightarrow \beta \) then \( M_D \models SD \land \alpha \not\models \neg \beta \).
While the result is trivial, the proposition states the following. Given the same system description—descriptive enough to draw a Theorist is explanation for $\beta$—then, the same Theorist explanation is an excuse (i.e. a diagnosis in consistency based). We conclude that the “abductive” explanation is logically stronger than the consistency based explanation. However, as David Poole (1990),(1989b) spelled out, the explanations/diagnoses derivable from each framework depend on the knowledge provided in the system description.

In order to derive diagnoses from the consistency based framework the correct behaviour of the system (making explicit the $\neg ab$-assumptions) has to be defined. In order to derive diagnosis in Theorist (at least some) fault models should be encoded in the system description. It is clear that if no fault models are included in the system description SD, that is, SD just axiomatizes the correct behavior of the system, assuming some abnormalities $ab.C$ excuses an observation $\beta$ but will not entail that observation. Namely $\{\Lambda_{c\in\Delta} ab.C \} \cup SD \cup \beta$ is consistent, but $\{\Lambda_{c\in\Delta} ab.C \} \cup SD \not\models \beta$.

We conclude that when enough fault models are included as part of the system description, diagnoses arising from the consistency based framework become more “predictive”. Namely, when fault models are encoded, consistency based diagnoses do not just excuse the observation, but are inconsistent with the negation of the observation.

Konolige (1992b) extended the consistency based diagnosis framework allowing for diagnoses that are based not only on “abnormalities”. For instance, suppose we consider that “If the switch is off we expect no light”; however, we don’t regard the switch being off (or on) an abnormality. Konolige determines a set of possible primitive “causes” consisting of ab-literals and other literals as well. For the two-bulbs example, let’s add to the system description

$$\text{switch} \cdot \text{off} \supset \text{dark}$$

Let’s consider a set of possible primitive “causes” $\{ab.bulb1, ab.bulb2, switch.off\}$. Now $switch.off$ explains dark and involves no abnormalities. Konolige draws preferences among explanations, based on the notion of “maximizing normality”. Our account of
preferences among excuses seems to be related to Konolige’s concept of the \textit{adjunct} of an explanation.\footnote{An explanation \( A \) is associated with the set the normality conditions that \( A \) violates. Such normality conditions are collected in what is called the adjunct of \( A \). Then, explanations violating more normality conditions are not preferred (namely, explanations with minimal –subset related– adjuncts are preferred). Konolige’s concept of the adjunct of an explanation is not entirely semantic. For instance, assume the system description (considered undefeasible knowledge) consisting of \( ab\text{-plug} \supset ab\text{-bulb}, ab\text{-plug} \supset dark \) and \( ab\text{-bulb} \supset dark \). The explanation \( ab\text{-plug} \) is preferred over the explanation \( ab\text{-plug} \land ab\text{-bulb} \), and over \( ab\text{-bulb} \) arguing that the sentence \( ab\text{-plug} \) “excuses” \( ab\text{-bulb} \).}

He also extends dKMR’s framework allowing for normality conditions other than the expectations of components being normal. These normality conditions are modeled as defaults, and are explicitly stated in the system description. Our mapping of the consistency based diagnosis framework naturally accounts for these extensions, becoming more “Theorist like”.
Chapter 6

Conclusions

6.1 Summary

In this thesis we presented a general model of explanation based on the revision of the epistemic state of a program or agent. A natural ordering on explanations was provided based exclusively on semantic criteria. The result was an object-level account of abduction that respects the inherent defeasibility of explanations.

We proposed a notion of explanation relative to the background epistemic state of an agent or program. We first considered the background epistemic state as an agent’s beliefs about the actual state of the world. Explanations were modeled as subjunctive conditionals relative to the agent’s epistemic state. We expressed our desiderata on explanations as hypothetical changes in belief and translated them as subjunctive conditionals via the Ramsey test. We identified a family of explanations which we called “epistemic explanations”, modeled by different subjunctive conditional connectives (and respectively, different kinds of changes in belief). Varying in their predictive explanatory power, we distinguished “predictive”, “weak”, and “might” explanations. Predictive explanations were conceived as the strongest in the family and require that the explanation be sufficient to induce full belief in the explanandum. This predictive but “defasible” notion of explanation generalizes most current accounts, which require some deductive relationship between explanation and observation. “Might” explanations modeled as the weakest explanation, just guaranteeing that by believing in the explanation the explanandum not be discarded (namely, the explanandum be regarded possible).

Factual and counterfactual explananda were shown to be distinguishable and explainable in our framework. In this sense our model of abduction broadens the abilities of
most abductive frameworks. The majority of existing accounts of explanation do not admit explanations of counterfactual observations; namely, they must refrain to account for explananda that conflicts with the current facts.

As Rescher (1978) said “conjectural fancy is limitless”; then, lack preferences in any account of explanations may result chaotic. From the many possible competing explanations, some may be quite implausible. We proposed a very natural preference ordering on explanations. Founded on the idea that an explanation should not only make the observation plausible but itself be maximally plausible, we derived preferences on explanations. We based our account of explanation on Boutilier’s models for belief revision. The objective belief state of the agent, together with the revision policies (and possibly other preferences too) determine a plausibility ordering over possible worlds. The preference ordering on explanations in terms of plausibility arose from the plausibility ordering over possible worlds. The same conditional information that accounted for the defeasibility of explanations induced preferences.

We identified that not every sentence satisfying the semantic criteria resembles a “simplest” or “most informative” explanation. We observed that preferences in this respect vary among different applications (e.g. natural language, image interpretation), and go beyond semantical considerations. Trivial explanations, explanations including irrelevant factors, and explanations involving some disjunction with some implausible belief fall in this category. Interestingly, these explanations may be circumstantially preferred in certain applications. Contrasting to our semantic qualifications we referred to these considerations as pragmatics of explanations, but they haven’t been explored to any extent.

We then showed how we can capture the Theorist and the consistency based diagnosis frameworks, which are today’s canonical AI approaches to abduction based on default assumptions. By recasting them in our model of abduction we illustrated the expressive power and generality our framework. The relation between Theorist explanations and epistemic explanations arose by interpreting the background epistemic state of the agent
or program as a *theory of expectations* or a normative set of defaults. The Theorist framework was captured as a reflexive and transitive Kripke model (CT4O-structure), ranking worlds by counting default rule violations. In a similar fashion we provided a mapping for Brewka's extension of the Theorist system, that contemplates priorities among defaults. Our accomplishment in this aspect goes further than the exercise of providing a mapping of these frameworks in terms of possible worlds. We showed how the Theorist and Brewka's frameworks can be extended to provide predictive explanations and preferences among explanations.

We first identified Theorist explanations with the *paraconsistent* conditional connective in CT4O, and we drew the relation to weak epistemic explanations. We discussed the weak nature of Theorist explanations, showing them to be in some sense paraconsistent.

We proposed a stronger notion of explanation in Theorist (and Brewka) setting: a predictive Theorist/Brewka explanation. This predictive notion was based on the idea that an explanation together with *every* maximal set of expectations (defaults) should account for the observation. We elaborated on the effect of priorities when a default framework is used abductively, noticing that predictive explanations become more meaningful in the prioritized setting, and also that the number of predictive explanations may increase.

Our extension of Theorist to account for preferences was based on a simple and principled conception. Explanations satisfying more (under set inclusion) expectations (defaults) were preferred. We showed that this notion of preference precisely matches our notion of preference in terms of plausibility on epistemic explanations. The correspondence was immediate given the relation between models for belief revision and models for default reasoning ((Boutilier 1992a), (1992c), (1992f)). The theory of expectations (defaults) induces a fixed ordering over possible worlds, interpreted as a *normality ordering*. Worlds satisfying more (under set inclusion) expectations (defaults) are more normal in the ordering. A comparative preference ordering on explanations naturally arose based exclusively on the conditionals (defaults) sanctioned by the expectation theory.
Chapter 6. Conclusions

We then mapped the consistency based diagnosis framework onto a Theorist model by assuming a set of defaults consisting in each component not being abnormal. We showed consistency based diagnoses to be related to "might" epistemic explanations when interpreting the background epistemic state as the expectation that each component be working correctly. By recasting consistency based diagnosis and Theorist on the same kind of model, we were able to highlight connections between the two.

We proposed a very clear and natural semantics for our model of abduction, and expressed our desiderata of explanations as purely logic constraints. The development of this thesis was done entirely using existing modal conditional logics for belief revision and default reasoning. By using Boutilier's logics CO and CT4O, we inherited a principled semantics and a logical calculus for our semantic considerations about explanations.

6.2 Future Research

We have explored very little of what we called the pragmatics of explanation. More work needs to be done in terms of studying the possibility of filtering explanations to avoid triviality, irrelevant (but consistent) information, background information and the problem of disjunctions with implausible beliefs. Also, worthy of further investigation is the problem of preferences among explanations in the case of the explanandum being accepted or indeterminate in the epistemic state where the inquiry arises. We have shown that all explanations that are epistemically possible with respect to a background epistemic state are ranked as maximally plausible. Hence the plausibility criterion is not adequate for drawing a preference ordering between epistemically possible explanations. Other criteria might be of use. For instance, Levesque (1989) proposes a syntactic criteria for simplicity.

In our development of epistemic explanations we have committed to providing an explanation that is accepted (in the epistemic state) whenever the explanandum is accepted. Then, if $\beta$ was already accepted in the agent's epistemic state, whatever reason for $\beta$ being accepted should be an explanation for $\beta$. In many cases that explanation can
be a disjunction of all possible implicants of $\beta$, or ultimately the trivial explanation. As Boutilier pointed out, when modeling a notion of scientific explanation, an observation $\beta$ may have been accepted with no explanation for it. Then, when seeking an explanation for some already believed explanandum $\beta$, some not yet accepted belief $\alpha$ can be a candidate scientific explanation. A notion of scientific explanation may be modeled by relaxing our commitment for explanation be believed when explanandum is.

Our account of explanation has limited itself to explaining objective beliefs\(^1\) with other objective beliefs. That is, explanation and observation have to be propositional sentences. It would be interesting to allow for an explanandum belonging to the bimodal language ($\beta$ in $L_B$). For example, in the question, “Why if postmen were on strike would my letter not arrive?”, the object of inquiry is itself a conditional sentence. Along the same line of generalization, the study of explanations that are conditionals (or other statements in $L_B$) may prove interesting. For example, assume the following beliefs: “Fred hates the rain” and “If it rains, Fred is unhappy”. Suppose “It rains”. If we question “How would Fred be happy?”, one answer can be “were not raining”. But another explanatory answer can be “if it were not the case that if it rains Fred is unhappy”. In this case, the denial of a conditional, may provide an explanation. An account of conditional explanations allowing $\alpha, \beta \in L_B$, would require the machinery of iterated conditionals (nested revisions), formulating a very general account of explanation.

Our notion of predictive explanation is based on AGM revision semantics. The AGM revision models changes in mistaken beliefs about a static world. In contrast, Katsuno and Mendelzon’s update semantics focuses on changes in beliefs about a changing or dynamic world. Understanding the dual or “abductive” side of belief updates may give rise to a different notion of explanation. Suppose we know that Fred loves either rock music or blues, but not both. We could explain Fred liking rock if we knew that he did not like blues. Two years later we met Fred at a party and we see him captivated by rock music. What do we conclude about his like for blues? Only if he didn’t change his

\(^1\)Objective beliefs are those expressible in the object language.
musical taste, we conclude that he doesn’t like blues. By revision semantics, we ought to assume a static world, and we could explain Fred’s not liking blues by his liking rock. But things could have changed. By update semantics, this is not a valid “explanation”.

A parallel between abduction as belief revision and planning as belief update seems worthy of study. Revising a belief set with an explanation should make the observation (explanandum) believed. By updating the state of the world by an action, a goal can be reached. Namely, the interpretation of sequences actions as sequences of updates can lead to a principled semantic approach to planning in AI.

On another frontier, following the same line of this work, it seems possible to attack the counterpart problem of causation, maybe following Lewis’s (1973) counterfactual analysis of causation. Causal relations may be understood as a special kind of a “normative conditional” as opposed to a subjunctive conditional. The following example explains the distinction. Assume the causal relation that “low pressure causes low barometric readings”, and suppose that “today there is low pressure but the barometer reads high”. Evidently the barometer is malfunctioning. As sanctioned by controllability concerns, “low pressure” causes “low barometer reading”. In contrast, solely “low pressure” does not explain (the counterfactual circumstance) “the barometric reading low”. This is because today the pressure is low, but the barometer reads high. In terms of epistemic explanations, “low pressure and barometer not malfunctioning” explains (in a hypothetical fashion) “low barometer reading”.

Finally, a comparison of our approach with quantitative frameworks of causation/explanation may prove to be enlightening.
References


REFERENCES


REFERENCES


Appendix A

Proofs of Theorems Chapter 3

Proposition 3.3 Let $\alpha$ be a predictive explanation for $\beta$ such that $\alpha$ is indeterminate in $K$. Then, Predict and Absent are equivalent.

Proof $\alpha, \neg \alpha \not\in K$. By $\text{EpStatus}$, $\beta, \neg \beta \not\in K$. Then, $M \models \alpha \Rightarrow \beta$ iff

1. $M \models B(\alpha \supset \beta)$
2. $M \models B(\neg \beta \supset \neg \alpha)$
3. $M \models \neg \beta \Rightarrow \neg \alpha$.

Proposition 3.4 Let $\alpha$ be a predictive explanation for $\beta$, such that $\neg \beta \in K$. Then, $\neg \beta \in (K^-\beta)^*_{\alpha}$ iff $M \models \beta \Rightarrow \alpha$.

Proof $M \models B(\neg \alpha, B \neg \beta, \alpha \Rightarrow \beta)$. $\|K^-\beta\| = \|K\| \cup \text{Pl}(\beta)$. Then, $\alpha \not\in K^-\beta$, therefore, $(K^-\beta)^*_{\alpha}$ is a consistent revision. Then,

1. $M \models \beta \Rightarrow \alpha$ iff
2. for every $w \in \text{Pl}(\beta)$, $w \models \alpha$ iff
3. $(\beta \supset \alpha) \in K^-\beta$ iff
4. $(\neg \alpha \supset \neg \beta) \in K^-\beta$ iff
5. $\neg \beta \in (K^-\beta)^*_{\alpha}$.

Proposition 3.5 Let $\alpha$ be a predictive explanation for $\beta$, such that $\beta \in K$. Then, $\alpha \in (K^+\alpha)^*_{\beta}$ iff $M \models \neg \alpha \Rightarrow \neg \beta$.

Proof $M \models B\alpha, B\beta, \neg \beta \Rightarrow \neg \alpha$. $\|K^+\alpha\| = \|K\| \cup \text{Pl}(\neg \alpha)$. Then, $\neg \beta \not\in K^+\alpha$, therefore, $(K^+\alpha)^*_{\beta}$ is a consistent revision. Then,

1. $M \models \neg \alpha \Rightarrow \neg \beta$ iff
2. for every $w \in \text{Pl}(\neg \alpha)$, $w \models \neg \beta$ iff
3. $(\neg \alpha \supset \neg \beta) \in K^+\alpha$ iff
4. $(\beta \supset \alpha) \in K^+\alpha$ iff
5. $\alpha \in (K^+\alpha)^*_{\beta}$.

Proposition 3.6 Let $\alpha$ be a predictive explanation for $\beta$, such that $\beta$ is indeterminate in $K$. Then, Correl and Cover are equivalent.
Proof \(\beta, \neg \beta \notin K\). By EpStatus \(\alpha, \neg \alpha \notin K\). Then,
\[
\begin{align*}
M \models \beta & \Rightarrow \alpha \text{ iff } M \models B(\beta \supset \alpha) \text{ iff } \\
M \models B(\neg \alpha \supset \neg \beta) & \text{ iff } \\
M \models \neg \alpha & \Rightarrow \neg \beta.
\end{align*}
\]

Proposition 3.7 \(\beta\) is always a preferred (quasi)-predictive explanation for \(\beta\).

Proof According to Definition 3.7 \(\beta\) is a preferred explanation for \(\beta\) iff there is no explanation \(\alpha\) such that \(\alpha < \beta\).
Suppose there is an (quasi)-predictive explanation \(\alpha\) such that \(\alpha < \beta\). Then \(\alpha\) satisfies \(M \models \alpha \Rightarrow \beta\), so every \(R\)-minimal \(\alpha\)-world satisfies \(\beta\); but this contradicts \(\alpha < \beta\). Then such \(\alpha\) can not exist.

Proposition 3.8 For any predictive or quasi-predictive explanation \(\alpha\) for \(\beta\), \(\beta \leq \alpha\).

Proof If \(\alpha\) is a quasi-predictive or predictive explanation, then \(M \models \alpha \Rightarrow \beta\). Then each \(R\)-minimal \(\alpha\)-world is a \(\beta\)-world. Then, \(M \models \Box(\alpha \supset \Diamond \beta)\).

Proposition 3.9 Let \(\alpha\) be a (quasi)-predictive explanation for \(\beta\). \(\alpha \leq \beta\) iff \(M \models \beta \nLeftarrow \neg \alpha\).

Proof By hypothesis, \(M \models \alpha \Rightarrow \beta\).
\[
M \models \beta \Leftarrow \neg \alpha \text{ iff } \\
\text{at each minimal } \beta\text{-cluster there is some } \alpha\text{-world iff } \\
M \models \Box(\beta \supset \Diamond \alpha).
\]

Proposition 3.10 \(\alpha\) is a preferred quasi-predictive explanation for \(\beta\) iff \(\alpha\) is a preferred predictive explanation for \(\beta\).

Proof (left to the right). Let \(\alpha\) be a preferred quasi-predictive explanation for \(\beta\); then \(\alpha, \beta\) satisfy Conditions Commit, Predict and Absent.
By Predict \(\alpha \Rightarrow \beta\), then \(\Box(\alpha \supset \Diamond \beta)\), hence \(\alpha \leq \beta\).
By Proposition 3.8 \(\beta\) is a preferred quasi-predictive explanation for \(\beta\), then \(\alpha \not< \beta\); so \(\alpha \leq \beta\) and \(\beta \leq \alpha\).
Then, if \(\neg B\neg \beta\) then there is at least a \(\beta\)-world in \(\|K\|\),
then since \(\alpha \leq \beta\), \(\neg B\neg \beta \supset \neg B\neg \alpha\).
By contraposition, \(B\neg \alpha \supset B\neg \beta\).
And, by Absent, \(B\neg \beta \supset B\neg \alpha\)
by Commit, \(B\beta \supset B\alpha\), and
by Predict, \(B\alpha \supset B\beta\).
Then \((B\alpha \equiv B\beta) \wedge (B\neg\alpha \equiv B\neg\beta)\), thus \(\alpha\) satisfies \textit{Epstatus}, hence \(\alpha\) is a preferred predictive explanation for \(\beta\).

The proof right to left is trivial.

**Proposition 3.11** Let \(\neg\beta \notin K\), (i.e., \(\beta\) accepted or indeterminate), and \(\alpha, \alpha'\) any predictive explanations for \(\beta\). Then, \(\alpha \leq \alpha'\) and \(\alpha' \leq \alpha\).

**Proof** By hypothesis \(\neg\beta \notin K\). By \textit{EpStatus} \(\neg\alpha, \neg\alpha' \notin K\). Then there is some \(v, w \in \|K\|\) such that \(w \models \alpha\), and some \(v \models \alpha'\). \(v, w\) are \(R\)-minimal worlds in \(M\). Then, \(wRv\) and \(wRv\), so \(\alpha \leq \alpha'\) and \(\alpha' \leq \alpha\).

**Proposition 3.12** Let \(\neg\beta \notin K\). If \(\alpha, \alpha'\) are predictive explanations for \(\beta\), then \(\alpha, \alpha'\) are equally plausible in \(M_{\beta}\).

**Proof (sketch)** We present the case for explanandum indeterminate in \(K\). Exactly the same reasoning applies if \(\beta\) is accepted in \(K\).

Let \(\beta\) be indeterminate in \(K\). Then, there are some \(\beta\)-worlds in the minimal cluster in \(M\). As \(\alpha, \alpha'\) are predictive explanations for \(\beta\), then \(\alpha, \alpha'\) are also indeterminate in \(K\). Then there are some \(\alpha\)-worlds and some \(\alpha'\)-worlds in the minimal cluster in \(M\).

Let \(M_{\beta}\) the model after the revision of \(K\) by \(\neg\beta\). Then, the minimal cluster in \(M_{\beta}\) is formed by all \(\neg\beta\)-worlds satisfying \(\neg\alpha, \neg\alpha'\), since by hypothesis \(M \models \neg\beta \Rightarrow \neg\alpha\), \(M \models \neg\beta \Rightarrow \neg\alpha'\).

According to the natural revision function, the next to the minimal cluster is formed precisely by all the \(\beta\)-worlds from \(K\). Among those \(\beta\)-worlds are the \(\alpha\)-worlds and the \(\alpha'\)-worlds. Then \(M_{\beta} \models \Box(\alpha' \cup \Diamond \alpha)\), and \(M_{\beta} \models \Box(\alpha \cup \Diamond \alpha')\). Thus \(\alpha, \alpha'\) are equally preferred.

**Proposition 3.13** A trivial explanation for \(\beta\) is a predictive covering explanation for \(\beta\).

**Proof** The proof is trivial. \(\beta\) has the same epistemic status as \(\beta\). Besides, \textit{Predict} and \textit{Absent}, \textit{Correl} and \textit{Cover}, are trivially satisfied since \(M \models \beta \Rightarrow \beta\) and \(M \models \neg\beta \Rightarrow \neg\beta\) (ID is a theorem in CT4O, and CO).
Appendix B

Proofs of Theorems Chapter 4

Proposition 4.1 $M_D$ is a $CT_4O^*$ model.

Proof We have to show that $R$ is reflexive and transitive; this is obvious from definition 4.2.

$wRw$ for all $w \in W$ since $V(w) \subseteq V(w)$; then $R$ is reflexive.

Assume $wRv$ and $vRu$. Then, $V(w) \subseteq V(v)$, and $V(v) \subseteq V(u)$. Therefore $V(w) \subseteq V(u)$; so $wRu$. Hence $R$ is transitive.

Proposition 4.2 Let $M_D = (W, R, \varphi)$ be the Theorist model for $D$. Let each single default formula $d \in D$ be satisfiable. $M_D$ possesses a single $R$-minimum cluster iff $D$ is consistent.

Proof Suppose $w, v$ are minimal in $R$. Since $D$ is consistent both $V(w) = \emptyset$ and $V(v) = \emptyset$. So $wRv$ and $vRw$.

there is a unique minimal cluster.

Lemma 4.3 $S$ is an extension of $(X, D)$ iff there is a minimal $X$-cluster that characterizes $S$.

Proof (left to right) Assume $S$ is an extension of $(X, D)$, and let $\|S\| \subseteq W$ the proposition denoted by $S$ in $M_D$.

$S = Cn(X \cup D)$, s.t. $D$ is a maximal subset of $D$ consistent with $X$, then

each $w \in \|S\|$, $w \models X$ and $w \models D$, so

each $w \in \|S\|$ violates the same defaults $d' \in (D - D)$, thus

for any $v, w \in \|S\|$, $V(w) = V(v)$, hence $wRv$ and $vRw$.

Since $D$ is a maximal subset of $D$ consistent with $X$, then there can be no $u \in W$, $u \models X, u \models D'$ such that $D \subset D' \subseteq D$. Then, $\|S\|$ is an $R$-minimal $X$-cluster.

(right to left) Let $\|S\| \subseteq W$ a minimal $X$-cluster in $M_D$.

Since $\|S\|$ is a cluster, for every $u, w \in \|S\|$ $wRu$ and $uRw$, thus $V(w) = V(u)$, then

$u, w$ violate the same defaults; hence, $u, w$ satisfy the same defaults from $D$.

Since $\|S\|$ is an $X$-cluster minimal in $R$, then for each $w \in \|S\|

103
\[ w \models X, \text{ and for every other } v \in W \text{ such that } wRv \text{ but not } vRw, v \not\models X, \text{ then} \]
\[ \text{if } V(w) \subset V(v) \text{ then } v \text{ does not satisfy } X, \text{ then} \]
\[ w \models D \text{ such that } D \text{ is a maximal subset of } D \text{ consistent with } X. \text{ Hence,} \]
\[ \|S\| \text{ characterizes } S = \text{Cn}(X \cup D) \text{ s.t. } D \text{ is a maximal subset of } D \text{ consistent with} \]
\[ X, \text{ thus, } S \text{ is an extension of } (X, D). \]

**Theorem 4.4** \( \gamma \) is predicted in Theorist sense from \( \langle F, D \rangle \) iff \( M_D \models F \Rightarrow \gamma \), and \( F \) is satisfiable.

**Proof** \( \gamma \) is predicted in Theorist sense iff
\( \gamma \) belongs to all extensions of \( \langle F, D \rangle \) and \( F \) consistent iff
by Lemma 4.3, \( \gamma \) is satisfied by each \( R \)-minimal \( F \)-cluster in \( M_D \) and \( F \) satisfiable iff
by Proposition 2.10, \( M_{Th} \models F \Rightarrow \gamma \) and \( F \) satisfiable.

**Theorem 4.5** Let \( D \subseteq D \) and \( C \subseteq C \). \( D \cup C \) is a Theorist explanation for \( \beta \) iff \( M_D \models (F \cup C) \rightarrow \beta \) and \( (F \cup C) \) is consistent.

**Proof** \( D \cup C \) is a Theorist explanation for \( \beta \) iff
\( F \cup D \cup C \models_{CPL} \beta \) and \( F \cup D \cup C \) is consistent iff
by Poole’s theorem 2.1, \( \beta \) is in some extension of \( (\langle F \cup C \rangle, D) \) iff
by Lemma 4.3, there is an \( R \)-minimal \( (F \cup C) \)-cluster in \( M_D \) satisfying \( \beta \), iff
by Proposition 2.11 \( M_D \models (F \cup C) \rightarrow \beta \), and \( (F \cup C) \) is consistent.

**Corollary 4.6** If \( F \models_{CPL} \neg \beta \) then there is no \( C \) such that \( M_D \models (F \cup C) \rightarrow \beta \) and \( (F \cup C) \) is consistent.

**Proof** We show the contrapositive of the assertion. Namely, if \( \beta \) is explainable then \( F \not\models \neg \beta \).
Assume \( \beta \) is explainable. By theorem 4.5, there exists \( C \) such that \( M_D \models (F \cup C) \rightarrow \beta \) and \( (F \cup C) \) is consistent.
Then, there exists \( w \in W \) such that \( M_D \models_w (F \cup C \cup \beta) \); so, there exists a world \( w \) satisfying \( F \) and \( \beta \); then, \( F \not\models \neg \beta \).

**Proposition 4.7** If \( M_D \models F \Rightarrow \beta \) and \( F \) is satisfiable then (a) \( \beta \) is explainable with no conjectures, and (b) \( \neg \beta \) is not explainable with no conjectures.

**Proof** Assume \( M_D \models F \Rightarrow \beta \), and \( F \) satisfiable.
(a) we have to show that $M_D \models F \rightarrow \beta$, but this is obvious since $A \Rightarrow B \supset A \rightarrow B$ is a theorem in CT4O.

(b) Since $M_D \models F \Rightarrow \beta$, by Proposition 2.10 every minimal $F$-cluster satisfies $\beta$.

Then, there is no minimal $F$-cluster that satisfies $\neg \beta$, while $F$ is satisfiable. By Proposition 2.11, $M_D \models F \not\rightarrow \neg \beta$.

By theorem 4.5, $\neg \beta$ is not explainable from $\langle F, D, \emptyset \rangle$.

**Theorem 4.8** $C$ is a predictive explanation for $\beta$ iff $M_D \models (F \cup C) \Rightarrow \beta$, and $(F \cup C)$ is consistent.

**Proof** $C$ is a predictive explanation for $\beta$ iff $\beta$ belongs to every extension of $\langle (F \cup C), D \rangle$ iff by Lemma 4.3, $\beta$ is satisfied by every $R$-minimal $(F \cup C)$-cluster in $M_D$ iff by Proposition 2.10, $M_D \models (F \cup C) \Rightarrow \beta$.

**Proposition 4.9** Let $C \subseteq C$. If $C$ is a predictive explanation for $\beta$ then (a) there is a Theorist explanation for $\beta$ assuming $C$; and (b) $\neg \beta$ is not explainable assuming $C$.

**Proof** By hypothesis $C$ is a predictive explanation for $\beta$.

Then, $M_D \models (F \cup C) \Rightarrow \beta$, and $(F \cup C)$ is consistent.

(a) Since $A \Rightarrow B \supset A \rightarrow B$ is a theorem in CT4O, then $(F \cup C) \Rightarrow \beta \supset (F \cup C) \rightarrow \beta$.

By theorem 4.5 there is $D \subseteq D$, such that $(C \cup D)$ is an explanation for $\beta$.

(b) By Proposition 2.10 all $R$-minimal $(F \cup C)$-clusters in $M_D$ are $\beta$-clusters.

Then no minimal $(F \cup C)$-cluster verifies $\neg \beta$.

by Proposition 2.11 $M_D \models (F \cup C) \not\rightarrow \neg \beta$,

by theorem 4.5 there is no $D \in D$ such that $C \cup D$ is a Theorist explanation for $\neg \beta$.

**Proposition 4.10** $M_B$ is a CT4O$^*$ model.

**Proof** We have to show that $R$ is reflexive and transitive.

For every $w \in W$ $wRw$ iff $\forall i, 1 \leq i \leq n \ V'_w = V'_w$. By definitions 4.5 and 4.7, $R$ is reflexive.

Assume $wRv$, and $vRu$.

There are three possible cases:
Appendix B. Proofs of Theorems Chapter 4

(a) \( \min(u, v) = \min(v, w) \)

\[ V^\min(u,v) \subseteq V^\min(v,u), \text{ and } V^\min(v,u) \subseteq V^\min(u,v), \]

then \( \min(u, w) = \min(u, v) = \min(v, w) \), and

\[ V^\min(u,w) \subseteq V^\min(u,v) \text{ and } V^k = V^k \text{, } \forall k, 1 \leq k < \min(u, v); \text{ hence, } wRu. \]

(b) \( \min(u, v) > \min(v, w) \)

\[ V^\min(u,v) \subseteq V^\min(v,u), \text{ and } V^k = V^k, \text{ } \forall k, 1 \leq k < \min(u, v); \]

in particular, \( V^\min(v,w) = V^\min(v,w) \)

\[ V^\min(v,w) \subseteq V^\min(v,u), \text{ then } V^\min(v,w) \subseteq V^\min(v,w) \]

then \( \min(u, w) = \min(v, w) \)

and \( \forall k, 1 \leq k < \min(v, w), V^k = V^k = V^k; \text{ hence, } wRu. \)

(c) \( \min(u, v) < \min(v, w) \)

\[ V^\min(v,w) \subseteq V^\min(v,u), \text{ and } V^k = V^k, \text{ } \forall k, 1 \leq k < \min(v, w); \]

in particular, \( V^\min(u,v) = V^\min(v,u) \)

\[ V^\min(u,v) \subseteq V^\min(u,v), \text{ then } V^\min(u,v) \subseteq V^\min(u,v) \]

then \( \min(u, w) = \min(u, v) \)

and \( \forall k, 1 \leq k < \min(u, v), V^k = V^k = V^k; \text{ hence, } wRu. \)

Lemma 4.11 \( S \) is an extension of \( (X, B) \) iff there is a minimal \( X \)-cluster characterizing \( S \).

Proof The proof is exactly alike the one of Lemma 4.3.

Proposition 4.12 Let each single default formula \( d \in B \) be satisfiable. \( M_B \) possesses a single \( R \)-minimum cluster iff \( B \) is consistent.

Proof The proof is exactly alike the one of Proposition 4.2.

Theorem 4.13 \( \mathcal{F} \models_{BS} \gamma \) iff \( M_B \models F \Rightarrow \gamma \), and \( F \) satisfiable.

Proof \( \mathcal{F} \models_{BS} \gamma \) iff

every extension of \( \langle \mathcal{F}, B \rangle \) contains \( \gamma \) and \( \mathcal{F} \) is consistent iff

by Lemma 4.11, every \( R \)-minimal \( F \)-cluster satisfies \( \gamma \) and \( F \) is satisfiable iff

by Proposition 2.10, \( M_B \models F \Rightarrow \gamma \), and \( F \) satisfiable.

Theorem 4.14 \( \mathcal{F} \cup \alpha \models_{BW} \beta \) iff \( M_B \models (F \wedge \alpha) \rightarrow \beta \) and \( (F \wedge \alpha) \) satisfiable.
Proof $\mathcal{F} \cup \alpha \vdash_{BW} \beta$ iff
some extension of $((\mathcal{F} \cup \alpha), \mathcal{B})$ contains $\beta$ and $\mathcal{F} \cup \alpha$ is consistent iff
by Lemma 4.11, some $R$-minimal $F \wedge \alpha$-cluster satisfies $\beta$ iff
by Proposition 2.11, $M_{\mathcal{B}} \models (F \wedge \alpha) \rightarrow \beta$, and $(F \wedge \alpha)$ satisfiable.

Proposition 4.15 If $wR_{\mathcal{D}}v$ then $wR_{\mathcal{B}}v$.
Proof Since $\mathcal{D} = \bigcup_{i=1}^{n} B_i$, then for every $u \in W$ $V(u) = \bigcup_{i=1}^{n} V_u^i$
Assume $wR_{\mathcal{D}}v$. By definition 4.2 $V(v) \subseteq V(w)$, then
$V_u^i \subseteq V_w^i$, $\forall i, 1 \leq i \leq n$, then $wR_{\mathcal{B}}v$.

Proposition 4.16 $M_{\mathcal{D}}$ and $M_{\mathcal{B}}$ have the same clusters.
Proof We have to show that for any $v, w \in W$ $wR_{\mathcal{D}}v$ and $vR_{\mathcal{D}}w$ iff $wR_{\mathcal{B}}v$ and $vR_{\mathcal{B}}w$.
By Proposition 4.15, if $wR_{\mathcal{D}}v$ and $vR_{\mathcal{D}}w$ then $wR_{\mathcal{B}}v$ and $vR_{\mathcal{B}}w$.
By definition 4.7 if $wR_{\mathcal{B}}v$ and $vR_{\mathcal{B}}w$ then $V_u^i = V_w^i$, $\forall i, 1 \leq i \leq n$
Since $\mathcal{D} = \bigcup_{i=1}^{n} B_i$, then for every $u \in W$ $V(u) = \bigcup_{i=1}^{n} V_u^i$; then,
$V(v) = V(w)$; thus $wR_{\mathcal{D}}v$ and $vR_{\mathcal{D}}w$.

Proposition 4.17 Every $R_{\mathcal{B}}$-minimal cluster in $M_{\mathcal{B}}$ is a $R_{\mathcal{D}}$-minimal cluster in $M_{\mathcal{D}}$.
Proof By Proposition 4.16 $M_{\mathcal{D}}$ and $M_{\mathcal{B}}$ possess the same clusters.
Since $\mathcal{D} = \bigcup_{i=1}^{n} B_i$, then for every $u \in W$ $V(u) = \bigcup_{i=1}^{n} V_u^i$; then,
Let $U \subseteq W$ a cluster in $M_{\mathcal{D}}$ that is not $R_{\mathcal{D}}$-minimal.
Then there is some $R_{\mathcal{D}}$-minimal world $w \in W$, $w \notin U$, such that for each $u \in U$,
$uR_{\mathcal{D}}w$ and not $wR_{\mathcal{D}}u$ iff
for each $u \in U$, $V(w) \subseteq V(u)$, then
$V_u^i \subseteq V_w^i$, $\forall i, 1 \leq i \leq n$
by definition 4.7, for each $u \in U$, $uR_{\mathcal{B}}w$ but not $wR_{\mathcal{B}}u$, then
$U$ is not an $R_{\mathcal{B}}$-minimal cluster in $M_{\mathcal{B}}$.

Proposition 4.18 If $\mathcal{D}$ is consistent, then $M_{\mathcal{D}}$ and $M_{\mathcal{B}}$ have the same unique minimal cluster.
Proof Assume $\mathcal{D}$ is consistent.
By Propositions 4.2, and 4.12, $M_{\mathcal{D}}$ and $M_{\mathcal{B}}$ have a single minimal cluster.
Then, for every $u$ such that $V(u) = \emptyset$ $u$ is in the $R_{\mathcal{D}}$-minimal cluster.
Since $\mathcal{D} = \bigcup_{i=1}^{n} B_i$, then for every $u \in W$ $V(u) = \bigcup_{i=1}^{n} V_u^i$.
Hence, $M_{\mathcal{D}}$ and $M_{\mathcal{B}}$ have the same unique minimal cluster.
Proposition 4.19 If $M_D \models \alpha \Rightarrow \beta$ then $M_B \models \alpha \Rightarrow \beta$.

Proof If $M_D \models \alpha \Rightarrow \beta$ then every $R_D$-minimal $\alpha$-cluster satisfies $\beta$;

By Proposition 4.17 every $R_B$-minimal $\alpha$-cluster in $M_B$ is a $R_D$-minimal $\alpha$-cluster in $M_D$.

Since every every $R_D$-minimal $\alpha$-cluster in $M_D$ satisfies $\beta$, then each $R_B$-minimal $\alpha$-cluster satisfies $\beta$, then,

by Proposition 2.10 $M_B \models \alpha \Rightarrow \beta$.

Proposition 4.20 If $M_B \models \alpha \rightarrow \beta$ then $M_D \models \alpha \rightarrow \beta$.

Proof Assume $M_B \models \alpha \rightarrow \beta$

by Proposition 2.11, there is some $R_B$-minimal $\alpha$-cluster in $M_B$ satisfying $\beta$;

by Proposition 4.17, every $R_B$-minimal cluster in $M_B$ is an $R_D$-minimal cluster in $M_D$,

so there is some $R_D$-minimal $\alpha$-cluster in $M_D$ satisfying $\beta$,

then by Proposition 2.11 $M_D \models \alpha \rightarrow \beta$.

Theorem 4.21 Let $\alpha, \alpha'$ be predictive explanations for $\beta$. $\alpha \leq_F \alpha'$ iff $M \models 
\Box((\alpha' \land F) \supset \Diamond(\alpha \land F))$.

Proof By hypothesis $(F \land \alpha) \Rightarrow \beta$, $(F \land \alpha') \Rightarrow \beta$, and $(F \land \alpha)$, $(F \land \alpha')$ are satisfiable.

$\alpha \leq_F \alpha'$ iff

each maximal set of defaults consistent with $F \land \alpha'$ is contained in some maximal set of defaults consistent with $\alpha$ iff

each $R$-minimal $(F \land \alpha')$-world $w$ has access to some $R$-minimal $(F \land \alpha)$-world $v$,

namely $wRv$ iff

$M \models \Box((\alpha' \land F) \supset \Diamond(\alpha \land F))$.

Corollary 4.22 The preference relation $\leq_F$ is reflexive and transitive.

Proof This is evident since Boutiler's plausibility ordering is reflexive and transitive.

Proposition 4.23 Let $\alpha$ a predictive explanation for $\beta$, then $\beta \leq_F \alpha$.

Proof By hypothesis $(F \land \alpha) \Rightarrow \beta$, then

by Proposition 2.10, every $R$-minimal $(F \land \alpha)$-cluster, satisfies $\beta$, then

each $(F \land \alpha)$-world has access to some $R$-minimal $(F \land \alpha \land \beta)$-world, then

$M \models \Box((F \land \alpha) \supset \Diamond(F \land \beta))$. 
Appendix C

Proofs of Theorems Chapter 5

Theorem 5.1 $\alpha$ is an excuse for $\beta$ iff $\alpha \land \gamma$ is a diagnosis for $\beta$, where $\alpha = \bigwedge_i ab.c_i$ for $c_i \in \Delta$, and $\gamma = \bigwedge_j \neg ab.c_j$ for $c_j \in COMPS - \Delta$.

Proof Let $\alpha = \bigwedge_i ab.c_i$ for $c_i \in \Delta \subseteq COMPS$.

$\alpha$ is an excuse for $\beta$ iff

$M_D \models SD \land \alpha \not\models \neg \beta$ iff

there is some $w \in W$ such that $w$ is a $R$-minimal $SD \land \alpha$-world, and $w$ satisfies $\beta$ iff

there is some $w \in W$ such that $w \models SD \land \alpha \land \beta$ and $w$ violates minimal set of defaults consistent with $SD \land \alpha$, namely $w \models \gamma$, $\gamma = \bigwedge_j \neg ab.c_j$ for $c_j \in COMPS - \Delta$ iff

$SD \land \alpha \land \gamma \land \neg \beta$ is satisfiable iff

$\alpha$ is a diagnosis for $\beta$.

Proposition 5.2 There exists an excuse for $\beta$ iff $SD \land \beta$ is satisfiable.

Proof There is no $\alpha$ such that $SD \land \alpha \not\models \neg \beta$ iff

there is no $w \in W$ such that $w$ is a $R$-minimal $SD \land \alpha$-world satisfying $\beta$ iff

for every $w \in W$ such that $w \models SD$, then $w \models SD \land \alpha$, for different $\alpha$; hence, there is no $w \in W$, satisfying $SD \land \beta$, iff $SD \land \beta$ is not satisfiable.

Theorem 5.3 $\alpha$ is a preferred excuse for $\beta$ iff $\alpha \land \gamma$ is a minimal diagnosis, where $\alpha = \bigwedge_i ab.c_i$ for $c_i \in \Delta$, and $\gamma = \bigwedge_j \neg ab.c_j$ for $c_j \in COMPS - \Delta$.

Proof According to the definition 4.2, each cluster in $M_D$ satisfies satisfies the same defaults. As defaults have the form $\neg ab.c_j$ and excuses have the form $\bigwedge ab.c_i$; then, each cluster uniquely determines an excuse “complementary” to the conjunction of all the normality assumptions in that cluster.

$\alpha$ is a preferred excuse for $\beta$, $\alpha = \bigwedge_i ab.c_i$ for $c_i \in \Delta$ iff

$\alpha$ is an excuse for $\beta$ and there is no excuse $\alpha'$, such that $(SD \land \alpha') < (SD \land \alpha)$ in $M_D$ iff
Appendix C. Proofs of Theorems Chapter 5

\[ \alpha \text{ is an excuse for } \beta \text{ and there is no excuse } \alpha' \text{ satisfying strictly more defaults than } \alpha \text{ iff } \]
\[ \alpha \text{ is an excuse for } \beta \text{ and there is no excuse } \alpha' = \bigwedge_i ab.c_i \text{ such that for } c_i \in \Delta' \subset \Delta \text{ iff } \]
\[ \alpha \land \gamma \text{ is a diagnosis for } \beta \text{ (theorem 5.1), and for no proper subset } \Delta' \subset \Delta \text{ is } \alpha' \land \gamma' \text{ a diagnosis (where } \alpha' = \bigwedge_i ab.c_i \text{ for } c_i \in \Delta' \text{ and } \gamma' = \bigwedge_j \neg ab.c_j \text{ for } c_j \in \text{COMPS} - \Delta') \text{ iff } \]
\[ \alpha \land \gamma \text{ is a minimal diagnosis.} \]

**Proposition 5.5** \( \alpha = T \text{ is a preferred excuse for } \beta \text{ iff } M_D \models SD \neq \neg \beta \text{ and } SD \cup \{ \neg ab(c) \} \text{ is consistent for all components } c \in \text{COMPS}. \)

**Proof** \( M_D \models SD \neq \neg \beta \text{ and } M_D \models T \neq SD \text{ iff } \)
\[ \text{there is some minimal } SD\text{-world } w \text{ in } M_D \text{ such that } V(w) = \{ \} \text{ (w violates no defaults } D) \text{ and } w \models \beta \text{ iff } \]
\[ \alpha = T \text{ is an excuse for } \beta \text{ and there is no other } \alpha' \text{ involving strictly less abnormalities iff } \]
\[ \alpha = T \text{ is a minimal diagnosis for } \beta. \]

**Theorem 5.7** \( M_D \models SD \land \beta \rightarrow \alpha \text{ iff } \alpha \land \gamma \text{ is a minimal diagnosis for } \beta. \)
(\( \text{where } \alpha = \bigwedge_i ab.c_i \text{ for } c_i \in \Delta \text{ and } \gamma = \bigwedge_j \neg ab.c_j \text{ for } c_j \in \text{COMPS} - \Delta \).)

**Proof** Since \( M_D \) divides worlds into clusters that satisfy the same defaults, then each cluster characterizes the same excuse.
Then \( \beta \)-clusters are always subclusters of \( \alpha \)-clusters.
\( M_D \models SD \land \beta \rightarrow \alpha \text{ iff } \)
\[ \text{there is an } R\text{-minimal } SD\land \beta \text{-cluster } C \text{ satisfying } \alpha \text{ where } \alpha = \bigwedge_i ab.c_i \text{ for } c_i \in \Delta. \]
Then cluster \( C \) satisfies \( \gamma = \bigwedge_j \neg ab.c_j \text{ for } c_j \in \text{COMPS} - \Delta \text{ iff } \)
\[ SD \land \beta \land \alpha \land \gamma \text{ is satisfiable and for any } w \in W \text{ and for any } u \in C, \]
such that \( uRw \text{ and not } wRu, w \models \neg(SD \land \beta). \)
Then \( w \) is strictly more normal that \( u \), that is, \( w \models \bigwedge \alpha' \text{ and } w \models \bigwedge_{c \in \Delta'} \text{COMPS} \neg ab.c \)
where \( \alpha' = \bigwedge_i ab.c_i \text{ for } c_i \in \Delta' \subset \Delta. \)
then, \( \alpha \land \gamma \text{ is a diagnosis for } \beta \) but there is no \( \alpha' \land \gamma' \text{ that is diagnosis for } \beta \), s.t. \( \alpha' = \bigwedge_i ab.c_i \text{ for } c_i \in \Delta' \subset \Delta \text{ and } \gamma' = \bigwedge_j \text{COMPS} - \Delta' \subset \Delta \text{ iff } \)
\[ \alpha \land \gamma \text{ is a minimal diagnosis for } \beta. \]
Corollary 5.8 Let $\Psi = \{ \alpha \text{ s.t. } M_D \models SD \wedge \beta \rightarrow \alpha, \text{ where } \alpha = \bigwedge_i ab.c_i \text{ for } c_i \in \Delta \}$. 

$M_D \models SD \wedge \beta \rightarrow \bigvee \Psi$ iff for each $\alpha \in \Psi$, $\alpha \wedge \gamma$ is a minimal diagnosis for $\beta$, s.t. $\gamma = \bigwedge_j \neg ab.c_j$ for $c_j \in COMPS - \Delta$.

Proof Different clusters satisfy different defaults, and a set of defaults determines a set of excuses, then each minimal $SD \wedge \beta$-cluster satisfies a different $\alpha$.

Let $\Psi = \{ \alpha \text{ s.t. } M_D \models SD \wedge \beta \rightarrow \alpha, \text{ where } \alpha = \bigwedge_i ab.c_i \text{ for } c_i \in \Delta \}$. 

$M_D \models SD \wedge \beta \rightarrow \bigvee \alpha$ iff each $R$-minimal $SD \wedge \beta$-world satisfies $\bigvee \Psi$ iff by theorem 5.7, each $R$-minimal $SD \wedge \beta$-cluster determines a preferred excuse $\alpha$ iff by theorem 5.3 each $\alpha \wedge \gamma$ is a minimal diagnosis for $\beta$, where $\alpha \in \Psi$, $\alpha = \bigwedge_i ab.c_i$ for $c_i \in \Delta$ and $\gamma = \bigwedge_j \neg ab.c_j$ for $c_j \in COMPS - \Delta$.

Proposition 5.9 If $M_D \models SD \wedge \alpha \rightarrow \beta$ then $M_D \models SD \wedge \alpha \not\models \neg \beta$.

Proof $A \rightarrow B \supset A \not\models \neg B$ is a theorem in CT4O.