To appear, *Artificial Intelligence,* 1995

Abduction as Belief Revision*

Craig Boutilier and Verónica Becher
Department of Computer Science
University of British Columbia
Vancouver, British Columbia
CANADA, V6T 1Z4
email: cebly@cs.ubc.ca, becher@cs.ubc.ca

Abstract

We propose a model of abduction based on the revision of the epistemic state of an agent. Explanations must be sufficient to induce belief in the sentence to be explained (for instance, some observation), or ensure its consistency with other beliefs, in a manner that adequately accounts for factual and hypothetical sentences. Our model will generate explanations that *nonmonotonically predict* an observation, thus generalizing most current accounts, which require some deductive relationship between explanation and observation. It also provides a natural preference ordering on explanations, defined in terms of normality or plausibility. To illustrate the generality of our approach, we reconstruct two of the key paradigms for model-based diagnosis, abductive and consistency-based diagnosis, within our framework. This reconstruction provides an alternative semantics for both and extends these systems to accommodate our predictive explanations and semantic preferences on explanations. It also illustrates how more general information can be incorporated in a principled manner.

*Some parts of this paper appeared in preliminary form as "Abduction as Belief Revision: A Model of Preferred Explanations," *Proc. of Eleventh National Conf. on Artificial Intelligence (AAAI-93)*, Washington, DC, pp.642–648 (1993).

1 Introduction

It has become widely recognized that a lot of reasoning does not proceed in a "straightforward" deductive manner. Reasonable conclusions cannot always be reached simply by considering the logical consequences (relative to some background theory) of some known facts. A common pattern of inference that fails to conform to this picture is *abduction*, the notion of finding an *explanation* for the truth of some fact. For instance, if the grass is wet, one might explain this fact by postulating that the sprinkler was turned on. This is certainly not a *deductive consequence* of the grass being wet (it may well have rained).

Abduction has come to play a crucial role in knowledge representation and reasoning, across many areas of AI. In discourse interpretation, one often wants to ascribe beliefs to a speaker that explain a particular utterance, perhaps gaining insight into the speaker's intentions [30]. More generally, plan recognition often proceeds abductively. In high-level scene interpretation [51], an interpretation can be reached by postulating scene objects that explain the appearance of objects in an image. Probably the most common use of abductive inference in AI is in the area of model-based diagnosis. Given unexpected observations of the behavior of an artifact or system, a diagnosis is usually taken to be some set of components, the malfunctioning of which explains these observations [14, 24, 17, 49, 43].

Traditionally, the process of abduction has been modeled by appeal to some sort of deductive relation between the *explanandum* (or fact to be explained) and the *explanation* (the fact that renders the explanandum plausible). Hempel's [29] *deductive-nomological* explanations fall into this category, requiring that the explanation entail the explanandum relative to some background knowledge. Broadly speaking, this picture of abduction can be characterized as follows: an explanation for β relative to background theory T will be any α that, together with T, entails β (usually with the additional constraint that $\{\alpha\} \cup T$ be consistent). Such a picture is adopted in much research on abduction [54, 35, 50]. Theories of this type are, unfortunately, bound to the unrelenting nature of deductive inference. There are three directions in which such theories must be generalized.

First, we should not require that an explanation deductively entail its observation (even relative to some background theory). There are very few explanations that do not admit exceptions. The sprinkler being on can explain the wet grass; but the sprinkler being on with a water main broken is not a reasonable explanation. Yet this exceptional condition does not make the initial explanation any less compelling. Rather it illustrates that explanations may entail their conclusions in a *defeasible* or *nonmonotonic* sense.

Second, while there may be many competing explanations for a particular observation, certain of these may be relatively implausible. While a tanker truck exploding in front of the yard may explain

the wet grass in the sense described above, this is certainly not as reasonable an explanation as the sprinkler being turned on. Thus, we require some notion of preference to chose among these potential explanations.

Third, the deductive picture of explanation does not allow one to explain facts that are inconsistent with the background theory. Such explanations are, in fact, among the most important; for it is facts that conflict with existing expectations that most urgently require explanation. This is the case in diagnostic applications, for example, where observations to be explained contradict our belief that a system is performing according to specification.

The first two of these problems can be addressed using, for example, probabilistic information [29, 17, 46, 41]. We might simply require that an explanation render the observation sufficiently probable. Explanations might thus be *nonmonotonic* in the sense that α may explain β, but $\alpha \wedge \gamma$ may not (e.g., $P(\beta|\alpha)$ may be sufficiently high while $P(\beta|\alpha \wedge \gamma)$ may not). For instance, it is highly likely that the grass becomes wet when the sprinkler is turned on, but it is unlikely to become wet if the water main is broken. Preference can also be given to explanations that are more likely. A tanker truck exploding in front of the yard is much less probable than the sprinkler being turned on. There have been proposals to address these issues in a more qualitative manner using "logic-based" frameworks also. Peirce (see Rescher [52]) discusses the "plausibility" of explanations, as do Quine and Ullian [48]. Consistency-based diagnosis [49, 16] uses abnormality assumptions to capture the context-dependence of explanations; and preferred explanations are those that minimize abnormalities. Poole's [44] assumption-based framework captures some of these ideas by explicitly introducing a set of default assumptions to account for the nonmonotonicity of explanations.

In this paper we propose a semantic framework and logical specification of abduction that captures the spirit of probabilistic proposals, but does so in a qualitative fashion. Explanations are given a defeasible aspect through the use of techniques for default reasoning and belief revision. Furthermore, explanations are viewed as more or less plausible according to a qualitative notion of plausibility, a relation naturally induced by the preferences associated with our defaults. Finally, by relying on existing theories of belief revision, explanations for facts that conflict with existing beliefs can be provided. In particular, such conflicting observations will require explanations that themselves force an agent to revise its beliefs.

Our account will take as central subjunctive conditionals of the form $A \Rightarrow B$, which can be interpreted as asserting that, if an agent were to believe A it would also believe B. Such a conditional can be consistently held even if A is believed to be false. This is the cornerstone of our notion of explanation: if believing A is sufficient to induce belief in B, then A *explains* B. This determines a strong, *predictive* sense of explanation; but weaker forms of explanation can also be captured.

Semantically, such conditionals are interpreted relative to an ordering of plausibility or normality over possible worlds. This ordering is taken to represent the epistemic state of an agent; thus all forms of explanation we describe can be classified as *epistemic explanations*. Our conditional logic, described in earlier work as a representation of belief revision and default reasoning [3, 7, 9], has the desired nonmonotonicity and induces a natural preference ordering on sentences (hence explanations).

In the next section we describe abduction, belief revision, our conditional logics and other necessary logical preliminaries. In Section 3, we discuss the concept of explanation, its epistemic nature, and how different types of explanations can be captured in our framework. We also introduce the notion of *preferred explanations*, showing how the same conditional information used to represent the defeasibility of explanations induces a natural preference ordering. To demonstrate the expressive power of our model, in Section 4 we show how Poole's [43, 44] Theorist framework (without constraints) and Brewka's [12] extension of Theorist can be captured in our logics. This reconstruction explains semantically the non-predictive and *paraconsistent* nature of explanations in Theorist. It also illustrates the correct manner in which to augment Theorist with a notion of predictive explanation and how one should capture semantic preferences on Theorist explanations. These two abilities have until now been unexplored in this canonical abductive framework. In Section 5, we reconstruct a canonical theory of *consistency-based diagnosis* due to de Kleer, Mackworth and Reiter [16, 49] in our logics. This again suggests extensions of the theory and illustrates the natural similarities and distinctions between consistency-based and abductive diagnosis.

Proofs of main theorems may be found in the appendix.

2 Abduction and Belief Revision

In this section, we briefly discuss some previous work on abduction, drawing attention to the aspects of these various proposals that influence our approach. We also describe the *AGM model* of belief revision of Alchourrón, Gärdenfors and Makinson [2]; and we present the conditional logics required to capture this theory of revision, due to Boutilier [9]. This will provide the logical apparatus required to describe the process of abduction in terms of belief revision.

2.1 Abduction

Abduction is the process of inferring certain facts and/or laws that render some sentence plausible, that explain some phenomenon or observation. The sentence to be explained is often denoted the *explanandum*. We will use the term "observation" instead, for typically we are interested in explaining some observed fact. This is merely suggestive, however, for hypothetical possibilities can be explained

as well. The sentences (facts or laws) doing the explaining are often dubbed the *explanans* sentences. Though the term is often used to characterize this inference process, we will use "explanation" more simply to refer to the explanans sentences. Thus, an explanation renders an observation plausible (in some yet to be determined sense).

The most basic and, in some idealized sense, the most compelling form of abduction is represented by Hempel's [29] *deductive-nomological* explanations. Such explanations consist of certain specific facts and universal generalizations (scientific laws) that, taken together, deductively entail a given observation. For example, the observation "This thing flies" can be explained by the fact "This thing is a bird" and the law "All birds fly." As Hempel observes, often parts of the explanation are left unstated with the explicitly provided explanation being elliptical. If it is understood among participants in some discourse that all birds fly, then "This thing is a bird" alone is a reasonable explanation. Suppose we take T to be some theory capturing the relevant background knowledge (this may be some scientific or commonsense theory). Then the sentence α explains observation β just when

$$\{\alpha\} \cup T \models \beta$$

We will be less concerned with the nomological aspects of abduction, assuming that relevant laws are captured in some background theory.[1] Thus, our notion of explanation will be elliptical in this sense, taking background information for granted.

The criteria for deductive explanations are clearly too strong to allow wide applicability. In commonsense reasoning and scientific inquiry very few explanations have such strength. One accepts as a reasonable explanation for wet grass that the sprinkler was turned on; but this explanation is not (deductively) conclusive. The grass may have been covered by a tarpaulin, the water pressure may have fallen at a crucial instance, any of a number of other exceptional conditions can defeat this inference. Of course, we may claim that "the sprinkler was turned on" is elliptical, implicitly assuming that none of these exceptional circumstances hold, and that the *true* explanation includes the denial of these. However, this runs into the *qualification problem* of default reasoning, the problem of having to know that such conditions are false [38]. This view is also untenable when such qualifications cannot be listed, or the phenomenon in question is inherently probabilistic (at least, given our current knowledge). To take an example of Hempel, Jim's close exposure to his brother who has the measles explains Jim catching the measles; but it certainly doesn't imply Jim catching the measles.

A number of methods for specifying probabilistic explanations have been proffered. Hempel [29]

[1] In fact, as we will see in Section 3, the "theory" is implicit in the epistemic state of our reasoning agent. We will have a few things to say about laws in our framework in the concluding section.

requires that the explanation make the observation highly probable. Thus, probabilistic explanations still retain the essential predictive power of deductive explanations. Other accounts make less stringent requirements. For instance, Gärdenfors [22] insists only that the explanation render the observation more probable than it is *a priori*. A key component of the Gärdenfors theory is that the judgements of probability are rendered with respect to the epistemic state of an agent. We return to this in Section 3.

Because of their probabilistic nature, such explanations are *nonmonotonic* or *defeasible*. It may be that `SprinklerOn` explains `WetGrass`, since this observation is very probable given the explanation. But the stronger proposition $\texttt{SprinklerOn} \wedge \texttt{Covered}$ is not a reasonable explanation, for the probability of wet grass is quite low in this case. Our goal is to capture this type of explanation in a qualitative fashion. Rather than relying on probabilistic information, we will provide an account of defeasible explanations based on the "default rules" held by an agent.

Both deductive and probabilistic models of abduction typically give rise to a number of competing explanations for a given observation. The propositions `Rain` and `SprinklerOn` both explain `WetGrass`. If an agent has to choose among competing explanations, there must exist some criteria for this choice. An obvious preference criterion on explanations is based on the likelihood of the explanations themselves. An agent should choose the most probable explanation relative to a given context. Such accounts are often found in diagnosis [46, 15] and most probable explanations are discussed by Pearl [41]. In a more qualitative sense, one might require that adopted explanation(s) be among the most "plausible." This view is advocated by Peirce (see Rescher [52]) and Quine and Ullian [48]. The notion of *minimal diagnosis* in the consistency-based models of diagnosis [49] is an attempt to qualitatively characterize most probable diagnoses. We will provide a formal framework in which such qualitative judgements of plausibility can be made.

One of the areas of AI that most frequently appeals to abductive inference is *model-based diagnosis*. Given a theory describing the correct behavior of some system or artifact, one can make predictions about its behavior based on some given information. One might expect a certain observation based on information about other parts of the system. For example, given the inputs to a digital circuit, the background theory (or *system description*) allows one to deduce the value of the outputs. Should the actual observation differ from the expected observation then the system must not conform to the system description (assuming the input values are correct). The goal of model-based diagnosis is to discover an explanation for the aberrant behavior, usually some set of components of the system that, if behaving abnormally, will entail or excuse the actual observation. The two main paradigms for model-based diagnosis are the *abductive* approaches, of which Poole's [43, 44] Theorist framework is representative, and *consistency-based* models such as that of de Kleer, Mackworth and Reiter [16, 49]. These will be discussed in detail in Sections 4 and 5.

2.2 Conditionals and Belief Revision

The account of abduction we propose relies heavily on the notion of belief revision. For instance, a *predictive explanation* requires that belief in the explanation be sufficient to induce belief in the observation. Therefore we must be able to test the epistemic state of an agent after it (hypothetically) adopts a potential explanation, or test a knowledge base once it is revised to incorporate the explanation. A theory of belief revision thus lies at the core of epistemic explanation.

We assume an agent to have a deductively closed set of beliefs K taken from some underlying language. For concreteness, we will assume this language $\mathbf{L}_{CPL}$ to be that of classical propositional logic generated by some set of variables **P**. We will often take K to be the closure of some finite set of premises, or *knowledge base, KB*; so $K = Cn(KB)$. The *expansion* of K by new information A is the belief set $K_A^+ = Cn(K \cup \{A\})$. This is a seemingly reasonable method of belief change when $K \not\models \neg A$. More troublesome is the revision of K by A when $K \models \neg A$. Some beliefs in K must be given up before A can be accommodated. The problem lies in determining which part of K to give up. Alchourrón, Gärdenfors and Makinson [2] have proposed a theory of revision (the AGM theory) based on the following observation: the least "entrenched" beliefs in K should be given up and A added to this *contracted* belief set.

We use K_A^* to denote the belief set resulting when K is revised by A. The AGM theory logically delimits the scope of acceptable revision functions. To this end, the AGM postulates below are maintained to hold for any reasonable notion of revision [22].

(R1) K_A^* is a belief set (i.e. deductively closed).

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

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

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

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

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

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

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

The semantics of AGM revision functions will be described below.

An alternative model of revision is based on the notion of *epistemic entrenchment* [22]. Given a belief set K, we can characterize the revision of K by ordering beliefs according to our willingness

to give them up when necessary. If one of two beliefs must be retracted in order to accommodate some new fact, the least entrenched belief will be relinquished, while the most entrenched persists. Gärdenfors [22] presents five postulates for such an ordering and shows that these orderings determine exactly the space of revision functions satisfying the AGM postulates. We let $B \leq_E A$ denote the fact that A is at least as entrenched as B in theory K. A complete set of sentences of this form is sufficient to specify a revision function. We note that the dual of an entrenchment ordering is a plausibility ordering on sentences. A sentence A is more plausible than B just when $\neg A$ is less entrenched than $\neg B$, and means that A would be more readily accepted than B if the opportunity arose. Grove [28] studied this relationship and its connection to the AGM theory.

Another form of belief change studied within the AGM theory is the process of *contraction*, or rejecting a belief in a belief set. When the belief set K is contracted by A, the resulting belief set K^-_A is such that A is no longer held. The AGM theory provides a set of postulates for contraction as well. This process is related to revision via the Levi and Harper identities:

$$K^-_A = K \cap K^*_{\neg A} \quad \text{and} \quad K^*_A = (K^-_{\neg A})^+_A$$

2.2.1 The Logics CO and CO*

Boutilier [9] presents a family of bimodal logics suitable for representing and reasoning about the revision of a knowledge base. We briefly review the logics and associated possible worlds semantics for revision. We refer to [9] for further details and motivation.

Semantically, the process of revision can be captured by considering a *plausibility ordering* over possible worlds. We can reason about such structures, as well as AGM revision (and several generalizations of it), using a family of bimodal logics. The language $\mathbf{L}_B$ is a bimodal language formed from a denumerable set $\mathbf{P}$ of propositional variables, together with the usual classical connectives and two modal operators $\Box$ and $\overleftarrow{\Box}$. Intuitively, $\Box A$ is read as "A holds at all equally or more plausible worlds," while $\overleftarrow{\Box} A$ is read "A holds at all less plausible worlds." We denote by $\mathbf{L}_{CPL}$ the propositional sublanguage of $\mathbf{L}_B$. We will define four bimodal logics based on this language.

Our semantics is based on structures consisting of a set of possible worlds W and a binary ordering relation $\leq$ over W, reflecting the relative degree of plausibility of worlds. The interpretation of $\leq$ is as follows: $v \leq w$ iff v is *at least as plausible* as w.[2] As usual, v is *more plausible* than w ($v < w$) iff $v \leq w$ but not $w \leq v$. Plausibility is a pragmatic measure that reflects the degree to which one is willing to accept w as a possible state of affairs. If v is more plausible than w, loosely speaking, v is

[2]Having "more" plausible elements denoted as "lesser" in the ordering is consistent with the usual AI practice of preferring minimal elements in some ordering — in this case, the more plausible worlds.

"more consistent" with an agent's beliefs than w. We take reflexivity and transitivity to be minimal requirements on $\leq$, dubbing any such model a CT4O-model.

Definition 2.1 [7] A *CT4O-model* is a triple $M = \langle W, \leq, \varphi \rangle$, where W is a set (of possible worlds), $\leq$ is a reflexive, transitive binary relation on W (the ordering relation), and φ maps $\mathbf{P}$ into 2^W ($\varphi(A)$ is the set of worlds where A is true).

Sentences in $\mathbf{L}_B$ are interpreted in the usual way, with the truth of a modal formula at world w in M (where $M \models_w A$ means A is true at w) given by

1. $M \models_w \Box A$ iff for each v such that $v \leq w$, $M \models_v A$.
2. $M \models_w \overleftarrow{\Box} A$ iff for each v such that $v \not\leq w$, $M \models_v A$.

If $M \models_w A$ we say that M *satisfies* A at w. For any sentence A, we use $\|A\|$ to denote the set of worlds $w \in W$ that satisfy A (assuming some fixed M). Each world in this set is an A*-world*. For an arbitrary set of formulae S, we use $\|S\|$ to denote those worlds satisfying each $A \in S$ and refer to these as S-worlds. Somewhat loosely we dub those worlds that falsify some $A \in S$ to be $\neg S$-worlds. We now define several new connectives as follows:

$$\Diamond A \equiv_{\text{df}} \neg\Box\neg A \; ; \;\; \overleftarrow{\Diamond} A \equiv_{\text{df}} \neg\overleftarrow{\Box}\neg A \; ; \;\; \overleftrightarrow{\Box} A \equiv_{\text{df}} \Box A \wedge \overleftarrow{\Box} A \; ; \;\; \overleftrightarrow{\Diamond} A \equiv_{\text{df}} \Diamond A \vee \overleftarrow{\Diamond} A$$

It is easy to verify that these connectives have the following truth conditions:

(a) $M \models_w \Diamond A$ iff for some v such that $v \leq w$, $M \models_v A$.

(b) $M \models_w \overleftarrow{\Diamond} A$ iff for some v such that $v \not\leq w$, $M \models_v A$.

(c) $M \models_w \overleftrightarrow{\Box} A$ iff for all $v \in W$, $M \models_v A$.

(d) $M \models_w \overleftrightarrow{\Diamond} A$ iff for some $v \in W$, $M \models_v A$.

These connectives have the obvious readings: $\Box A$ means "A is true at all equally or more plausible worlds"; $\Diamond A$ means "A is true at some equally or more plausible world"; $\overleftarrow{\Box} A$ means "A is true at all less plausible (and incomparable) worlds"; $\overleftarrow{\Diamond} A$ means "A is true at some less plausible (or incomparable) world"; $\overleftrightarrow{\Box} A$ means "A is true at all worlds, whether more or less plausible"; finally, $\overleftrightarrow{\Diamond} A$ means "A is true at some world, whether more or less plausible." Validity and satisfiability are defined in a straightforward manner and a sound and complete axiomatization for the logic CT4O is provided in [7].

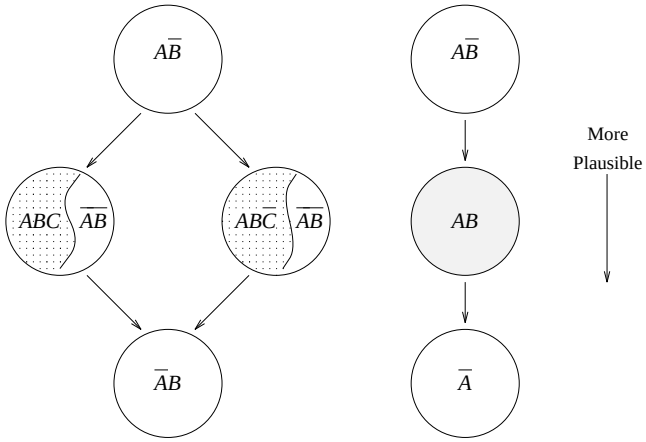

Figure 1: CT4O and CO models

A natural restriction on the ordering of plausibility is *connectedness*; that is, for any pair of worlds w, v, either $v \leq w$ or $w \leq v$. In other words, all worlds must have comparable degrees of plausibility. This restriction gives rise to the logic CO (again axiomatized in [7]).

Definition 2.2 [7] A *CO-model* is a triple $M = \langle W, \leq, \varphi \rangle$, where M is a CT4O-model and $\leq$ is totally connected.

In any reflexive, transitive Kripke frame, a *cluster* is any maximal mutually accessible set of worlds [53]: a set $\mathcal{C} \subseteq W$ is a cluster just when $v \leq w$ for all $v, w \in \mathcal{C}$ and no extension $\mathcal{C}' \supset \mathcal{C}$ has this property. We note that CO-structures consist of a totally-ordered set of clusters of equally plausible worlds, while CT4O-models consist of a partially-ordered set of clusters. Figure 1 illustrates this, where each large circle denotes a cluster of equally plausible worlds and arrows point in the direction of increasing plausibility.

Finally, both CT4O and CO can be extended by restricting attention to those structures in which all logically possible worlds are represented. No matter how implausible, each should be somehow ranked and should occur in our models. This property turns out to be crucial in characterizing the AGM theory of belief revision.

Definition 2.3 [7] Let $M = \langle W, \leq, \varphi \rangle$ be a Kripke model. For all $w \in W$, w^* is defined as the map from **P** into $\{0, 1\}$ such that $w^*(A) = 1$ iff $w \in \varphi(A)$ (w^* is the valuation associated with w).

CT4O*-models and CO*-models are (respectively) CT4O-models and CO-models satisfying the

condition that

$$\{f : f \text{ maps } \mathbf{P} \text{ into } \{0,1\}\} \subseteq \{w^* : w \in W\}.$$

This restriction is captured axiomatically determining the logics CT4O* and CO* [7].

2.2.2 Modeling Belief Revision

Assume we have a fixed (CO- or CT4O-) model M. We use *min*(α) to denote the set of *most plausible* α-worlds in M:[3]

$$\textit{min}(\alpha) = \{w : w \models \alpha, \text{ and } v < w \text{ implies } v \not\models \alpha\}$$

In both models in Figure 1, the shaded regions denote the worlds that make up *min*(A).

The revision of a belief set K can be represented using CT4O- or CO-models that reflect the degree of plausibility accorded to worlds by an agent in such a belief state. To capture revision of K, we insist that any such K*-revision model* be such that $\|K\| = \textit{min}(\top)$; that is, the model must have a (unique) minimal cluster formed by $\|K\|$.[4] This reflects the intuition that all and only K-worlds are most plausible for an agent with belief set K [9], and corresponds to a form of *only knowing* [36, 4]. The CT4O-model in Figure 1(a) is a K-revision model for $K = \textit{Cn}(\neg A, B)$, while the CO-model in Figure 1(b) is suitable for $K = \textit{Cn}(\neg A)$.

To revise K by A, we construct the revised set K^*_A by considering the set *min*(A) of most plausible A-worlds in M. In particular, we require that $\|K^*_A\| = \textit{min}(A)$; thus $B \in K^*_A$ iff B is true at each of the most plausible A-worlds. We can define a conditional connective $\Rightarrow$ such that $A \Rightarrow B$ is true in just such a case:

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

This is equivalent to the requirement that

$$\textit{min}(A) \subseteq \|B\|$$

Both models in Figure 1 satisfy $A \Rightarrow B$, since B holds at each world in *min*(A), the shaded regions of the models.

The *Ramsey test* [57] provides acceptance conditions for subjunctive conditionals of the form "If A were the case, then B would hold" by appeal to belief revision. Indeed, the conditional should be accepted just when an agent, hypothetically revising its beliefs by A, accepts B. Thus, we can equate

[3] We assume, for simplicity, that such a (limiting) set exists for each $\alpha \in \mathbf{L}_{CPL}$, though the following technical developments do not require this [7, 9].

[4] This constraint can be expressed in the object language $\mathbf{L}_B$; see [9, 4].

the conditional $A \Rightarrow B$ with the statement $B \in K^*_A$ and interpret our conditional as a certain type of epistemic subjunctive conditional. For a specific K-revision model we can define the revised belief set K^*_A as

$$K^*_A = \{B \in \mathbf{L}_{CPL} : M \models A \Rightarrow B\}.$$

Boutilier [9] shows that the revision functions determined by CO*-models are exactly those that satisfy the AGM postulates. The revision functions captured by the weaker logics impose slightly weaker constraints on the revision functions: CT4O and CT4O* fail to satisfy postulate (R8), while CT4O and CO satisfy slightly weaker versions of most of the postulates. Intuitively, a K-revision model captures the epistemic state of an agent, both its beliefs and its revision policies. A belief connective can be defined in the object language:[5]

$$\mathsf{B}(A) \equiv_{\text{df}} \top \Rightarrow A$$

We briefly describe the *contraction* of K by $\neg A$ in this semantic framework. To retract belief in $\neg A$, we simply accept the worlds in *min*(A) as epistemically possible without rejecting the possibility of K-worlds. In other words,

$$K^-_A = \|K\| \cup \mathit{min}(A)$$

This is due to the fact that certain A-worlds must become epistemically possible if $\neg A$ is *not* to be believed, and the principle of minimal change suggests that only the most plausible A-worlds should be accorded this status. The belief set $K^-_{\neg A}$ does not contain $\neg A$, and this operation captures the AGM model of contraction if we restrict our attention to CO*-models. In Figure 1(a) $K^-_{\neg A} = \mathit{Cn}(B)$, while in Figure 1(b) $K^-_{\neg A} = \mathit{Cn}(A \supset B)$.

A key distinction between CT4O and CO-models is illustrated in Figure 1: in a CO-model, all worlds in *min*(A) must be equally plausible, while in CT4O this need not be the case. Indeed, the CT4O-model shown has two maximally plausible sets of A-worlds (the shaded regions), yet these are incomparable. We denote the set of such incomparable subsets of *min*(A) by $Pl(A)$:

$$Pl(A) = \{\mathit{min}(A) \cap \mathcal{C} : \mathcal{C} \text{ is a cluster}\}$$

Thus, we have that *min*$(A) = \cup Pl(A)$. Taking *each* such subset (each element of $Pl(A)$) to be a plausible revised state of affairs rather than their union, we can define a weaker notion of revision using the following connective. It reflects the intuition that the consequent C holds within *some*

[5] See [4] for a more comprehensive definition of belief and a proof of correspondence to the belief logic weak S5.

element of $Pl(A)$:

$$(A \rightarrow C) \equiv_{\text{df}} \overleftrightarrow{\Box}(\neg A) \vee \overleftrightarrow{\Diamond}(A \wedge \Box(A \supset C))$$

The model in Figure 1(a) shows the distinction: it satisfies neither $A \Rightarrow C$ nor $A \Rightarrow \neg C$, but both $A \rightarrow C$ and $A \rightarrow \neg C$. There is a set of comparable most plausible A-worlds that satisfies C and one that satisfies $\neg C$. Notice that this connective is *paraconsistent* in the sense that both C and $\neg C$ may be "derivable" from A, but $C \wedge \neg C$ is not. However, $\rightarrow$ and $\Rightarrow$ are equivalent in CO, since *min*(A) must lie within a single cluster. This weak connective will be primarily of interest when we examine the Theorist system in Section 4.

We define the *plausibility* of a proposition by appealing to the plausibility ordering on worlds. We judge a proposition to be just as plausible as the most plausible world at which that proposition holds. For instance, if A is consistent with a belief set K, then it will be maximally plausible — the agent considers A to be epistemically possible. We can compare the relative plausibility of two propositions semantically: A is at least as plausible as B just when, for every B-world w, there is some A-world that is at least as plausible as w. This is expressed in $\mathbf{L}_B$ as $\overleftrightarrow{\Box}(B \supset \Diamond A)$. If A is (strictly) more plausible than B, then as we move away from $\|K\|$, we will find an A-world before a B-world; thus, A is qualitatively "more likely" than B. In each model in Figure 1, $A \wedge B$ is more plausible than $A \wedge \neg B$. We note that in CO-models plausibility totally orders propositions; but in CT4O, certain propositions may be incomparable by this measure.

2.2.3 Default Rules and Expectations

The subjunctive conditionals defined above have many properties one would expect of default rules. In particular, the conditional is defeasible. For instance, one can assert that if it rains the grass will get wet ($R \Rightarrow W$), but that it won't get wet if the grass is covered ($(R \wedge C) \Rightarrow \neg W$). As subjunctive conditionals, these refer to an agent adopting belief in the antecedent and thus accepting the consequent. In this case, the most plausible R-worlds must be different from the most plausible $R \wedge C$-worlds.

These conditionals have much the same character as default rules. Recently, a number of conditional logics have been proposed for default reasoning [18, 26, 33, 34]. In particular, Boutilier [7] has proposed using the logics CT4O and CO together with the conditional $\Rightarrow$ for default reasoning. To use the logics for this purpose requires simply that we interpret the ordering relation $\leq$ as ranking worlds according to their degree of normality. On this interpretation, $A \Rightarrow B$ means that B holds at the *most normal* A-worlds; that is, "If A then normally B." These default logics are shown to be equivalent to the preferential and rational consequence operations of Lehmann [33, 34]. They are also

equivalent to the logic of arbitrarily high probabilities proposed by Adams [1] and further developed by Goldszmidt and Pearl [26], and can be given a probabilistic interpretation [7].

Boutilier [9] also shows how default reasoning based on such a conditional logic can be interpreted as a form of belief revision, hence explaining the equivalence of the conditional logic representation of both processes. Gärdenfors and Makinson's [23] notion of expectation inference adopts a similar viewpoint. Roughly, we think of default rules of the form $A \Rightarrow B$ as inducing various expectations about the normal state of affairs. In particular, for any such default an agent *expects* the sentence $A \supset B$ to be true in the most normal state of affairs. An agent without specific knowledge of a particular situation should then adopt, as a "starting point," belief in this theory of expectations. In other words, an agent's "initial" beliefs should be precisely its default expectations. When specific facts F are learned, the agent can revise this belief set according to the revision model capturing its default rules. The revised belief set will then correspond precisely to the set of default conclusions the agent would reach by performing conditional default reasoning from this set of facts using its conditional default rules (see [9] for details). For this reason, our theory of explanation can be used in one of two ways. We may think of explanations relative to the epistemic state of an agent. This is the viewpoint adopted in Section 3 where we present our theory. We may also interpret the conditionals involved in explanation as default rules. This interpretation will be implicit in Sections 4 and 5 in our reconstruction of model-based diagnosis, where plausibility orderings are in fact normality orderings.

3 Epistemic Explanations

Often scientific explanations are postulated relative to some background theory consisting of various scientific laws, principles and facts. In commonsense domains, this background theory should be thought of as the belief set of some agent. We will therefore define explanations relative to the epistemic state of some agent or program. We assume this agent to possess an objective (or propositional) belief set K. We also assume the agent to have certain judgements of plausibility and entrenchment at its disposal to guide the revision of its beliefs. These may be reflected in the conditionals held by the agent, explicit statements of plausibility, or any other sentences in the bimodal language that constrain admissible plausibility orderings. Such a theory may be complete — in the sense that it determines a unique plausibility ordering — or incomplete. For simplicity, we assume (initially) that an agent's theory is complete and that its epistemic state is captured by a single K-revision model. We discuss later how one might compactly axiomatize such a categorical theory, and how explanations are derived for incomplete theories.

Defining explanations relative to such structured epistemic states extends the usual deductive and

probabilistic approaches. There an explanation must be added to an agent's "theory" to account for an observation. This restrictive view precludes meaningful explanations of observations other than those consistent with K. In fact, it is often explanations for observations that *conflict* with our current beliefs in which we are most interested. Thus, a model of belief revision seems crucial for explanations of this sort. In order to account for such explanations, one must permit the belief set (or background theory) to be revised in some way that allows consistent explanations of such observations. Gärdenfors [22] has proposed a model of abduction that relies crucially on the epistemic state of the agent doing the explaining. Our model finds its origins in his account, but there are several crucial differences. First, Gärdenfors's model is probabilistic whereas our model is qualitative. As well, our model will provide a predictive notion of explanation (in a sense described below). In contrast, Gärdenfors makes no such requirement, counting as explanations facts that only marginally affect the probability of an observation. However, we share with Gärdenfors the idea that explanations may be evaluated with respect to states of belief other than that currently held by an agent.

Levesque's [35] account of abduction is also based on the notion of an epistemic state. Levesque allows the notion of "belief" to vary (from the standard deductively-closed notion) within his framework in order to capture different types of explanation (e.g., a syntax-motivated notion of simplest explanation). Our model is orthogonal in that the notion of "implication" between explanation and observation is weakened.

In this section, we introduce several forms of epistemic explanation and their characterization in terms of revision. There are two key dimensions along which these forms of explanation are compared, *predictive power* and the *epistemic status of the observation to be explained*.

If belief in the explanation is sufficient to induce belief in the observation, the explanation is said to be *predictive*. Deductive-nomological explanations have this form, as do probabilistic explanations based on high probability. However, weaker, *non-predictive* explanations are also of interest. These must simply render the observation reasonable, without necessarily predicting it. Consistency-based diagnosis adopts this perspective. Exposure to a virus may explain one's having a cold without having the predictive power to induce the belief that one will catch cold (prior to observing the cold). Predictive and non-predictive explanations are discussed in Sections 3.1 and 3.2, respectively. We will also distinguish two forms of non-predictive explanations: *weak* explanations and the even weaker *might* explanations.

Explanations may also be categorized according to the epistemic status of the explanandum, or "observation" to be explained. There are two types of sentences that we may wish to explain: beliefs and non-beliefs. If β is a belief held by the agent, it requires a *factual* explanation, some other belief α that might have caused the agent to accept β. This type of explanation is clearly crucial in many

reasoning applications. An intelligent program will provide conclusions of various types to a user; but a user should expect a program to be able to *explain* how it reached such a belief, or to justify its reasoning. We may ask a robot to explain its actions, or an expert system to explain its predictions. The explanation should clearly be given in terms of *other* (perhaps more fundamental) beliefs held by the program. When explaining belief in β, a program or agent that offers a disbelieved sentence α is performing in a misleading manner. A second type of explanation is *hypothetical*: even if β is not believed, we may want an explanation for it, some new belief the agent *could* adopt that would be sufficient to ensure belief in β. This counterfactual reading turns out to be quite important in AI, for instance, in diagnostic tasks (see below), planning, and so on [25]. For example, if turning on the sprinkler explains the grass being wet and an agent's goal is to wet the grass, then it may well turn on the sprinkler. We can further distinguish hypothetical explanations into those where observation β is *rejected* in K (i.e., $\neg\beta \in K$) and those where observation β is *indeterminate* in K (i.e., $\beta \notin K$ and $\neg\beta \notin K$). Regardless of the predictive power required of an explanation, factual and hypothetical explanations will require slightly different treatment.

The type of explanation one requires will usually depend on the underlying application. For instance, we will see that hypothetical explanations, whether predictive or non-predictive, play a key role in diagnosis. Whatever the chosen form of explanation, certain explanations will be deemed more plausible than others and will be preferred on those grounds. We will introduce a model of preference in Section 3.3 that can be used to further distinguish explanations in this fashion.

3.1 Predictive Explanations

In very many settings, we require that explanations be *predictive*; that is, if an agent were to adopt a belief in the explanation, it would be compelled to accept the observation. In other words, the explanation should be sufficient to induce belief in the observation. Legal explanations, discourse interpretation, goal regression in planning, and diagnosis in certain domains all make use of this type of explanation.

To determine an appropriate definition of predictive explanation, we consider the factual and hypothetical cases separately. If the observation β is believed, as argued above, we require that a suitable explanation α also be believed. For example, if asked to explain the belief `WetGrass`, an agent might choose between `Rain` and `SprinklerOn`. If it believes the sprinkler is on and that it hasn't rained, then `Rain` is not an appropriate explanation. This leads to our first condition on explanations: if observation β is accepted (i.e., $\beta \in K$) then any explanation α must also be accepted (i.e., $\alpha \in K$).

If β is not believed, it may be rejected or indeterminate. In the first instance, where β is rejected,

we insist that any explanation α also be rejected (i.e., $\neg\alpha \in K$). If this were not the case then α would be consistent with K. According to the AGM theory and our model of revision, this means that accepting α would be tantamount to adding α to K, and $\neg\beta$ would still be believed. For example, suppose an agent believes the grass is not wet and that the sprinkler may or may not be on. To explain (or ensure) wet grass, it should not accept the sprinkler being on (or turn it on), for according to its beliefs the sprinkler may well be on — yet the grass is not believed to be wet.

In the second instance, where β is indeterminate, we insist that any explanation also be indeterminate (i.e., $\alpha \notin K$ and $\neg\alpha \notin K$). If $\alpha \in K$, clearly accepting α causes no change in belief and does not render β believed. Dismissing explanations α where $\neg\alpha \in K$ requires more subtle justification. Intuitively, when β is indeterminate, it is an epistemic possibility for the agent: for all the agent knows β could be true. If this is the case, it should be explained with some sentence that is also epistemically possible. If $\neg\alpha \in K$ the agent knows α to be false, so it should not be willing to accept it as an explanation of some fact β that might be true. Since learning β conflicts with none of its beliefs, so too should a reasonable explanation be consistent with its beliefs. For example, suppose an agent is unsure whether or not the grass is wet, but believes that it hasn't rained. Upon learning the grass is wet, accepting rain as an explanation seems unreasonable.[6]

Combining these criteria for both factual and hypothetical explanations, we have the following condition relating the epistemic status of observation β and explanation α:

(ES) $\alpha \in K$ iff $\beta \in K$ and $\neg\alpha \in K$ iff $\neg\beta \in K$

Assuming an agent to possess a unique revision model M reflecting its current epistemic state, we can express this in the object language as

$$M \models (\mathsf{B}\alpha \equiv \mathsf{B}\beta) \wedge (\mathsf{B}\neg\alpha \equiv \mathsf{B}\neg\beta)$$

If the epistemic state is captured by some (possibly incomplete) theory in the language $\mathbf{L}_B$, we can test this condition using entailment in the appropriate bimodal logic.

We note here that this condition relating the epistemic status of explanation and observation is at odds with one prevailing view of abduction, which takes only non-beliefs to be valid explanations. On this view, to offer a *current* belief α as an explanation is uninformative; abduction should be an "inference process" allowing the derivation of *new* beliefs. We take a somewhat different view, assuming that observations are not (usually) accepted into a belief set until some explanation is found and accepted. In the context of its other beliefs, observation β is unexpected to a greater or

[6] Below we will briefly explanations where this condition is weakened.

lesser degree. Unexplained "belief" in β places the agent in a state of cognitive dissonance. An explanation relieves this dissonance when it is accepted [22]. After this process both explanation and observation are believed. Thus, the abductive *process* should be understood in terms of *hypothetical* explanations: when it is realized what *could* have caused belief in an (unexpected) observation, both observation and explanation are incorporated. In this sense, our use of the term observation is somewhat nontraditional — it is a fact that has yet to be accepted (in some sense) as a belief. *Factual* explanations are retrospective in the sense that they (should) describe "historically" what explanation was *actually* adopted for a certain belief. We will continue to call such beliefs "observations," using the term generally to denote a fact to be explained.

Apart from the epistemic status of observation and explanation, we must address the predictive aspect of explanations. In particular, we require that adopting belief in the explanation α be sufficient to induce belief in the observation β. The obvious criterion is the following predictive condition:

(P) $\beta \in K^*_\alpha$

which is expressed in the object language as $\alpha \Rightarrow \beta$. This captures the intuition that *If the explanation were believed, so too would be the observation* [37]. For hypothetical explanations, this seems sufficient, but for factual explanations (where $\beta \in K$), this condition is trivialized by the presence of **(ES)**. For once we insist that a valid explanation α be in K, we have $K^*_\alpha = K$; and clearly $\beta \in K^*_\alpha$ for *any* belief α. But surely arbitrary beliefs should not count as valid explanations for other beliefs. The belief that grass is green should not count as an explanation for the belief that the grass is wet.

In order to evaluate the predictive force of factual explanations, we require that the agent (hypothetically) give up its belief in β and then find some α that would (in this new belief state) restore β. In other words, we contract K by β and evaluate the conditional $\alpha \Rightarrow \beta$ with respect to this contracted belief state:

(PF) $\beta \in (K^-_\beta)^*_\alpha$

Thus, when we hypothetically suspend belief in β, if α is sufficient to restore this belief then α counts as a valid explanation. The contracted belief set K^-_β might fruitfully be thought of as the belief set held by the agent before it came to accept the observation β.

An (apparently) unfortunate consequence of this condition is the difficulty it introduces in evaluation. It seems to require that one generate a new epistemic state, reflecting the hypothetical belief set K^-_β, against which to evaluate the conditional $\alpha \Rightarrow \beta$. Thus, **(PF)** requires two successive changes in belief state, a contraction followed by a revision.[7] However, it turns out that the condition **(ES)**

[7] This is especially problematic, for the AGM theory provides no guidance as to the conditionals an agent should adopt

ensures that one can effectively test **(PF)** without resorting to hypothetical contraction. We first note that **(PF)** reduces to **(P)** for hypothetical explanations; for if $\beta \notin K$ then $K^-_\beta = K$. For factual explanations, **(ES)** requires that both α and β are believed. The following proposition shows that **(PF)** can be evaluated without iterated belief change.

Proposition 3.1 *If* $\alpha, \beta \in K$*, then* $\beta \in (K^-_\beta)^*_\alpha$ *iff* $\neg\alpha \in K^*_{\neg\beta}$.

Thus condition **(PF)**, in the presense of **(ES)**, is equivalent to the following condition pertaining to the absence of the observation:

(A) $\neg\alpha \in K^*_{\neg\beta}$

which is expressed in the object language as $\neg\beta \Rightarrow \neg\alpha$. This captures the intuition that *If the observation had been absent, so too would be the explanation.*

This condition is now vacuous when the observation is rejected in K, for $K^*_{\neg\beta} = K$ and we must have $\neg\alpha \in K$ by **(ES)**. It seems plausible to insist that an agent ought to imagine the explanation to be possible and then test if rejection of the observation leads to rejection of the explanation; in other words:

(AR) $\neg\alpha \in (K^-_{\neg\alpha})^*_{\neg\beta}$

However, just as **(PF)** reduces to **(A)**, so too does **(AR)** reduce to **(P)**.

Proposition 3.2 *If* $\neg\alpha, \neg\beta \in K$*, then* $\neg\alpha \in (K^-_{\neg\alpha})^*_{\neg\beta}$ *iff* $\beta \in K^*_\alpha$.

Thus, we are lead to the notion of a predictive explanation, relative to some epistemic state.

Definition 3.1 Let M be a K-revision model reflecting the epistemic state of an agent with belief set K. A *predictive explanation* for observation β (relative to M) is any $\alpha \in \mathbf{L}_{CPL}$ such that:

(ES) $M \models (\mathsf{B}\alpha \equiv \mathsf{B}\beta) \wedge (\mathsf{B}\neg\alpha \equiv \mathsf{B}\neg\beta)$;

(P) $M \models \alpha \Rightarrow \beta$; and

(A) $M \models \neg\beta \Rightarrow \neg\alpha$.

---

in this contracted belief state. Very little can be known about the content of belief sets that are changed more than once as required by **(PF)**. The AGM theory does not provide a method for determining the structure of the resulting epistemic state, even if the original epistemic state and belief set K are completely known (but for a recently developed model that captures such iterated revision, see [6]).

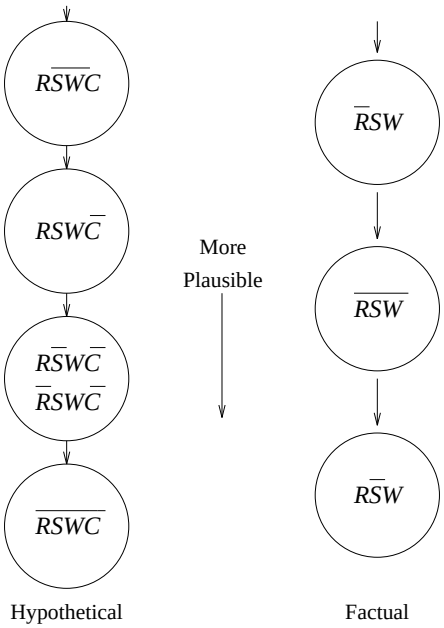

Figure 2: Explanations for "Wet Grass"

The reductions afforded by Propositions 3.1 and 3.2 are crucial, for they allow an agent to test whether an explanation is valid relative to its current epistemic state (or its current set of simple conditionals). An agent is not required to perform hypothetical contraction.

This definition captures both factual and hypothetical predictive explanations. Furthermore, once the epistemic status of β is known we need only test one of the conditions **(A)** or **(P)**.

Proposition 3.3 *If $\alpha, \beta \in K$ then α (predictively) explains β iff $\neg\beta \Rightarrow \neg\alpha$.*

Proposition 3.4 *If $\alpha, \beta, \neg\alpha, \neg\beta \notin K$ then α (predictively) explains β iff $\alpha \Rightarrow \beta$ iff $\neg\beta \Rightarrow \neg\alpha$.*

Proposition 3.5 *If $\neg\alpha, \neg\beta \in K$ then α (predictively) explains β iff $\alpha \Rightarrow \beta$.*

Example 3.1 Figure 2 illustrates both factual and hypothetical explanations. In the first model, the agent believes the following are each false: the grass is wet (W), the sprinkler is on (S), it rained (R) and the grass is covered (C). W is explained by sprinkler S, since $S \Rightarrow W$ holds in that model. So should the agent observe W, S is as possible explanation; should the agent desire W to be true (and have control over S) it can ensure W by causing S to hold. Similarly, R explains W, as does $S \wedge R$. Thus, there may be competing explanations; we discuss preferences

on these below. Intuitively, α explains β just when β is true at the most plausible situations in which α holds. Thus, explanations are *defeasible*: W is explained by R; but, R together with C does not explain wet grass, for $R \wedge C \Rightarrow \neg W$. Notice that R *alone* explains W, since the "exceptional" condition C is normally false when R holds, thus need not be stated. This defeasibility is a feature of explanations that has been given little attention in many logic-based approaches to abduction.

The second model illustrates factual explanations for W. Since W is believed, explanations must also be believed. R and $\neg S$ are candidates, but only R satisfies the condition on factual explanations: if we give up belief in W, adding R is sufficient to get it back. In other words, $\neg W \Rightarrow \neg R$. This does not hold for $\neg S$ because $\neg W \Rightarrow S$ is false. ■

The crucial features of predictive explanations illustrated in this example are their defeasibility, the potential for competing explanations, and the distinction between factual and hypothetical explanations.

Notice that if we relax the condition **(ES)** in the factual example above, we might accept S as a hypothetical explanation for factual belief W. Although, we believe R, W and $\neg S$, one might say that "Had the sprinkler been on, the grass (still) would have been wet." This slightly more permissive form of predictive explanation, called *counterfactual explanation*, is not explored further here (but see [10] for further details).

3.1.1 Causal Explanations

The notion of explanation described here cannot be given a truly causal interpretation. In the factual model in Figure 2, we suggested that rain explains wet grass. However, it is also the case that wet grass explains rain. Explanations are simply beliefs (whether factual or hypothetical) that induce belief in the fact to be explained. The connection may be causal (belief in R induces belief in W) or evidential (belief in W induces belief in R).

Ultimately, we would like to be able to distinguish causal from non-causal explanations in this conditional model. Lewis [37] has proposed a counterfactual analysis of causation, whereby a theory of conditionals might be used to determine causal relations between propositions. More recently, and perhaps more compelling, is the use of *stratified rankings* on conditional theories by Goldszmidt and Pearl [27] to represent causation. Incorporating such considerations in our model brings to mind Shoham's [55] epistemic account of causality, whereby a causal theory is expressed in terms of the knowledge of an agent, and can be nonmonotonic. Whether or not causality is an epistemic notion (cf. the critique of Galton [20]), it is clear that perceived causal relations will have a dramatic impact

on the conditional beliefs of an agent. Furthermore, it is the epistemic state of an agent with respect to which causal predictions and explanations must be derived. In this regard, an epistemic theory of causal explanation is consistent with Shoham's viewpoint. However, a more sophisticated account of causation is necessary in order to distinguish causal from evidential relations among an agent's beliefs.[8] A more suitable theory should include some account of actions, events, and "intervention" [27]. For instance, if a (possibly hypothetical) mechanism exists for independently wetting the grass (W) and making it rain (R), this can be exploited to show that W does not cause R, but that R causes W, according to the plausibility judgements of an agent. Such experimentation or experience can be used to distinguish causal from evidential explanations.

Another similarity between conditionals and Shoham's causal statements are their context-sensitivity. Simon [56] argues that one potential drawback in Shoham's theory is the necessity of distinguishing *causal* from *contextual* conditions and the asymmetry this introduces. While this may or may not be a necessary feature of "true" causal relations, it is a fact of life in any useful epistemic account, for we naturally communicate and acquire our causal knowledge making such distinctions. Simon finds disquieting the fact that the roles of cause and contextual condition are sometimes reversed; but the dependence of the form of causal utterances on circumstances is exactly what we capture when we evaluate causal statements with respect to an epistemic state. Imagine an agent possesses two conditionals $R \Rightarrow W$ and $R \wedge C \Rightarrow \neg W$: the grass gets wet when it rains unless it's covered. Taking $\neg C$ to be the normal case, it seems natural to offer R as a causal explanation (or cause) for W, and take $\neg C$ to be a contextual condition. This offers a certain economy in thinking about and communicating causes. However, in a different epistemic setting, without altering the underlying physical causal relations (whatever they may be, or if they even exist), these roles may reversed. If the grass is typically covered, we may have $R \Rightarrow \neg W$ and $R \wedge \neg C \Rightarrow W$. Supposing that it usually rains, $\top \Rightarrow R$, an intuitive causal explanation for W relative to this epistemic state is now $\neg C$, someone uncovered the grass. R is relegated to the role of contextual condition. This asymmetry, far from being problematic, is natural and desirable. We do not delve further into causal explanations here, but we conjecture that conditional logics will provide a natural and flexible mechanism for representing causal relations and generating causal explanations with an epistemic flavor.

[8] Temporal precedence, one mechanism available in Shoham's theory, cannot resolve such issues in general. For instance, the truth of E at time t may be evidence for the truth of fact F at time $t+1$ without having caused it.

3.2 Nonpredictive Explanations

3.2.1 *Might* Explanations

Very often we are interested in weaker types of explanation that do not predict an observation, but merely "allow" it. For example, suppose Fred has a choice of three supermarkets at which to shop, one very close (A), and two rather farther away (B and C). We expect Fred to shop at the closest A, but observe that he actually chooses to shop at C. We might explain Fred's choice by claiming that (D) Fred dislikes the service at market A. However, explanation D does not *predict* that Fred will choose C, for he may well have chosen B. That is, we do not accept the conditionals $D \Rightarrow C$ or $D \Rightarrow B$, but only $D \Rightarrow B \vee C$. In a sense D "excuses" or *permits* C but does not *predict* C. If we learned D, we would claim that Fred *might* go to C. Upon learning C, we adopt the explanation D. A similar example is captured by the hypothetical model in Figure 2: here W permits both R and S without predicting them. *Might explanations* of this type play an important role in consistency-based diagnosis without fault models as well (see Section 5).

Intuitively, a might explanation reflects the slogan *If the explanation were believed, the observation would be a possibility*. The sense of "possible" here is naturally that of epistemic possibility. If an agent accepts explanation α, the observation β becomes consistent with its new belief set. The might condition is simply

(M) $\neg\beta \notin K^*_\alpha$

which is expressed as $\alpha \not\Rightarrow \neg\beta$.

For hypothetical explanations of rejected β (where $\neg\beta \in K$), might explanations require nothing further. However, for explanations of indeterminate β, we must weaken the condition **(ES)**. If β is indeterminate, it is already a possibility for the agent, and we should not rule out beliefs $\alpha \in K$ as potential might explanations: if α is believed (it is!) then β is possible (it is!). Such might explanations are not very informative, however, so we take the principle case for might explanations to be that where β is rejected. Thus, **(ES)** is again replaced by **(F)**:[9]

(F) If $\beta \in K$ then $\alpha \in K$.

Definition 3.2 Let M be a K-revision model reflecting the epistemic state of an agent with belief set K. A *might explanation* for observation β (relative to M) is any $\alpha \in \mathbf{L}_{CPL}$ such that:

(F) $M \models \mathsf{B}\beta \supset \mathsf{B}\alpha$; and

[9] This weakening of **(ES)** does not affect the principle case where $\neg\beta \in K$. If $\neg\alpha \notin K$, then $\alpha \not\Rightarrow \neg\beta$ cannot hold. So $\neg\beta \in K$ ensures $\neg\alpha \in K$ for all might explanations.

(M) $M \models \alpha \not\Rightarrow \neg\beta$.

Intuitively, the epistemic state induced by acceptance of α must contain β-worlds, hence rendering β possible. If it contains only β-worlds then α is a predictive explanation. Predictive explanations are therefore a special (stronger) case of might explanations.

Proposition 3.6 *If α is a predictive explanation for β then α is a might explanation for β.*

We take might explanations to be the primary form of non-predictive explanation.

3.2.2 A Variant of Might Explanations

In this section we describe a form of might explanation that is of particular relevance to CT4O-models, where clusters of equally plausible worlds are partially ordered rather than totally ordered. This form of explanation is somewhat difficult to motivate independently, but in Section 4 we will see that it is precisely the type of explanation used by Theorist.

Clearly, a sentence α can be a might explanation for both β and $\neg\beta$. This is similar to the behavior of the weak conditional connective $\rightarrow$, where $\alpha \rightarrow \beta$ and $\alpha \rightarrow \neg\beta$ can be held consistently. Recall that a sentence $\alpha \rightarrow \beta$ holds just when β holds at all worlds in some element of $Pl(\alpha)$, (i.e., at minimal cluster of α-worlds). We call α a *weak explanation* for β just in case it is a might explanation such that $\alpha \rightarrow \beta$.

Definition 3.3 Let M be a K-revision model reflecting the epistemic state of an agent with belief set K. A *weak explanation* for observation β (relative to M) is any $\alpha \in \mathbf{L}_{CPL}$ such that:

(F) $M \models \mathsf{B}\beta \supset \mathsf{B}\alpha$; and

(W) $M \models \alpha \rightarrow \beta$.

Intuitively, weak explanations lie between predictive and might explanations. They are stronger than might explanations, for they require, at some cluster of most plausible α-worlds, that β holds. All other most plausible α-worlds are of incomparable plausibility, so in some sense α is "potentially predictive" (it "could" be that the relevant cluster is actually *min*(α), if only one could render all worlds comparable). On the other hand, weak explanations are weaker than predictive explanations in the sense that certain *min*(α)-worlds do not (in the principle case) satisfy the observation. Weak explanations are therefore a special (stronger) case of might explanations.

Proposition 3.7 *If α is a weak explanation for β then α is a might explanation for β.*

Naturally, in the logic CO, since $\alpha \rightarrow \beta$ iff $\alpha \Rightarrow \beta$, weak explanations are predictive. Therefore, weak explanations will only be used in the context of CT4O-models. In the CT4O-model in Figure 1(a), A is a weak explanation for both C and $\neg C$.

3.3 Preferences

The explanations defined above carry the explanatory force we expect, whether predictive or not, yet are more flexible than deductive explanations. They exhibit the desired defeasibility, allowing exceptions and more specific information to override their explanatory status. However, the criteria we propose admit many explanations for a given observation in general: any α sufficient to induce belief in β counts as a valid predictive explanation. For instance, rain explains wet grass; but a tanker truck full of milk exploding in front of the yard also explains wet grass. If you *could* convince someone that such an event occurred, you would convince them that the grass was wet.

Certainly some explanations should be preferred to others on grounds of likelihood or plausibility. In probabilistic approaches to abduction, one might prefer most probable explanations. In consistency-based diagnosis, explanations with the fewest abnormalities are preferred on the grounds that (say) multiple component failures are unlikely. Such preferences can be captured in our model quite easily. Our CT4O- and CO-structures rank worlds according to their degree of plausibility, and reasonable explanations are simply those that occur at the most plausible worlds. We recall from Section 2.2 the notion of plausibility as applied to propositions. A is at least as plausible as B just when, for every B-world w, there is some A-world that is at least as plausible as w. For CO-models, this totally orders propositions; but for CT4O-models, two propositions may have incomparable "degrees" of plausibility.

An adopted explanation is not one that simply makes an observation less surprising, but one that is itself as unsurprising as possible. We use the plausibility ranking to judge this degree of surprise.

Definition 3.4 If α and α' both explain β then α is *at least as preferred as* α' (written $\alpha \leq_P \alpha'$) iff $M \models \vec{\Box}(\alpha' \supset \Diamond\alpha)$. The *preferred explanations* of β are those α such that for no explanation α' is it the case that $\alpha' <_P \alpha$.

Preferred explanations are those that are most plausible, that require the "least" change in belief set K in order to be accepted. Examining the hypothetical model in Figure 2, we see that R, S and $R \wedge S$ each explain W; but R and S are preferred to $R \wedge S$ (it may not be known whether the sprinkler was on or it rained, but it's unlikely that the sprinkler was on in the rain). Any world in which a tanker truck explodes is less plausible than these other worlds, so that explanation is given relatively less credibility.

By basing the notion of preference on the relative plausibility of explanations, we lose the ability to distinguish factual explanations from one another. The conditions **(ES)** and **(FS)** ensure that every valid explanation of a factual observation is believed, and all beliefs are equally (and maximally) plausible for an agent. Thus, *each* candidate explanation is preferred. This fits well with the point of view adopted above: an agent, when accepting β, also accepts its most plausible explanation(s). There is no need, then, to rank factual explanations according to plausibility – all explanations in K are equally plausible. If one wanted to distinguish possible explanations of some belief β, one might distinguish the hypothetical explanations of β in the contracted belief state K^-_β. Most plausible explanations are then those that the agent judged to be most plausible before accepting β. However, such a move serves no purpose, for the most preferred explanations in state K^-_β must be beliefs in K.

Proposition 3.8 *Let $\beta \in K$ and α be a predictive explanation for β. Then α is a preferred (hypothetical) explanation for β in K^-_β.*

It is not hard to see that preferences cannot be applied to hypothetical explanations of indeterminate β for precisely the same reason: all valid explanations must be epistemically possible, and therefore maximally plausible, this because **(ES)** requires $\neg\alpha \notin K$. For these reasons, when describing preferences, we restrict our attention to hypothetical explanations of *rejected* β.

A predictive explanation needn't be compared to all other explanations in order to determine if it is most preferred. The following proposition indicates a simpler test for preference.

Proposition 3.9 *Let α be a predictive explanation for β relative to model M. Then α is a preferred explanation iff $M \models \beta \not\rightarrow \neg\alpha$.*

This test simply says that in any cluster of most normal β-worlds, if α is a preferred explanation of β, then an α-world must occur somewhere in that cluster, for this is (potentially) the most plausible cluster of situations in which the observation holds.

The test is greatly simplified, and much clearer, for totally-ordered CO-models. This due to the equivalence of $\rightarrow$ and $\Rightarrow$ under CO.

Proposition 3.10 *Let α be a predictive explanation for β relative to CO-model M. Then α is a preferred explanation iff $M \models \beta \not\Rightarrow \neg\alpha$.*

In this case, α is a preferred explanation iff belief in β does not preclude the possibility of α. Preferred explanations are those that are most plausible, that require the "least" change in belief set K in order to be accepted. Examining the hypothetical model in Figure 2, we see that $W \not\Rightarrow \neg R$ and $W \not\Rightarrow \neg S$ holds, but $W \not\Rightarrow \neg(S \wedge R)$ is false. So R and S are preferred explanations, while explanation $S \wedge R$

is not.[10]

3.4 The Pragmatics of Explanation

In any actual system for explanation, ultimately a *sentence* must be returned which explains the given observation. The semantic conditions we have proposed admit explanations that are intuitively unsatisfying in some circumstances. Of the many explanations, some may be preferred on grounds other than plausibility. Natural criteria such as simplicity and informativeness are often used to rule out certain explanations in certain contexts [47]. Levesque [35] has proposed criteria for judging the simplicity of explanations. Hobbs *et al* [30] argue that in natural language interpretation most specific explanations are often required, rather than simple explanations. In diagnostic systems, often this problem is circumvented, for explanations are usually drawn from a prespecified set of conjectures [44] (see Sections 4 and 5).

It is clear that the exact form an explanation should take is influenced by the application one has in mind. Therefore, we do not include such considerations in our semantic account of abduction. Rather, we view these as *pragmatic* concerns, distinct from the semantic issues involved in predictiveness and plausibility (cf. Levesque [35]). Providing an account of the pragmatics of explanations is beyond the scope of this paper; but we briefly review two such issues that arise in our framework: trivial explanations and irrelevant information.

3.4.1 Trivial Explanations

A simple theorem of CT4O and CO is $\beta \Rightarrow \beta$. This means that β is always a predictive (and preferred) explanation for itself. While this *trivial explanation* may seem strange, upon reflection it is clear that no other proposition has a stronger claim on inducing belief in an observation than the observation itself. This makes the task of explanation quite simple! Unfortunately, a system that provides uninformative trivial explanations will not be deemed especially helpful.

We expect pragmatic considerations, much like Gricean maxims, to rule out uninformative explanations where possible. For instance, one might require that an explanation be semantically distinct from the observation it purports to explain. However, the semantics should not rule out trivial explanations. In some applications a trivial explanation may be entirely appropriate. Consider causal explanations in a causal network. One might expect a causal explanation for a node having a particular value to consist of some assignment of values to its ancestors. However, when asked to explain a

[10] When there are several disjoint preferred explanations (e.g., R, S), we may be interested in *covering explanations*, that capture *all* of the plausible *causes* of an observation. We refer to [10] for a discussion of this notion.

root node, no explanation but the trivial explanation seems appropriate. Presumably, in any abstract model of a domain, causes (hence causal explanations) cannot be traced back *ad infinitum*.[11]

3.4.2 Irrelevant Information

Very often one can strengthen or weaken an explanation with extraneous information and not affect its explanatory power. But such constructions often result in explanations that are intuitively unsatisfying. Suppose as usual that the sprinkler being on explains wet grass, so $S \Rightarrow W$. Suppose furthermore that the conditionals $S \Rightarrow O$ and $S \Rightarrow \neg O$ are both rejected by the agent, where O stands for "Fred's office door is open," something about which our agent has no information. A simple inference ensures that $(S \wedge O) \Rightarrow W$ and $(S \wedge \neg O) \Rightarrow W$ both hold. Thus, $S \wedge O$ and $S \wedge \neg O$ both explain W. Yet, intuitively both of these explanations are unappealing — they contain information that is *irrelevant* to the conclusion at hand.

In order to rule out such explanations, we expect the pragmatic component of an abductive system to filter out semantically correct explanations that are inappropriate in a given context. In Poole's Theorist system, for example, explanations are drawn from a prespecified set of conjectures. We can view this as a crude pragmatic "theory." Levesque [35] embeds a syntactic notion of *simplicity* in his semantics for abduction. In our conditional framework one can define conditions under which a proposition is deemed irrelevant to a conditional [21, 3].

Explanations can also be strengthened with "background information" that, while not irrelevant, can be left unstated. For instance, returning to the example given by the factual model in Figure 2, we can see that R explains W, and $R \wedge \neg C$ explains W as well. However, since $\neg C$ normally holds when R holds (i.e., $R \Rightarrow \neg C$), it can be left as a tacit assumption. Certainly, $\neg C$ is relevant, for $R \wedge C \Rightarrow \neg W$, but it needn't be stated as part of the explanation. This suggests that logically weak explanations are to be (pragmatically) preferred. It also suggests a mechanism whereby an abductive system can elaborate or clarify its explanations. Should an explanation be questioned, the system can identify tacit knowledge that is deemed relevant to the explanation and elaborate by providing these facts.

One can weaken explanations by disjoining certain information to valid explanations, retaining explanatory power. In general, if A explains B, and C is less plausible that A, then $A \vee C$ explains B as well. Since $(A \vee C) \Rightarrow A$ (because C is less plausible than A), we must have $(A \vee C) \Rightarrow B$. If rain explains wet grass, so does "It rained or the lawn was covered," since C is less plausible than R. Once again, we view the weaker explanation as violating (something like) the Gricean maxim of

[11] "Why is the grass wet?" "Because it rained." "Why did it rain?" "It just did!"

Informativeness: the explanation R is certainly more informative that the weaker $R \vee C$ (but still relevant). The explanation $R \vee C$ also carries with it the unwanted implicature that both disjuncts are (individually) valid explanations. This is strongly related to the following issue that arises in the study of conditional logics: sentences with the linguistic form $(A \vee C) \Rightarrow B$ are usually intended to represent an assertion with the logical form $(A \Rightarrow B) \vee (C \Rightarrow B)$ [40].

4 Abductive Models of Diagnosis

One of the main approaches to model-based diagnostic reasoning and explanation are the so-called "abductive" theories. Representative of these models is Poole's [43, 44] Theorist framework for explanation and prediction, and Brewka's [12] extension of it. In this section, we describe both models, how they can be embedded within our framework, and how the notions we defined in the last section can be used to define natural extensions of the Theorist framework. This also provides an object-level semantic account of Theorist.

4.1 Theorist and Preferred Subtheories

Poole [43, 44] presents a framework for hypothetical reasoning that supports explanation and default prediction. Theorist is based on *default theories*, pairs $\langle \mathcal{F}, \mathcal{D} \rangle$ where $\mathcal{F}$ and $\mathcal{D}$ are sets of sentences.[12] The elements of $\mathcal{F}$ are *facts*, known to be true of the situation under investigation. We take $\mathcal{D}$ to be a set of *defaults*, sentences that are normally true, or *expectations* about typical states of affairs. Although nothing crucial depends on this, we assume $\mathcal{D}$ to be consistent. Poole also uses a set $\mathcal{C}$ of *conjectures* that may be used in the explanation of observations, but should not be used in default prediction.[13]

Definition 4.1 [44] An *extension* of $\langle \mathcal{F}, \mathcal{D} \rangle$ is any set *Cn*$(\mathcal{F} \cup D)$ where D is a maximal subset of $\mathcal{D}$ such that $\mathcal{F} \cup D$ is consistent.

Intuitively, extensions are formed by assuming as many defaults as possible. Since defaults are expected to be true, each extension corresponds to a "most normal" situation at which $\mathcal{F}$ holds. A

[12] Poole's presentation is first-order, using ground instances of formulae in the definitions to follow. For simplicity, we present only the propositional version.

[13] The following definitions are slightly modified, but capture the essential spirit of Theorist. We ignore two aspects of Theorist, *constraints* and *names*. While constraints can be used to rule out undesirable extensions for prediction, it is generally accepted that *priorities*, which we examine below, provide a more understandable mechanism for resolving conflicts. The role of constraints in explanation has largely been ignored. Named defaults add no expressive power to Theorist; they can be captured by introducing the names themselves as the only (atomic) defaults.

(skeptical) notion of default prediction is defined by considering what is true at each such normal situation.

Definition 4.2 [44] Sentence A is *predicted* by $\langle \mathcal{F}, \mathcal{D} \rangle$ iff A is in each extension of $\langle \mathcal{F}, \mathcal{D} \rangle$.

Conjectures play a key role in abduction, and can be viewed as possible hypotheses that (together with certain defaults) explain a given observation β.

Definition 4.3 [44] $C \cup D$ is a (Theorist) *explanation* for observation β (w.r.t. $\langle \mathcal{F}, \mathcal{D}, \mathcal{C} \rangle$) iff $C \subseteq \mathcal{C}$, $D \subseteq \mathcal{D}$, $C \cup D \cup \mathcal{F}$ is consistent and $C \cup D \cup \mathcal{F} \models \beta$.

Since we take defaults to be assumptions pertaining to the normal course of events, the set C of adopted conjectures carries the bulk of the explanatory force of a Theorist explanation. Just as we ignore "causal rules" and "scientific laws" in our earlier definition of predictive explanation, here we take the default component of an explanation to be "understood," and take a set C of conjectures to be a Theorist explanation iff there is some set of defaults D that satisfies the required relation. We assume sets C, D and $\mathcal{F}$ are finite and sometimes treat them as the conjunction of their elements.

Example 4.1 Let $\mathcal{F} = \{U, A\}$, $\mathcal{D} = \{U \supset A, A \supset E, U \supset \neg E, R \supset \neg P\}$ and $\mathcal{C} = \{U, A, E\}$, where U, A, E, R and P stand for *university student*, *adult*, *employed*, *Republican* and *Pacifist*, respectively. The extensions of this default theory are

$$Cn\{U, A, U \supset A, A \supset E, R \supset \neg P\}$$
$$Cn\{U, A, U \supset A, U \supset \neg E, R \supset \neg P\}$$

Thus A is predicted, but neither E nor $\neg E$ are predicted.

Suppose now that $\mathcal{F} = \emptyset$. The conjecture A explains E, but does not explain $\neg E$. Thus, if one adopted belief in A, one would predict E. In a similar fashion, U explains E; but U also explains $\neg E$. Notice that $\neg P$ is not explainable. ■

The last explanation in this example illustrates that Theorist explanations are, in a certain sense, *paraconsistent*: a conjecture may explain both a proposition and its negation. Certainly, such explanations cannot be construed as predictive. Notice also that certain propositions may not have explanations of the type defined by Theorist, but can be explained (nontrivially) if we allow explanations that do not lie within the set of conjectures. Intuitively, we might want to admit R as a valid explanation of $\neg P$ even though it is not listed among our assumable hypotheses in $\mathcal{C}$.

In the example above, the second extension is more satisfying than the first. The fact that university students are a specific subclass of adults suggests that the default rule $U \supset \neg E$ should be applied instead of $A \supset E$. Brewka [12] has extended the Theorist framework for default prediction by introducing priorities on defaults to handle such a situation.

Definition 4.4 A *Brewka theory* is a pair $\langle \mathcal{F}, \langle \mathcal{D}_1, \cdots, \mathcal{D}_n \rangle \rangle$ where $\mathcal{F}$ is a set of facts and each $\mathcal{D}_i$ is a set of defaults.

Intuitively, $\langle \mathcal{D}_1, \cdots, \mathcal{D}_n \rangle$ is an ordered set of default sets, where the defaults in the lower ranked sets have higher priority than those in the higher ranked sets. We will say that default $d \in \mathcal{D}_i$ has priority over default $e \in \mathcal{D}_j$ if $i < j$. When constructing extensions of such a theory, if two default rules conflict, the higher priority rule must be used rather than the lower priority rule. Multiple extensions of a theory exist only when default rules of the same priority conflict with the facts or higher priority rules. A Theorist default theory (with no conjectures) $\langle \mathcal{F}, \mathcal{D} \rangle$ is a Brewka theory with a single priority level. The *reduction* of a Brewka theory to a (Theorist) default theory is $\langle \mathcal{F}, \mathcal{D} \rangle$, where $\mathcal{D} = \mathcal{D}_1 \cup \ldots \cup \mathcal{D}_n$. Brewka's *preferred subtheories* (hereafter dubbed extensions) are constructed in the obvious way.

Definition 4.5 An *extension* of a Brewka theory $\langle \mathcal{F}, \langle \mathcal{D}_1, \cdots, \mathcal{D}_n \rangle \rangle$ is any set

$$E = Cn(\mathcal{F} \cup D_1 \cup \ldots \cup D_n)$$

where, for all $1 \leq k \leq n$, $\mathcal{F} \cup D_1 \cup \ldots \cup D_k$ is a maximal consistent subset of $\mathcal{F} \cup \mathcal{D}_1 \cup \ldots \cup \mathcal{D}_k$.

Thus, extensions are constructed by adding to $\mathcal{F}$ as many defaults from $\mathcal{D}_1$ as possible, then as many defaults from $\mathcal{D}_2$ as possible, and so on. The following proposition should be clear:

Proposition 4.1 *Every extension of a Brewka theory $\langle \mathcal{F}, \langle \mathcal{D}_1, \cdots, \mathcal{D}_n \rangle \rangle$ is a Theorist extension of its reduction $\langle \mathcal{F}, \mathcal{D} \rangle$.*

Prediction based on a Brewka theory is defined in the obvious way, as membership in all extensions. It then becomes clear that:

Proposition 4.2 *A is predicted by a Brewka theory $\langle \mathcal{F}, \langle \mathcal{D}_1, \cdots, \mathcal{D}_n \rangle \rangle$ if it is predicted by its reduction $\langle \mathcal{F}, \mathcal{D} \rangle$.*

In other words, Brewka theories allow (typically strictly) more predictions than their Theorist counterparts. In the example above, should we divide $\mathcal{D}$ into priority levels by placing $U \supset A$ and $U \supset \neg E$

in $\mathcal{D}_1$ and $A \supset E$ in $\mathcal{D}_2$, we are left with a single extension $Cn\{U, U \supset A, U \supset \neg E, R \supset \neg P\}$, and $\neg E$ is predicted.

Brewka does not provide a notion of explanation, but the Theorist definition of explanation will suffice. That is, α explains β iff $\{\alpha\} \cup \mathcal{F}$ is consistent with some set of defaults $D \subseteq \mathcal{D}_1 \cup \ldots \cup \mathcal{D}_k$ such that $\{\alpha\} \cup \mathcal{F} \cup D \models \beta$. Again, we will often draw explanations from a prespecified set of conjectures. This definition retains the essential properties of the Theorist definition, in particular, its paraconsistent flavor.

4.2 Capturing Theorist in CT4O

Our goal is to represent and extend the notion of explanation in Theorist by embedding it within our conditional framework. This will have the effect of providing a semantic interpretation in our conditional logic for Theorist's notion of explanation and prediction. In what follows, we assume a fixed, consistent set of defaults $\mathcal{D}$, but the sets $\mathcal{F}$ and $\mathcal{C}$ of facts and conjectures, respectively, will be allowed to vary.[14]

The definitions of extension and prediction in Theorist suggest that the more defaults a situation satisfies, the more normal that situation is. We capture the normality criterion implicit in Theorist by ranking possible worlds according to the default sentences they falsify (or *violate*).

Definition 4.6 For any possible world $w \in W$, the set of defaults *violated* by w is

$$V(w) = \{d \in \mathcal{D} : w \models \neg d\}$$

If we interpret defaults as normality assumptions, clearly the ordering of worlds should be induced by set inclusion on these violation sets. This gives rise to a suitable CT4O*-model, the *Theorist structure,* for a set of defaults $\mathcal{D}$.

Definition 4.7 The *Theorist structure* for $\mathcal{D}$ is $M_{\mathcal{D}} = \langle W, \leq, \varphi \rangle$ where W is the set of truth assignments suitable for $\mathbf{L}_{CPL}$; φ is the valuation function induced by W; and $v \leq w$ iff $V(v) \subseteq V(w)$.

Proposition 4.3 *$M_{\mathcal{D}}$ is a CT4O*-model.*

The model $M_{\mathcal{D}}$ divides worlds into clusters of equally plausible worlds that violate the same set of defaults in $\mathcal{D}$. If $V(w) = V(v)$ then $w \leq v$ and $v \leq w$. Otherwise, v and w must be in different clusters.

[14] The consistency of the set $\mathcal{D}$ is not crucial to our representation, but allows the presentation to be simplified. We will point out various properties of our model that depend on this assumption and how they are generalized when $\mathcal{D}$ is not consistent.

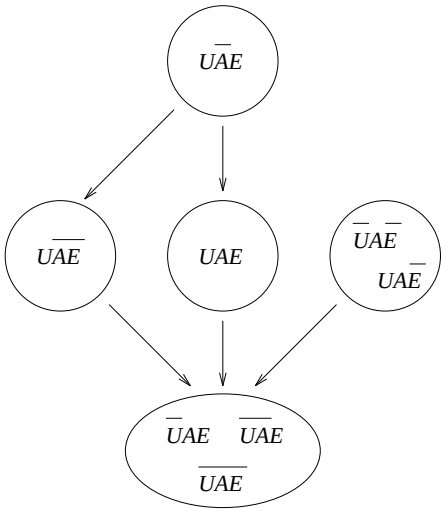

Figure 3: Theorist Model for the University Students Example

Proposition 4.4 *$\mathcal{C}$ is a cluster in the model $M_{\mathcal{D}}$ iff for some $D \subseteq \mathcal{D}$*

$$\mathcal{C} = \{w : w \models d \text{ if } d \in D \quad \textit{and} \quad w \models \neg d \text{ if } d \in \mathcal{D} - D\}$$

Since $\mathcal{D}$ is finite, any model $M_{\mathcal{D}}$ consists of a finite set of clusters. Figure 3 depicts the Theorist model for the default set $\mathcal{D} = \{U \supset A, A \supset E, U \supset \neg E\}$. The bottom cluster contains those worlds that violate no defaults, that is, the most normal worlds. The middle clusters (from left to right) violate the default sets $\{U \supset A\}$, $\{U \supset \neg E\}$ and $\{A \supset E\}$, respectively. The least plausible worlds violate the default set $\{U \supset A, U \supset \neg E\}$. Notice that the model $M_{\mathcal{D}}$ is sensitive to the syntactic structure of the default set $\mathcal{D}$. Logically equivalent sets of defaults can result in drastically different models, reflecting the syntax-sensitivity exhibited by Theorist.

To interpret this model, we view the defaults in $\mathcal{D}$ as expectations held by an agent, statements regarding the most normal or plausible states of affairs. If an agent has no "factual beliefs," it would adopt this set of defaults as its only beliefs. Thus, the model $M_{\mathcal{D}}$ captures the epistemic state of an agent who has yet to encounter any default violations. In a diagnosis application, we might think of such a belief state as representing the normal functioning of a system. Notice that since $\mathcal{D}$ is consistent

the model $M_{\mathcal{D}}$ has a unique minimal cluster.[15]

The facts $\mathcal{F}$ play no role in the definition of the model $M_{\mathcal{D}}$. The manner in which we define prediction and explanation relative to this model below will account for $\mathcal{F}$ by using these facts in the antecedents of relevant conditionals. This allows a single model to be used for a variety of different sets of facts. One can explicitly account for $\mathcal{F}$ in the model by ruling out any worlds falsifying $\mathcal{F}$ (e.g., by using the axiom $\overleftrightarrow{\Box}\mathcal{F}$). However, we find the current formulation more convenient.

4.2.1 Prediction

Extensions of a default theory $\langle \mathcal{F}, \mathcal{D} \rangle$ are formed by considering maximal subsets of defaults consistent with the facts $\mathcal{F}$. Recall the definition of a *most plausible set* of A-worlds in a CT4O-model for some proposition A from Section 2:

$$Pl(A) = \{\mathit{min}(A) \cap \mathcal{C} : \mathcal{C} \text{ is a cluster}\}$$

By Proposition 4.4, the worlds in some most plausible set of A-worlds must violate exactly the same defaults. In the Theorist model, an extension must then correspond to a set of most plausible $\mathcal{F}$-worlds.

Proposition 4.5 *E is an extension of $\langle \mathcal{F}, \mathcal{D} \rangle$ iff $\|E\| = S$ for some $S \in Pl(\mathcal{F})$.*

Corollary 4.6 *A is in some extension of $\langle \mathcal{F}, \mathcal{D} \rangle$ iff $M_{\mathcal{D}} \models \mathcal{F} \rightarrow A$.*

Theorist predictions are those sentences true in all extensions. Since $\mathit{min}(\mathcal{F}) = \cup Pl(\mathcal{F})$, we have the following:

Theorem 4.7 *A is predicted (in Theorist sense) from default theory $\langle \mathcal{F}, \mathcal{D} \rangle$ iff $M_{\mathcal{D}} \models \mathcal{F} \Rightarrow A$.*

Thus, default predictions in Theorist correspond precisely to those sentences an agent would believe if it adopted belief in the facts $\mathcal{F}$. In other words, believing $\mathcal{F}$ induces belief in all (and only) default predictions.

Consider the example illustrated in Figure 3. We have $A \Rightarrow E$ and $A \Rightarrow \neg U$, corresponding to the Theorist predictions E and $\neg U$ when $\mathcal{F} = \{A\}$. Notice that $U \not\Rightarrow A$, $U \not\Rightarrow \neg A$, $U \not\Rightarrow E$ and $U \not\Rightarrow \neg E$ all hold, indicating that none of $A, \neg A, E, \neg E$ are predicted when $\mathcal{F} = \{U\}$. But $U \Rightarrow (E \supset A)$ holds so $E \supset A$ is predicted by Theorist when $\mathcal{F} = \{U\}$.

[15] If $\mathcal{D}$ is inconsistent, then we will have a minimal cluster corresponding to each maximal consistent subset of $\mathcal{D}$; i.e., a minimal cluster for each extension of $\mathcal{F} = \emptyset$.

4.2.2 Weak Explanations

To capture Theorist explanations, we assume the existence of a set $\mathcal{C}$ of conjectures from which possible explanations are drawn. Recall that $C \subseteq \mathcal{C}$ explains β (in the Theorist sense) iff C, together with $\mathcal{F}$ and some subset of defaults $D \subseteq \mathcal{D}$, entails β. When this relation holds, there clearly must exist a *maximal* such set of defaults consistent with C. This allows us to restrict our attention to such maximal subsets of $\mathcal{D}$. Essentially, we can exploit the result of Poole ensuring that β is explainable iff it is in some extension. The notion of weak explanation described in Section 3 precisely captures Theorist explanations.

Theorem 4.8 *Let $C \subseteq \mathcal{C}$. Then C is a Theorist explanation for β iff $M_{\mathcal{D}} \models (\mathcal{F} \wedge C) \rightarrow \beta$ and $\mathcal{F} \wedge C$ is consistent.*

In other words, C is a Theorist explanation iff $\mathcal{F} \wedge C$ is a weak explanation.[16]

The defeasibility of Theorist explanations is captured by the weak conditional $\rightarrow$. In $M_{\mathcal{D}}$ above we have that $A \rightarrow \neg U$, so A explains $\neg U$ (indeed, $\neg U$ is explainable with $\emptyset$). However, adding the fact $\neg E$ renders this explanation invalid, for $(A \wedge \neg E) \not\rightarrow \neg U$. The paraconsistent nature of Theorist explanations corresponds precisely to the paraconsistent nature of the connective $\rightarrow$. In the example above, we have $U \rightarrow E$ and $U \rightarrow \neg E$, so when $\mathcal{F} = \emptyset$, U explains both E and $\neg E$. If $\mathcal{F} = \{A\}$ then E is predicted; but U again explains $\neg E$, as well as E, for $U \wedge A \rightarrow E$ and $U \wedge A \rightarrow \neg E$ both hold.

4.2.3 Predictive Explanations

Some Theorist explanations do not exhibit this paraconsistent behavior. For instance, if $\mathcal{F} = \{U\}$ then E explains A since $U \wedge E \rightarrow A$. However, the even stronger relation $U \wedge E \Rightarrow A$ is true as well. Thus, given fact U, if E is adopted as a belief A becomes believed as well. The notion of *predictive explanation* as described in Section 3 seems especially natural and important. With respect to the Theorist model, we would expect a *predictive Theorist explanation* to be a set of conjectures C satisfying the relation $C \wedge \mathcal{F} \Rightarrow \beta$. While no such concept has been defined with the Theorist framework, we can extend Theorist with this capability.

Using the original ingredients of Theorist, a predictive explanation should be such that *all* (rather than some) extensions of the explanation (together with the given facts) contain the observation.

Definition 4.8 Let $C \subseteq \mathcal{C}$ and β be some observation. C is a *predictive explanation* for β iff $\beta \in E$ for all extensions E of $\langle \mathcal{F} \cup C, \mathcal{D} \rangle$.

[16] If explanations need not come from a prespecified pool of conjectures $\mathcal{C}$, then any α such that $\alpha \cup \mathcal{F}$ is a weak explanation will be considered a Theorist explanation.

Since prediction is based on considering the most normal situations consistent with some facts, predictive explanations should be evaluated with respect to all most normal situations satisfying that explanation. This definition reflects precisely the predictive explanations (in the CT4O sense) sanctioned by the Theorist model $M_{\mathcal{D}}$.

Theorem 4.9 *Let $C \subseteq \mathcal{C}$. Then C is a predictive Theorist explanation for β iff $M_{\mathcal{D}} \models (\mathcal{F} \wedge C) \Rightarrow \beta$ and $\mathcal{F} \wedge C$ is consistent.*

Notice that while the normative aspect of predictive Theorist explanations is explicitly brought out by Definition 4.8 (in particular, by the restriction to *maximal* subsets of defaults), it is *implicit* in the formulation $(\mathcal{F} \wedge C) \Rightarrow \beta$ of Theorem 4.9. This is due to the fact that the Theorist model $M_{\mathcal{D}}$ is constructed in such a way that maximal sets of defaults are "preferred," and the fact that $(\mathcal{F} \wedge C) \Rightarrow \beta$ is evaluated only in these most preferred situations satisfying $\mathcal{F} \wedge C$.

In our example above, A predictively explains E (with no facts) since $M_{\mathcal{D}} \models A \Rightarrow E$. Naturally, predictive explanations are defeasible: $M_{\mathcal{D}} \not\models U \wedge A \Rightarrow E$ so $U \wedge A$ fails to predictively explain E. If $\mathcal{F} = \{U\}$ then E predictively explains A since $U \wedge E \Rightarrow A$. The notion of predictive explanation described for epistemic explanations suggests a very natural and useful extension of the Theorist framework. Theorem 4.9 ensures that the predictive explanations defined in Definition 4.8 match our intuitions, while the definition itself demonstrates how our predictive explanations can be added directly to the Theorist framework.

4.2.4 Preferences

As with most approaches to abduction, Theorist admits a number of possible explanations, whether weak or predictive, and makes no attempt to distinguish certain explanations as preferred to others. Even if we restrict attention to explanations that are formed from elements of a conjecture set $\mathcal{C}$, certain explanations seem more plausible than others. For example, one may have a set of defaults

$$\mathcal{D} = \{R \supset W, S \supset W, R \supset \neg S\}$$

inducing the Theorist model pictured in Figure 4: rain and the sprinkler cause wet grass, and the sprinkler is on only if it isn't raining. Assuming $\mathcal{C} = \{R, S\}$ and $\mathcal{F} = \emptyset$, each of R, S and $R \wedge S$ (predictively) explain W. However, inspection of the model $M_{\mathcal{D}}$ suggests that, in fact, the explanation $R \wedge S$ should be less preferred than the others. This is due to the fact that the ordering of plausibility on propositions induced by $M_{\mathcal{D}}$ makes $R \wedge S$ less plausible than R or S.

Theorist provides no notion of preference of this type; but our definition of preference from Section 3 readily lends itself to application within the Theorist framework. In the parlance of

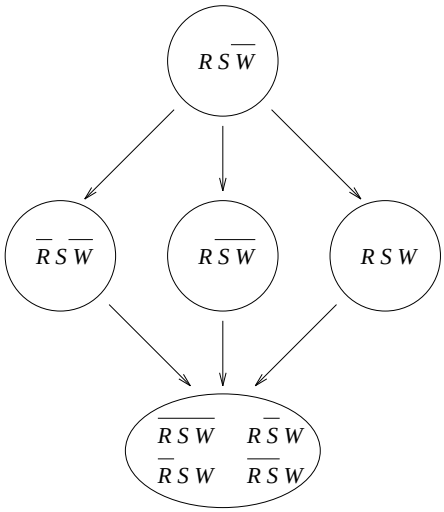

Figure 4: Theorist Model for the Wet Grass Example

Theorist, preferred explanations ought to be those that occur at the most plausible situations, or those that are consistent with as many defaults as possible. However, explanatory conjectures alone do not have the predictive force required — the facts $\mathcal{F}$ must also be considered.

Definition 4.9 Let $C, C' \subseteq \mathcal{C}$ be predictive Theorist explanations for β, relative to $\langle \mathcal{F}, \mathcal{D} \rangle$. C is *at least as preferred as* C' (written $C \leq_{\mathcal{F}} C'$) iff each maximal subset of defaults $D' \subseteq \mathcal{D}$ consistent with $C' \cup \mathcal{F}$ is contained in some subset of defaults $D \subseteq \mathcal{D}$ consistent with $C \cup \mathcal{F}$. Explanation C is a *preferred explanation* iff there is no explanation for β such that $C' <_{\mathcal{F}} C$.

In our example, R and S are equally preferred explanations since both are consistent with the entire set of defaults $\mathcal{D}$. The explanation $R \wedge S$ is less preferred because it conflicts with the default $R \supset \neg S$.

It is possible, due to the fact that the plausibility relation determined by $M_{\mathcal{D}}$ is not total, that two explanations are incomparable. If asked to explain $(R \vee S) \wedge \neg W$, predictive explanations $R \wedge \neg W$ and $S \wedge \neg W$ are preferred to the explanation $R \wedge S \wedge \neg W$. Yet these two preferred explanations are incomparable in the Theorist model. This notion of preference corresponds naturally to the plausibility ordering determined by $M_{\mathcal{D}}$.

Theorem 4.10 *Let $C, C' \subseteq \mathcal{C}$ be predictive Theorist explanations for β, relative to $\langle \mathcal{F}, \mathcal{D} \rangle$. Then $C \leq_{\mathcal{F}} C'$ iff $M_{\mathcal{D}} \models \vec{\Box}((C' \wedge \mathcal{F}) \supset \Diamond(C \wedge \mathcal{F}))$.*

Notice that the comparison of plausibility can be applied to nonpredictive explanations as well. We will see this in Section 5.

4.3 Capturing Preferred Subtheories in CT4O

The manner in which Theorist is embedded in our abductive framework also applies to Brewka's preferred subtheories. For a Brewka theory $\langle \mathcal{F}, \langle \mathcal{D}_1, \cdots, \mathcal{D}_n \rangle \rangle$ the plausibility of worlds is not determined solely by the number of rules violated, but also the priority of those rules. Implicit in the definition of an extension is the idea that any number of rules of lower priority may be violated if it allows a rule of higher priority to be satisfied. This gives rise to a new definition of rule violation.

Definition 4.10 For any possible world $w \in W$, the set of defaults of rank i *violated* by w is

$$V_i(w) = \{d \in \mathcal{D}_i : w \models \neg d\}$$

A world that violates fewer high priority defaults than another world should be considered more plausible, even if the second world violates fewer low priority defaults. This gives rise to the *Brewka structure* for an ordered set of defaults $\langle \mathcal{D}_1, \cdots, \mathcal{D}_n \rangle$.

Definition 4.11 Let $\langle \mathcal{D}_1, \cdots, \mathcal{D}_n \rangle$ be an ordered set of defaults, and let v, w be possible worlds. The minimal rank at which w and v *differ* is

$$\mathit{diff}(w, v) = \mathit{min}\{i : V_i(w) \neq V_i(v)\}$$

If $V_i(w) = V_i(v)$ for all $i \leq n$, by convention we let $\mathit{diff}(w, v) = n + 1$.

Thus, $\mathit{diff}(w, v)$ denotes the highest priority partition of default rules $\mathcal{D}_{diff(w,v)}$ within which w and v violate different rules. It is this set of rules that determines which of w or v is more plausible.

Definition 4.12 The *Brewka structure* for $\langle \mathcal{D}_1, \cdots, \mathcal{D}_n \rangle$ is $M_{\mathcal{B}} = \langle W, \leq, \varphi \rangle$ where W is the set of truth assignments suitable for $\mathbf{L}_{CPL}$; φ is the valuation function induced by W; and $v \leq w$ iff $V_{diff(w,v)}(v) \subseteq V_{diff(w,v)}(w)$.

Proposition 4.11 *$M_{\mathcal{B}}$ is a CT4O*-model.*

Let us denote by $\mathcal{D}$ the set $\mathcal{D}_1 \cup \ldots \cup \mathcal{D}_n$. The model $M_{\mathcal{B}}$, just as the Theorist model $M_{\mathcal{D}}$, divides worlds into clusters of equally plausible worlds that violate exactly the same set of defaults in $\mathcal{D}$.

Proposition 4.12 *$\mathcal{C}$ is a cluster in the model $M_{\mathcal{B}}$ iff for some $D \subseteq \mathcal{D}$*

$$\mathcal{C} = \{w : w \models d \textit{ if } d \in D \quad \textit{and} \quad w \models \neg d \textit{ if } d \in \mathcal{D} - D\}$$

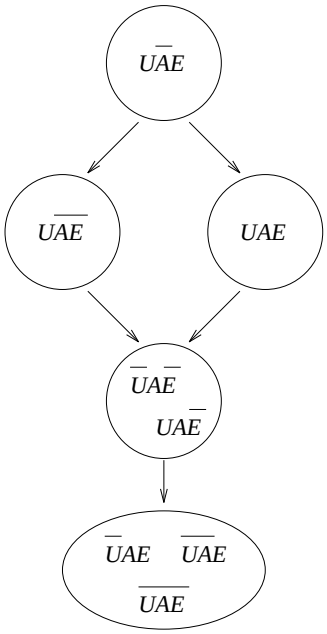

Figure 5: Brewka Model for the University Students Example

However, the ordering of clusters is determined differently. In the Theorist model, only set inclusion is used to determine relative plausibility. In contrast, the Brewka model may rank a world v more plausible than a world w, even if $V(v) \not\subseteq V(w)$. In particular, we may have that v violates a low priority rule that is satisfied by w. Figure 5 depicts the Brewka model for the default sets $\mathcal{D}_1 = \{U \supset A, U \supset \neg E\}$ and $\mathcal{D}_2 = \{A \supset E\}$. In contrast with the Theorist model for the "flat" version of this theory (see Figure 3), we see that worlds violating the rule $A \supset E$ are more plausible than worlds violating either of the other two rules (individually).

The notions of prediction and explanation in Brewka's framework correspond to our conditional models of prediction and explanation, allowing results to be shown that are entirely analogous to those demonstrated above for Theorist. We omit proofs of the following results; they can be verified in a straightforward way by extending the proofs of the corresponding results for Theorist to accommodate the more refined ordering of clusters provided in Definition 4.12.

Proposition 4.13 *E is an extension of $\langle \mathcal{F}, \langle \mathcal{D}_1, \cdots, \mathcal{D}_n \rangle \rangle$ iff $\|E\| = S$ for some $S \in Pl(\mathcal{F})$ in the model $M_{\mathcal{B}}$.*

Corollary 4.14 *A is in some extension of $\langle \mathcal{F}, \langle \mathcal{D}_1, \cdots, \mathcal{D}_n \rangle \rangle$ iff $M_{\mathcal{B}} \models \mathcal{F} \rightarrow A$.*

Theorem 4.15 *A is predicted from Brewka theory $\langle \mathcal{F}, \langle \mathcal{D}_1, \cdots, \mathcal{D}_n \rangle \rangle$ iff $M_{\mathcal{B}} \models \mathcal{F} \Rightarrow A$.*

Assuming some set of conjectures $\mathcal{C}$, we have

Theorem 4.16 *Let $C \subseteq \mathcal{C}$. Then C is a Theorist explanation for β relative to the Brewka theory $\langle \mathcal{F}, \langle \mathcal{D}_1, \cdots, \mathcal{D}_n \rangle \rangle$ iff $M_{\mathcal{B}} \models (\mathcal{F} \wedge C) \rightarrow \beta$ and $\mathcal{F} \wedge C$ is consistent.*

We define predictive explanations for a Brewka theory in the same fashion as for Theorist.

Definition 4.13 Let $C \subseteq \mathcal{C}$ and β be some observation. C is a *predictive explanation* for β iff $\beta \in E$ for all extensions E of $\langle \mathcal{F} \cup C, \langle \mathcal{D}_1, \cdots, \mathcal{D}_n \rangle \rangle$.

Intuitively, an observation is *predictively explained* by some conjectures if, for every "maximal" set of defaults consistent with C and $\mathcal{F}$, the observation is entailed by the facts $\mathcal{F}$ and the conjectures C, together with these defaults. However, Brewka explanations rely on a definition of "maximality" that includes the consideration of priority of default rules.

Theorem 4.17 *Let $C \subseteq \mathcal{C}$. Then C is a predictive Brewka explanation for β iff $M_{\mathcal{B}} \models (\mathcal{F} \wedge C) \Rightarrow \beta$ and $\mathcal{F} \wedge C$ is consistent.*

Finally, preferences on explanations are also defined in the same manner, but again taking priorities into account.

Definition 4.14 Let $C, C' \subseteq \mathcal{C}$ be predictive Brewka explanations for β, relative to the Brewka theory $\langle \mathcal{F}, \langle \mathcal{D}_1, \cdots, \mathcal{D}_n \rangle \rangle$. We call the set

$$\bigcup_{k \leq n} \{D_k : D_k \subseteq \mathcal{D}_k\}$$

a *maximal set of defaults* for C iff $\mathcal{F} \cup C \cup D_1 \cup \ldots \cup D_k$ is a maximal consistent subset of $\mathcal{F} \cup C \cup \mathcal{D}_1 \cup \ldots \cup \mathcal{D}_k$, for each $1 \leq k \leq n$. C is *at least as preferred as* C' (written $C \leq_{\mathcal{F}} C'$) iff for each maximal set of defaults $\cup_{k \leq n} \{D'_k\}$ for C' there is a maximal set $\cup_{k \leq n} \{D_k\}$ for C such that $D'_k \subseteq D_k$ for each $1 \leq k \leq n$.

Theorem 4.18 *Let $C, C' \subseteq \mathcal{C}$ be predictive Brewka explanations for β, relative to the Brewka theory $\langle \mathcal{F}, \langle \mathcal{D}_1, \cdots, \mathcal{D}_n \rangle \rangle$. Then $C \leq_{\mathcal{F}} C'$ iff $M_{\mathcal{B}} \models \overleftrightarrow{\Box}((C' \wedge \mathcal{F}) \supset \Diamond(C \wedge \mathcal{F}))$.*

If we compare the Brewka model $M_{\mathcal{B}}$ in Figure 5 with the Theorist model for the same (unprioritized) set of defaults $M_{\mathcal{D}}$ in Figure 3, the differences in structure induced by priorities become clear. In a sense, the Brewka model has increased "connectivity". While worlds that are comparable

in the Theorist model remains so in $M_{\mathcal{B}}$, certain clusters of worlds that are incomparable become comparable in $M_{\mathcal{B}}$. This leads, for instance, to the fact that U predicts $\neg E$ in $M_{\mathcal{B}}$, but does not in $M_{\mathcal{D}}$. This increased connectivity is, in fact, necessarily the case.

Proposition 4.19 *Let $M_{\mathcal{B}}$ be the Brewka model for $\langle \mathcal{D}_1, \cdots, \mathcal{D}_n \rangle$ and $M_{\mathcal{D}}$ the Theorist model for its reduction $\mathcal{D} = \mathcal{D}_1 \cup \ldots \cup \mathcal{D}_n$. Then $v \leq w$ in $M_{\mathcal{B}}$ whenever $v \leq w$ in $M_{\mathcal{D}}$.*

Theorem 4.20 *If $M_{\mathcal{D}} \models \alpha \Rightarrow \beta$ then $M_{\mathcal{B}} \models \alpha \Rightarrow \beta$.*

Thus any predictive explanation in Theorist is also a predictive explanation when *any* set of priorities is introduced.

Intuitively, one would expect priorities to prune away possible explanations. For predictive explanations, the opposite may occur, since priorities can only increase the number of predictions admitted by a set of facts and conjectures. However, if we consider only *preferred* predictive explanations, we have more reasonable behavior. It becomes clear that priorities may, in fact, reduce the number of preferred explanations (and it cannot increase the number).

We note that the representation of Theorist and Brewka models for a given set of defaults does not require that one specify the ordering relation for the model explicitly for each pair of worlds. One may axiomatize the model (relatively) concisely using techniques described in [3]. The truth of conditionals determining explanations and preferences can then be tested against this theory. However, we are not suggesting that our conditional framework be used as a computational basis for explanations in simple Theorist-like theories. Rather, it brings to light the underlying semantic properties of Theorist and several principled extensions.

5 Consistency-Based Diagnosis

While the Theorist system may be used for diagnosis (as our examples in this section illustrate), it is presented more generally as a method for effecting arbitrary explanations. Another approach to model-based diagnosis is consistency-based diagnosis, which is aimed more directly at the diagnostic task, namely to determine why a correctly designed system is not functioning according to its specification. In this section, after presenting the fundamental concepts from Reiter's [49] and de Kleer, Mackworth and Reiter's [16] methodology for diagnosis, we show how these canonical consistency-based models can be embedded in our framework for epistemic explanations. This highlights many of the key similarities and differences in the abductive and consistency-based approaches. We also address the role fault models play within our semantics and how diagnoses can be made predictive.

5.1 A Logical Specification

de Kleer, Mackworth and Reiter [49, 16] assume that an appropriate model of a system or artifact consists of two parts. The first is a set of *components COMP*, the parts of a system that one is able to distinguish and that (more or less) independently can fail to function correctly. The second is a set of logical sentences *SD*, the *system description*, that describes precisely the intended or normal functioning of the system. For example, given a certain set of inputs to a circuit, the system description should allow one to predict the value of the outputs. Because certain components may fail, a system description that only allows for correct behavior will be inconsistent with observations of incorrect behavior. Therefore, *abnormality predicates* are introduced. For any component $c \in$ *COMP*, the literal *ab*(c) denotes the fact that component c is not functioning as required. Such a component is said to be abnormal; otherwise it is normal. We assume that components usually function correctly. However, because expected observations depend on this assumption, the system description will usually contain sentences in which anticipated behavior is explicitly predicated on this assumption. Thus sentences such as $\neg$*ab*$(c_i) \supset \alpha$ assert that, if component c_i is functioning correctly then behavior α will be observed. The correct functioning of a system is then more accurately characterized by the set of sentences

$$\mathit{CORRECT} = \mathit{SD} \cup \{\neg \mathit{ab}(c_i) : c_i \in \mathit{COMP}\}$$

Throughout we assume that this set *CORRECT* is consistent.

If an observation is obtained that is inconsistent with *CORRECT* then (assuming that both the observation and system description are accurate and correct), it must be that some of the components have failed; that is, *ab*(c_i) must hold for some members $c_i \in$ *COMP*.[17] A *diagnosis* for such an observation is any set of components whose abnormality (alone) makes the observation consistent with *SD*. More precisely, following [16], we have these definitions.[18]

Definition 5.1 Let $\Delta \subseteq$ *COMP* be a set of components. Define sentence $D(\Delta)$ to be

$$\bigwedge [\{\mathit{ab}(c) : c \in \Delta\} \cup \{\neg \mathit{ab}(c) : c \in \mathit{COMP} - \Delta\}]$$

$D(\Delta)$ expresses the fact that the components in Δ are functioning improperly while all other components are functioning correctly.

[17] We will make a few remarks at the conclusion of this section regarding the possibility that *SD* is an incorrect model.

[18] As usual, a "set" of observations will be assumed to be finite and conjoined into a single sentence β. For any set of sentences, such as *SD*, we will assume finiteness, and treat the set somewhat loosely as the conjunction of its elements. Context should make clear whether the sentence or the set is intended.

Definition 5.2 Let $\Delta \subseteq COMP$. A *consistency-based diagnosis* (*CB-diagnosis* for short) for observation β is any $D(\Delta)$ such that $SD \cup \{\beta, D(\Delta)\}$ is satisfiable.

Reiter's [49] "Principle of Parsimony" suggests that reasonable diagnoses are those that require as few faults as possible to explain the aberrant behavior. A *minimal diagnosis* is any diagnosis $D(\Delta)$ such that for no proper subset $\Delta' \subset \Delta$ is $D(\Delta')$ a diagnosis. In Reiter's original formulation, only minimal diagnoses are deemed essential. If the correct functioning of a system is all that is modeled in *SD*, then one can show, for any diagnosis $D(\Delta)$, that a larger component set $\Delta \subseteq \Delta'$ also determines a diagnosis $D(\Delta')$. Thus, minimal diagnoses characterize the set of all diagnoses.

Example 5.1 Imagine a simple system with two components, a plug and a light bulb. One can observe that the bulb is bright, dim or dark. *SD* captures the correct behavior of the system:

$$\neg ab(bulb) \wedge \neg ab(plug) \supset bright$$

We assume that the three possible observations are exhaustive and mutually exclusive (and that this fact is captured in *SD* as well). We expect to see a bright light (i.e., $bright$ is true), since this is entailed by *CORRECT*:

$$SD \cup \{\neg ab(bulb), \neg ab(plug)\} \models bright$$

If we observe dim, then the minimal diagnoses are $D(\{bulb\})$ and $D(\{plug\})$. The nonminimal diagnosis $D(\{bulb, plug\})$ also renders the observation dim consistent. Notice that each of these diagnoses applies to the observation $bright$ as well, even though this is the system's predicted behavior. That is, the diagnoses do not rule out the "correct" behavior. ■

The presence of *fault models* renders Reiter's characterization incorrect.[19] de Kleer, Mackworth and Reiter suggest a notion of *kernel* diagnosis that can be used to replace minimal diagnosis in the characterization of all diagnoses. Our goal here is not to investigate such characterizations, but rather investigate the semantics of diagnosis as explanation. Despite the failure of minimal diagnoses in this characterization task, the principle of parsimony (in the absence of more refined, say, probabilistic information) suggests that minimal diagnoses are to be preferred. We will simply point out the impact of fault models on diagnosis.

Intuitively, a fault model is a portion of the system description that allows predictions to be made when it is known or assumed that some component is faulty. In the example above, one cannot predict

[19] Similar remarks apply to *exoneration axioms*, which we do not discuss here.

anything about the brightness of the light if one of the components is abnormal. All observations are possible (consistent). Suppose we add the following axiom:

$$ab(bulb) \wedge ab(plug) \supset dark$$

While $D(\{bulb\})$ and $D(\{plug\})$ are both diagnoses for dim, the "larger" diagnosis $D(\{bulb, plug\})$ is not. Thus, in the presence of fault models, supersets of diagnoses need not themselves be diagnoses. de Kleer, Mackworth and Reiter do, however, formulate conditions under which this is guaranteed to be the case.

5.2 Capturing Consistency-Based Models in CT4O

Just as with our embedding of Theorist, we can provide a CT4O-model that captures the underlying intuitions of consistency-based diagnosis. We assume that the language in which the system description and observations are phrased is propositional, denoted $\mathbf{L}_{CPL}$. We will assume that for each component in *COMP* there is a proposition stating that the component has failed. We will, however, continue to use the first-order notation $ab(c)$ for such a proposition.[20]

The principle of parsimony carries with it the implicit assumption that situations in which fewer system components are abnormal are more plausible than those with more components failing. This suggests a natural ordering of plausibility on possible worlds.

Definition 5.3 Let w be a possible world suitable for $\mathbf{L}_{CPL}$ and *COMP* some set of system components. The *abnormality set* for w is the set

$$Ab(w) = \{c \in COMP : w \models ab(c)\}$$

Definition 5.4 The *consistency-based model* (the *CB-model*) for component set *COMP* is $M_{COMP} = \langle W, \leq, \varphi \rangle$ where W is the set of truth assignments suitable for $\mathbf{L}_{CPL}$; φ is the valuation function induced by W; and $v \leq w$ iff $Ab(v) \subseteq Ab(w)$.

Proposition 5.1 *M_{COMP} is a CT4O*-model.*

Notice that the CB-model for a set of components is exactly the Theorist model with the set of defaults

$$\mathcal{D} = \{\neg ab(c) : c \in COMP\}$$

[20] A first-order diagnostic model can be captured propositionally by using ground terms should the domain of components and other objects of interest be finite. A first-order version of our logics could be used but this is not relevant to our concerns here.

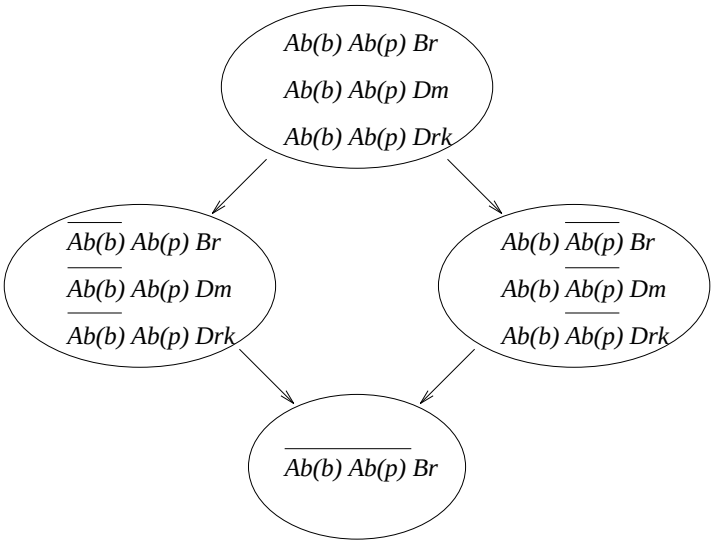

Figure 6: The CB-model for a Two Component System

We will exploit this fact below when comparing consistency-based and abductive diagnosis.

The model M_{COMP} does not rule out worlds violating *SD*. *SD*, much like $\mathcal{F}$ above, will be used explicitly in defining diagnoses. Worlds in which $\neg$*SD* holds will not play a role in consistency-based diagnosis; therefore, we could easily use a CT4O-model in which only *SD*-worlds are represented (e.g., using the axiom $\overleftrightarrow{\Box}$*SD*).[21]

Example 5.2 Figure 6 illustrates the model M_{COMP} for our simple light bulb example with two components. For simplicity, we show only those worlds that satisfy the system description *SD* provided in Example 5.1. As usual, worlds in the same cluster are those in which the same components have failed or work correctly. ■

The most plausible state of affairs in the model M_{COMP} is simply the set of worlds satisfying the theory *CORRECT*. Should an observation be made that conflicts with this theory, the system must be functioning abnormally and belief in the assumption $\neg ab(c)$ for at least one $c \in C$ must be retracted. A diagnosis is an explanation, given in terms of normal and abnormal components, for such an observation. Clearly a CB-diagnosis is not predictive, for it simply must ensure that the observation is rendered plausible. In Example 5.1, the sentences $D(\{bulb\})$ and $D(\{plug\})$ are both diagnoses

[21] However, one could imagine the diagnostic process including the debugging of *SD*, as takes place for instance in model verification, or even scientific theory formation.

of the observation dim. But neither of these diagnoses entails the observation dim. This leads to the notion of an *excuse*, which is simply a *might explanation*, as described in Section 3, consisting of possible component failures.

Definition 5.5 Let $\Delta \subseteq$ *COMP* be a set of components. Define sentence $AB(\Delta)$ to be

$$\bigwedge [\{ab(c) : c \in \Delta\}]$$

Thus, $AB(\Delta)$ asserts that all components in Δ are functioning abnormally. In contrast to the sentence $D(\Delta)$, $AB(\Delta)$ asserts nothing about the status of components not in Δ.

Definition 5.6 Let *COMP* and *SD* describe some system. An *excuse* for an observation β is any sentence $AB(\Delta)$ (where $\Delta \subseteq$ *COMP*) such that

$$M_{COMP} \models AB(\Delta) \wedge SD \not\Rightarrow \neg\beta$$

If belief in the excuse were adopted, the observation would not be disbelieved. For instance, the model M_{COMP} admits excuses $D(\{bulb\})$, $D(\{plug\})$ and $D(\{bulb, plug\})$ for the observation dim. Notice that $D(\emptyset)$ (which we assume to be $\top$) is not an excuse for dim since the belief $bright$ precludes it; that is, the conditional $\top \wedge SD \not\Rightarrow \neg dim$ is false.

Because of the ordering of plausibility built in to the CB-model, when a certain set of components is believed to have failed, other components are assumed to still be functioning correctly.

Proposition 5.2 $M_{COMP} \models AB(\Delta) \Rightarrow D(\Delta)$

This proposition ensures that a diagnosis in the CT4O framework (i.e., an excuse) can be given solely in terms of failing components. Thus, we have that an excuse determines a CB-diagnosis for an observation.

Theorem 5.3 *Let SD and COMP determine some system.* $D(\Delta)$ *is a CB-diagnosis for observation* β *iff* $AB(\Delta)$ *is an excuse for* β *relative to* M_{COMP}.

Naturally, we should not accept any might explanation for an observation as a reasonable diagnosis. Preferred diagnoses should be those that are most plausible, and the ordering of plausibility determined by the model M_{COMP} can be used for this purpose. Unsurprisingly, preferred diagnoses are precisely those that minimize the number of abnormal components.

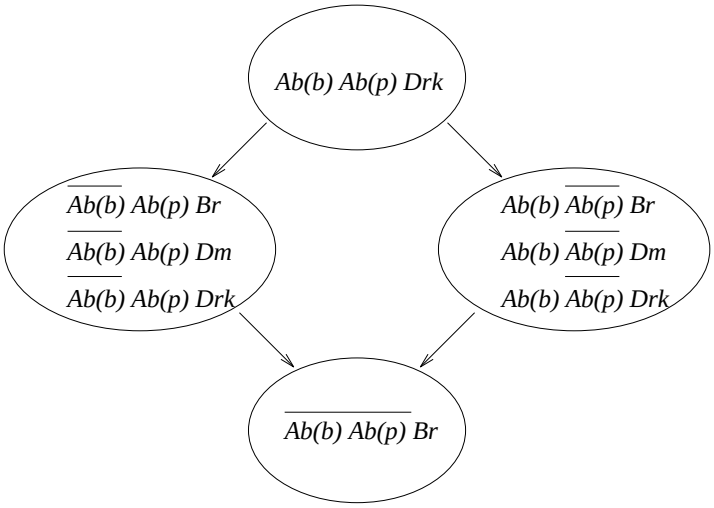

Figure 7: The Addition of a Fault Axiom

Definition 5.7 Let $D(\Delta)$ and $D(\Delta')$ be CB-diagnoses for observation β. $D(\Delta)$ is *at least as preferred as* $D(\Delta')$ (written $D(\Delta) \leq_{COMP} D(\Delta')$) iff

$$M_{COMP} \models \overleftrightarrow{\Box}(D(\Delta') \supset \Diamond D(\Delta))$$

$D(\Delta)$ is *preferred diagnosis* iff there is no diagnosis for β such that $D(\Delta') \leq_{COMP} D(\Delta)$.

Theorem 5.4 *$D(\Delta)$ is a preferred diagnosis iff $D(\Delta)$ is a minimal diagnosis.*

5.3 Predictive Diagnoses and Fault Models

Consider the light bulb example above with the additional axiom

$$ab(bulb) \wedge ab(plug) \supset dark$$

incorporated into the system description *SD*. Figure 7 illustrates the model M_{COMP} for this new system SD_F. This additional axiom will be dubbed a *fault axiom* or a partial *fault model*. If all axioms have a "positive form" (i.e., describing behavior based only on conditions of normality), then diagnoses (or assumptions of abnormality) can never be used to predict aberrant behavior. In other words, all "observations" are consistent with each (nonempty) diagnosis. Fault models change the nature of diagnosis by making it a more "nonmonotonic process." For instance, without this fault axiom, the two

excuses $AB(\{bulb\})$ and $AB(\{plug\})$ determine diagnoses for the observation dim, as does the larger diagnosis $AB(\{bulb, plug\})$. This forms the basis for Reiter's [49] characterization of all diagnoses in terms of minimal diagnoses. However, with the fault axiom, the sentence $AB(\{bulb, plug\})$ is not an excuse for dim:

$$M_{COMP} \models AB(\{bulb, plug\}) \wedge SD_F \Rightarrow \neg dim$$

This reflects the observation of de Kleer, Mackworth and Reiter that supersets of diagnoses need not be diagnoses themselves. In our terminology:

Proposition 5.5 *If $AB(\Delta)$ is an excuse for observation, $AB(\Delta')$ need not be, where $\Delta \subseteq \Delta'$.*

Fault models have another impact on the nature of diagnosis. Consider the observation $dark$. One diagnosis for this observation (relative to SD_F) is the excuse $AB(\{bulb, plug\})$:

$$M_{COMP} \models AB(\{bulb, plug\}) \wedge SD_F \not\Rightarrow \neg dark$$

Without fault axioms (i.e., using SD rather than SD_F), such an excuse renders the observation plausible, but does not preclude other observations. However, with the fault axiom we have an even stronger *predictive condition*:

$$M_{COMP} \models AB(\{bulb, plug\}) \wedge SD_F \Rightarrow dark$$

Not only does the diagnosis render $dark$ plausible, it also induces *belief* in the observation $dark$.

Naturally, one might extend the definition of a diagnosis by requiring not only that the observation be rendered consistent, but also that it be entailed by the diagnosis. Such diagnoses will be dubbed *predictive diagnoses*.

Definition 5.8 Let $\Delta \subseteq COMP$. A *predictive diagnosis* for observation β is any $D(\Delta)$ such that $SD \cup \{D(\Delta)\} \models \beta$ and $SD \cup \{D(\Delta)\}$ is consistent.

Theorem 5.6 *Assume $D(\Delta) \wedge SD$ is consistent. $D(\Delta)$ is a predictive diagnosis for β iff*

$$M_{COMP} \models AB(\Delta) \wedge SD \Rightarrow \beta$$

Predictive diagnoses are predictive explanations rather than might explanations, and as such carry many of the conceptual advantages of predictive explanations. Unfortunately, for most systems, one cannot expect diagnoses to be predictive in most circumstances. Typically, the knowledge of how a system fails is incomplete. One may know that a weak battery causes an LED display to show "strange" readings, but the *specific* observed display in such a circumstance is not usually predicted

by a diagnosis. However, with partial fault models one will have that certain diagnoses predict the observations they explain, rather than just excusing them.

If one has a complete fault model incorporated into *SD*, intuitively *every* diagnosis carries with it a prediction about the behavior that can be observed. Thus, one would expect that every CB-diagnosis, in the process of excusing the observation, would actually predict it. This leads to general circumstances under which every CB-diagnosis of an observation for a particular system is a predictive diagnosis. We assume that the system's behavior can be characterized by a given set of possible observations $\mathcal{O}$, the elements of which must be mutually exclusive and exhaustive (relative to *SD*).[22] We say that *SD* contains a *complete model of correct behavior* iff there exists a $\beta \in \mathcal{O}$ s.t.

$$\mathit{CORRECT} \models \mathcal{O}_\beta$$

where

$$\mathcal{O}_\beta = \beta \wedge \bigwedge\{\neg\gamma : \gamma \in \mathcal{O} - \{\beta\}\}$$

We say that *SD* contains a *complete fault model* iff for each diagnosis $D(\Delta)$ there is a β such that

$$D(\Delta) \wedge \mathit{SD} \models \beta \wedge \mathcal{O}_\beta$$

Notice that a complete fault model, on this definition, ensures that one has a complete model of correct behavior (simply set $\Delta = \emptyset$). If required, we could restrict Δ to nonempty sets of components, thus decoupling the model of faulty behavior from that of correct behavior.

If *SD* contains a complete model of correct behavior and a complete fault model, it is easy to see that each consistency-based diagnosis will be predictive. Consider our light bulb example once again, with observable behaviors $bright$, dim and $dark$ and the following axioms in *SD* (the first models correct behavior, the second and third are fault axioms):

$$(\neg \mathit{ab}(bulb) \wedge \neg \mathit{ab}(plug)) \supset bright$$
$$(\mathit{ab}(bulb) \equiv \neg \mathit{ab}(plug)) \supset dim$$
$$(\mathit{ab}(bulb) \wedge \mathit{ab}(plug)) \supset dark$$

[22] One may expect a number of possible observations of correct behavior, for instance, corresponding to the possible inputs to a circuit. However, we treat this as a single observation, the form of which will typically be a conjunction of implications or biconditionals. The antecedents will determine certain inputs and the consequents certain outputs (e.g., $on \wedge \neg ab(bulb) \supset bright$). Similar remarks apply to incorrect behavior. This is not the main point of our description so we do not pursue this issue further.

Clearly, any excuse we can make for a given observation will also predict that observation. In this example, every CB-diagnosis is a predictive diagnosis.

Proposition 5.7 *If SD includes a complete fault model then $D(\Delta)$ is a CB-diagnosis for β iff $D(\Delta)$ is a predictive diagnosis for β.*

Notice that to diagnose faulty behavior only, a model of correct behavior is not required — a complete fault model ensures that predictive explanations can be given for every "abnormal" observation. However, without any indication of correct behavior any observation is consistent with the assumption that all components work correctly. Thus, a complete model of correct behavior is required if CB-diagnoses are to be of any use. This is in accordance with the observation of Poole [45] who describes the categories of information required for consistency-based diagnosis and abductive diagnosis. Console and Torasso [13] have also addressed this issue. They suggest, as we have elaborated above, that consistency-based diagnosis is appropriate if fault models are lacking, while abductive approaches are more suitable if models of correct behavior are incomplete.

It is important to notice that the definition of complete fault model above relies crucially on the set of propositions one is allowed to explain, in other words, the set of "observables." For example, suppose we had only a single fault axiom:

$$(\mathit{ab}(bulb) \vee \mathit{ab}(plug)) \supset (dark \vee dim)$$

This fault model is incomplete relative to the original set of observables, for no CB-diagnosis for dim actually *predicts* dim. Each diagnosis, $D(\{bulb\})$, $D(\{plug\})$ and $D(\{bulb, plug\})$, allows the possibility of observation $dark$. However, suppose we "coalesce" the observations dim and $dark$ into a single category $notBright \equiv dim \vee dark$. If the observations a system is allowed to explain are restricted to $bright$ and $notBright$, this fault model is complete; any CB-diagnosis will then predict its observation. In this example, $D(\emptyset)$ predicts $bright$, while the other three diagnoses predict $notBright$. If users are allowed to make more refined observations, predictive diagnoses can be given if observations are mapped into coarse-grained explainable propositions.

5.4 On the Relationship to Abductive Diagnosis

Let us assume that we have a Theorist default theory for the diagnosis of a system where *SD* is taken to be the set of facts and the default set is

$$\mathcal{D} = \{\neg \mathit{ab}(c) : c \in \mathit{COMP}\}$$

As observed above, the Theorist model for such a theory is precisely the CB-model for this system. If we restrict Theorist explanations to those of the form used for consistency-based diagnosis, some interesting relationships emerge.

Suppose that Theorist explanations are restricted to have the form $AB(\Delta)$ (or $D(\Delta)$). We will call these explanations *Theorist diagnoses*. Such weak explanations are then guaranteed to be predictive. This is due simply to the fact that the most plausible worlds at which such an explanation holds must lie within a single cluster. In other words, Theorist diagnoses have a single extension.

Proposition 5.8 *Let* $\Delta \subseteq$ *COMP.* $M_{COMP} \models AB(\Delta) \rightarrow \beta$ *iff* $M_{COMP} \models AB(\Delta) \Rightarrow \beta$. *(Similarly for* $D(\Delta)$*.)*

Should we model a system in Theorist as we do for consistency-based diagnosis, then Theorist diagnoses are exactly *predictive diagnoses* as we have defined in the consistency-based framework.

As we have seen, many (if not most) observations cannot be predicted in the consistency-based framework, especially if fault-models are lacking or incomplete. This indicates that the abductive approach to diagnosis requires information of a form different from that used in the consistency-based approach. This is emphasized by Poole [45]. However, given complete fault models, Theorist diagnoses and consistency-based diagnoses will coincide. Konolige [31] has also examined the relationship between the two forms of diagnosis.

Without complete information, the Theorist system, in particular the notion of an extension, can still be used to effect consistency-based diagnosis. While a CB-diagnosis may not predict an observation, it does require that the observation is consistent with all other "predictions." In Theorist terms, the observation is consistent with the (single) extension of the diagnosis. In other words, these are might explanations in the Theorist model.

Theorem 5.9 *Let SD and COMP describe some system,* $\Delta \subseteq$ *COMP, and* $\mathcal{D}$ *be the set of defaults* $\{\neg ab(c) : c \in$ *COMP*$\}$. *Then* $D(\Delta)$ *is a CB-diagnosis for observation* β *iff* $\neg\beta \notin E$ *where* E *is the (only) Theorist extension of* $\langle$*SD* $\cup \{AB(\Delta)\}, \mathcal{D}\rangle$.

Corollary 5.10 $D(\Delta)$ *is a CB-diagnosis for observation* β *iff* $M_{\mathcal{D}} \models$ *SD* $\wedge AB(\Delta) \not\Rightarrow \neg\beta$.

Thus, consistency-based diagnosis can be captured in the Theorist abductive framework without requiring that the form of the system description be altered. *SD* is simply used as the set of facts $\mathcal{F}$. Poole [45] also defines a form of consistency-based diagnosis within Theorist. He shows that $AB(\Delta)$ is a "consistency-based diagnosis" iff $D(\Delta)$ is in some extension of *SD* $\cup \{\beta\}$. Our notion of consistency-based diagnosis in Theorist does not rely on forming extensions of the observation, but (more in the true spirit of abduction) examines extensions and predictions of the explanation itself.

This is important because our definition captures *all* CB-diagnoses. Poole's definition is based on Reiter's [49] definition of diagnosis in terms of minimal sets of abnormal components. It is not hard to see that, in fact, $D(\Delta)$ is in some extension of $SD \cup \{\beta\}$ iff $D(\Delta)$ is a *minimal* CB-diagnosis. While Poole's observation is correct for minimal diagnoses (and Reiter's formulation, in particular), it cannot be extended to the more general case subsequently developed by de Kleer, Mackworth and Reiter.

Console and Torasso [13] have also explored the distinction between abductive and consistency-based diagnosis and present a definition of explanation (in the style of Reiter) that combines both types. The set of observations to be explained are divided into two classes: those which must be predicted by an explanation and those which must simply be rendered consistent by the explanation. We can, of course, capture such explanations conditionally by using both predictive and weak explanations. Roughly, if β is the part of the observation that needs to be predicted and γ is the component that must be consistent with the explanation (and background theory) then we simply require that any explanation α be such that $\alpha \Rightarrow \beta$ and $\alpha \not\Rightarrow \neg\gamma$.

6 Concluding Remarks

We have presented some general conditions on epistemic explanations, describing a number of different types of explanations, and why certain explanations are to be preferred to others. Our account relies heavily on a model of belief revision and conditional sentences. The defeasible nature of explanations and preferences for plausible explanations are induced naturally by the properties of our revision model. We have also shown how the two main paradigms for model-based diagnosis can be embedded in our conditional framework.

A number of avenues remain to be explored. We are currently investigating how our model might be extended to incorporate causal explanations. Such explanations, especially in diagnostic and planning tasks, are of particular interest. Grafting a representation of causal influences onto our model of explanation, such as that of Goldszmidt and Pearl [27], seems like a promising way in which to (qualitatively) capture causal explanations. Konolige [32] has explored the use of causal theories in diagnosis as a means to obviate the need for fault models. His representation in terms of *default causal nets* allows both explanations and excuses; but the causal component of his representation remains essentially unanalyzed. The key features of Konolige's theories can be captured in our framework in a rather straightforward way. These include exemptions of "faults," distinguishing normality conditions from primitive causes and preferences for *normal* and *ideal* explanations. This is due to the flexibility of the conditional logic and the generality of plausibility orderings. We also hope to explore the issue

of designing tests to discriminate potential diagnoses, and the trade-off between further testing and repair. This is an issue that has recently attracted much attention [19, 39].

The pragmatics of explanation remains an important avenue to pursue. Ways in which to rule out weak or strong explanations, depending on context must be addressed. Another pragmatic concern has to do with the *elaboration* of explanations. We have assumed that explanations are given relative to background theory. If an explanation is questioned, or elaboration is requested, this may be due to the fact that certain background is not shared between the abductive system and the user requesting the explanation. Mechanisms with which the appropriate background knowledge can be determined, and offered as elaboration, would be of crucial interest. The manner in which an explanation is requested by a user can also provide clues as to what form an explanation should take [58].

Other forms of explanation cannot be captured in our framework, at least in its current formulation. An important type of explanation is of the form addressed by the theory of Gärdenfors [22]. There an explanation is simply required to render an observation more plausible than it was before the explanation was adopted. As an example, consider possible explanations for Fred's having developed AIDS (A). A possible (even reasonable) explanation is that Fred practiced "unsafe" sex (U). However, it would seem that adopting the belief U is not sufficient to induce the belief that Fred contracted HIV and developed AIDS. Furthermore, if the probability is low enough, this might not even be a valid *might explanation*; that is, $U \Rightarrow \neg A$. However, U does increase the likelihood of A (even if not enough to render A believable, or even epistemically possible). Such explanations might be captured by comparing the relative plausibility of A given U and A alone, without appeal to probabilities. Such an example may suggest a role for decision-theoretic versions of conditional defaults. While A may be unlikely given U, the consequences of developing AIDS are so drastic that one may adopt a default $U \Rightarrow A$: one should *act as if* A given U. Preliminary investigations of such defaults, in a conditional setting, may be found in [42, 8]. These may lead to a "practical" form of explanation, with some basis in rational action.

On a related note, our model can be extended with probabilistic information. Boutilier [5] shows how the notion of *counterfactual probabilities* can be grafted onto the conditional logic CO. Probabilistic information can then be used to determine explanations of the type described by Gärdenfors, explanations that are "almost predictive" and to distinguish equally plausible explanations on probabilistic grounds. This should allow a very general model of explanation and diagnosis. We should also remark that the conditional framework allows arbitrary orderings of preference. The orderings described above for Theorist and consistency-based diagnosis are merely illustrative. Generally, orderings need not be determined by default violation and set inclusion. One may, for example, decide that worlds violating the system description of some artifact are more plausible than

worlds where a large number of system components have failed. So if some observation can only be diagnosed with a large number of failures, one may prefer to adopt the hypothesis that the model of the system is in fact inaccurate. Such a viewpoint would be necessary in system design and verification.

Finally, we have neglected an important class of explanation, namely, observations that are explained by appeal to causal or scientific laws. Our explanations have taken for granted a background theory with appropriate conditional information. However, especially in the realm of scientific theory formation, explanations are often causal laws that explain observed correlations. Such explanations require a model of belief revision that allows one to revise a theory with new conditionals. One such model is proposed in [11] and may provide a starting point for such investigations.

7 Acknowledgements

We would like to thank Moisés Goldszmidt and David Poole for their helpful suggestions and discussion of this topic. Thanks also to the referees whose suggestions helped make parts of this paper clearer. This research was supported by NSERC Research Grant OGP0121843.

References

[1] Ernest W. Adams. *The Logic of Conditionals*. D.Reidel, Dordrecht, 1975.

[2] Carlos Alchourrón, Peter Gärdenfors, and David Makinson. On the logic of theory change: Partial meet contraction and revision functions. *Journal of Symbolic Logic*, 50:510–530, 1985.

[3] Craig Boutilier. Inaccessible worlds and irrelevance: Preliminary report. In *Proceedings of the Twelfth International Joint Conference on Artificial Intelligence*, pages 413–418, Sydney, 1991.

[4] Craig Boutilier. Epistemic entrenchment in autoepistemic logic. *Fundamenta Informaticae*, 17(1–2):5–30, 1992.

[5] Craig Boutilier. The probability of a possibility: Adding uncertainty to default rules. In *Proceedings of the Ninth Conference on Uncertainty in AI*, pages 461–468, Washington, D.C., 1993.

[6] Craig Boutilier. Revision sequences and nested conditionals. In *Proceedings of the Thirteenth International Joint Conference on Artificial Intelligence*, pages 519–525, Chambery, FR, 1993.

[7] Craig Boutilier. Conditional logics of normality: A modal approach. *Artificial Intelligence*, 1994. (in press).

[8] Craig Boutilier. Toward a logic for qualitative decision theory. In *Proceedings of the Fourth International Conference on Principles of Knowledge Representation and Reasoning*, Bonn, 1994. To appear.

[9] Craig Boutilier. Unifying default reasoning and belief revision in a modal framework. *Artificial Intelligence*, 1994. (in press).

[10] Craig Boutilier and Veronica Becher. Abduction as belief revision. Technical Report 93-23, University of British Columbia, Vancouver, 1993.

[11] Craig Boutilier and Moisés Goldszmidt. Revision by conditional beliefs. In *Proceedings of the Eleventh National Conference on Artificial Intelligence*, pages 649–654, Washington, D.C., 1993.

[12] Gerhard Brewka. Preferred subtheories: An extended logical framework for default reasoning. In *Proceedings of the Eleventh International Joint Conference on Artificial Intelligence*, pages 1043–1048, Detroit, 1989.

[13] Luca Console and Pietro Torasso. A spectrum of logical definitions of model-based diagnosis. *Computational Intelligence*, 7:133–141, 1991.

[14] Randall Davis. Diagnostic reasoning based on structure and behavior. *Artificial Intelligence*, 24:347–410, 1984.

[15] Johan de Kleer. Focusing on probable diagnoses. In *Proceedings of the Ninth National Conference on Artificial Intelligence*, pages 842–848, Anaheim, 1991.

[16] Johan de Kleer, Alan K. Mackworth, and Raymond Reiter. Characterizing diagnoses. In *Proceedings of the Eighth National Conference on Artificial Intelligence*, pages 324–330, Boston, 1990.

[17] Johan de Kleer and Brian C. Williams. Diagnosing multiple faults. *Artificial Intelligence*, 32:97–130, 1987.

[18] James P. Delgrande. An approach to default reasoning based on a first-order conditional logic: Revised report. *Artificial Intelligence*, 36:63–90, 1988.

[19] Gerhard Friedrich and Wolfgang Nejdl. Choosing observations and actions in model-based diagnosis/repair systems. In *Proceedings of the Third International Conference on Principles of Knowledge Representation and Reasoning*, pages 489–498, Cambridge, 1992.

[20] Antony Galton. A critique of Yoav Shoham's theory of causal reasoning. In *Proceedings of the Ninth National Conference on Artificial Intelligence*, pages 355–359, Anaheim, 1991.

[21] Peter Gärdenfors. On the logic of relevance. *Synthese*, 37(3):351–367, 1978.

[22] Peter Gärdenfors. *Knowledge in Flux: Modeling the Dynamics of Epistemic States*. MIT Press, Cambridge, 1988.

[23] Peter Gärdenfors and David Makinson. Nonmonotonic inference based on expectations. *Artificial Intelligence*, 65:197–245, 1994.

[24] M. R. Genesereth. The use of design descriptions in automated diagnosis. *Artificial Intelligence*, 24:411–436, 1984.

[25] Matthew L. Ginsberg. Counterfactuals. *Artificial Intelligence*, 30(1):35–79, 1986.

[26] Moisés Goldszmidt and Judea Pearl. On the consistency of defeasible databases. *Artificial Intelligence*, 52:121–149, 1991.

[27] Moisés Goldszmidt and Judea Pearl. Rank-based systems: A simple approach to belief revision, belief update, and reasoning about evidence and actions. In *Proceedings of the Third International Conference on Principles of Knowledge Representation and Reasoning*, pages 661–672, Cambridge, 1992.

[28] Adam Grove. Two modellings for theory change. *Journal of Philosophical Logic*, 17:157–170, 1988.

[29] Carl G. Hempel. *Philosophy of Natural Science*. Prentice-Hall, Englewood Cliffs, NJ, 1966.

[30] Jerry R. Hobbs, Mark Stickel, Douglas Appelt, and Paul Martin. Interpretation as abduction. Technical Note 499, SRI International, Menlo Park, December 1990.

[31] Kurt Konolige. Abduction versus closure in causal theories. *Artificial Intelligence*, 53:255–272, 1992.

[32] Kurt Konolige. Using default and causal reasoning in diagnosis. In *Proceedings of the Third International Conference on Principles of Knowledge Representation and Reasoning*, pages 509–520, Cambridge, 1992.

[33] Sarit Kraus, Daniel Lehmann, and Menachem Magidor. Nonmonotonic reasoning, preferential models and cumulative logics. *Artificial Intelligence*, 44:167–207, 1990.

[34] Daniel Lehmann. What does a conditional knowledge base entail? In *Proceedings of the First International Conference on Principles of Knowledge Representation and Reasoning*, pages 212–222, Toronto, 1989.

[35] Hector J. Levesque. A knowledge level account of abduction. In *Proceedings of the Eleventh International Joint Conference on Artificial Intelligence*, pages 1061–1067, Detroit, 1989.

[36] Hector J. Levesque. All I know: A study in autoepistemic logic. *Artificial Intelligence*, 42:263–309, 1990.

[37] David Lewis. Causation. *Journal of Philosophy*, 70:556–567, 1973.

[38] John McCarthy. Epistemological problems in artificial intelligence. In Ronald J. Brachman and Hector J. Levesque, editors, *Readings in Knowledge Representation*, pages 24–30. Morgan-Kaufmann, Los Altos, 1977. 1985.

[39] Sheila McIlraith and Ray Reiter. On experiments for hypothetical reasoning. In *Proc. 2nd International Workshop on Principles of Diagnosis*, pages 1–10, Milan, October 1991.

[40] Donald Nute. *Topics in Conditional Logic*. D.Reidel, Dordrecht, 1980.

[41] Judea Pearl. *Probabilistic Reasoning in Intelligent Systems: Networks of Plausible Inference*. Morgan Kaufmann, San Mateo, 1988.

[42] Judea Pearl. A calculus of pragmatic obligation. In *Proceedings of the Ninth Conference on Uncertainty in Artificial Intelligence*, pages 12–20, Washington, D.C., 1993.

[43] David Poole. A logical framework for default reasoning. *Artificial Intelligence*, 36:27–47, 1988.

[44] David Poole. Explanation and prediction: An architecture for default and abductive reasoning. *Computational Intelligence*, 5:97–110, 1989.

[45] David Poole. Normality and faults in logic-based diagnosis. In *Proceedings of the Eleventh International Joint Conference on Artificial Intelligence*, pages 1304–1310, Detroit, 1989.

[46] David Poole. Representing diagnostic knowledge for probabilistic horn abduction. In *Proceedings of the Twelfth International Joint Conference on Artificial Intelligence*, pages 1129–1135, Sydney, 1991.

[47] Karl R. Popper. *The Logic of Scientific Discovery*. Basic Books, New York, 1959.

[48] W.V. Quine and J.S. Ullian. *The Web of Belief*. Random House, New York, 1970.

[49] Raymond Reiter. A theory of diagnosis from first principles. *Artificial Intelligence*, 32:57–95, 1987.

[50] Raymond Reiter and Johan de Kleer. Foundations of assumption-based truth maintenance systems: Preliminary report. In *Proceedings of the Sixth National Conference on Artificial Intelligence*, pages 183–188, Seattle, 1987.

[51] Raymond Reiter and Alan K. Mackworth. A logical framework for depiction and image interpretation. *Artificial Intelligence*, 41:125–155, 1989.

[52] Nicholas Rescher. *Peirce's Philosophy of Science: Critical Studies in his Theory of Induction and Scientific Method*. University of Notre Dame Press, Notre Dame, 1978.

[53] Krister Segerberg. Modal logics with linear alternative relations. *Theoria*, 36:310–322, 1970.

[54] Murray Shanahan. Prediction is deduction but explanation is abduction. In *Proceedings of the Eleventh International Joint Conference on Artificial Intelligence*, pages 1140–1145, Detroit, 1989.

[55] Yoav Shoham. Nonmonotonic reasoning and causation. *Cognitive Science*, 14:213–252, 1990.

[56] Herbert A. Simon. Nonmonotonic reasoning and causation: Comment. *Cognitive Science*, 15:293–300, 1991.

[57] Robert C. Stalnaker. A theory of conditionals. In W.L. Harper, R. Stalnaker, and G. Pearce, editors, *Ifs*, pages 41–55. D. Reidel, Dordrecht, 1968. 1981.

[58] Bas C. van Fraassen. The pragmatics of explanation. *American Philosophical Quarterly*, 14:143–150, 1977.

A Proofs of Main Theorems

Proposition 3.1 *If* $\alpha, \beta \in K$*, then* $\beta \in (K^-_\beta)^*_\alpha$ *iff* $\neg\alpha \in K^*_{\neg\beta}$.

Proof Let M be an appropriate K-revision model for the contraction and revision function in question. We have $\beta \in (K^-_\beta)^*_\alpha$ iff β is true at each α-world in $\|(K^-_\beta)\|$, i.e., iff β holds at $(\|K\| \cup$ *min*$(\neg\beta)) \cap \|\alpha\|$ (since $\alpha \in K$). This holds iff there is no α-world in *min*$(\neg\beta)$ iff $\neg\alpha \in K^*_{\neg\beta}$. ■

Proposition 3.2 *If* $\neg\alpha, \neg\beta \in K$*, then* $\neg\alpha \in (K^-_{\neg\alpha})^*_{\neg\beta}$ *iff* $\beta \in K^*_\alpha$.

Proof The proof is similar to that of Proposition 3.1. ■

Proposition 3.3 *If* $\alpha, \beta \in K$ *then* α *(predictively) explains* β *iff* $\neg\beta \Rightarrow \neg\alpha$.

Proof If $\alpha, \beta \in K$ then condition **(A)**, $\alpha \Rightarrow \beta$, holds trivially (since $\|K\| =$ *min*$(\alpha) =$ *min*(β)). ■

Proposition 3.4 *If* $\alpha, \beta, \neg\alpha, \neg\beta \notin K$ *then* α *(predictively) explains* β *iff* $\alpha \Rightarrow \beta$ *iff* $\neg\beta \Rightarrow \neg\alpha$.

Proof If $\alpha, \beta, \neg\alpha, \neg\beta \notin K$, then *min*$(\alpha) \subseteq \|K\|$ and *min*$(\neg\beta) \subseteq \|K\|$. Thus, *min*$(\alpha) \subseteq \|\beta\|$ iff *min*$(\neg\beta) \subseteq \|\neg\alpha\|$. ■

Proposition 3.5 *If* $\neg\alpha, \neg\beta \in K$ *then* α *(predictively) explains* β *iff* $\alpha \Rightarrow \beta$.

Proof The proof is similar to that of Proposition 3.3. ■

Proposition 3.6 *If* α *is a predictive explanation for* β *then* α *is a might explanation for* β.

Proof The condition **(ES)** for predictive explanations guarantees the condition **(F)** for might explanations, while $\alpha \Rightarrow \beta \vdash_{CT4O} \alpha \not\Rightarrow \neg\beta$ (for satisfiable α). ■

Proposition 3.7 *If α is a weak explanation for β then α is a might explanation for β.*

Proof $\alpha \rightarrow \beta \vdash_{CT4O} \alpha \not\Rightarrow \neg\beta$ (for satisfiable α). ■

Proposition 3.8 *Let $\beta \in K$ and α be a predictive explanation for β. Then α is a preferred (hypothetical) explanation for β in K_β^-.*

Proof This follows immediately from Proposition 3.1 and the fact that α is epistemically possible in belief state K_β^- (due to the fact that any explanation α must be in K). ■

Proposition 3.9 *Let α be a predictive explanation for β relative to model M. Then α is a preferred explanation iff $M \models \beta \not\rightarrow \neg\alpha$.*

Proof This fact holds trivially for accepted and indeterminate β, since there is a unique minimal β-cluster (those β-worlds satisfying K), and it must intersect $\|\alpha\|$ if α is a predictive explanation. Suppose $\neg\beta \in K$.

If $\beta \rightarrow \neg\alpha$, then there is some minimal β-cluster $\mathcal{C}$ such that $M \models_w \neg\alpha$ for each $w \in \mathcal{C}$. Since β predictively explains itself (see below), we note that β is a strictly preferred to α. To see this, notice that for any $w \in \mathcal{C}$ we have $M \models_w \neg\Diamond\alpha$ (since $\alpha \Rightarrow \beta$, and any such w is in *min*(β)).

If $\beta \not\rightarrow \neg\alpha$, then each minimal β-cluster $\mathcal{C}$ contains some α-world. Thus, we have $M \models \overleftrightarrow{\Box}(\beta \supset \Diamond\alpha)$: α is at least as plausible as β. Clearly, no explanation α' of β is more plausible than β (for then $\alpha' \Rightarrow \beta$ is impossible). Thus, α is preferred. ■

Proposition 4.5 *E is an extension of $\langle \mathcal{F}, \mathcal{D} \rangle$ iff $\|E\| = S$ for some $S \in Pl(\mathcal{F})$.*

Proof By definition of $M_\mathcal{D}$ and Proposition 4.4, $S \in Pl(\mathcal{F})$ iff S consists of the set of worlds satisfying $\mathcal{F} \cup D$, where $D \subseteq \mathcal{D}$ is some maximal subset of defaults consistent with $\mathcal{F}$. By definition of an extension, $S = \|E\|$ for some extension E. ■

Theorem 4.7 *A is predicted (in Theorist sense) from default theory $\langle \mathcal{F}, \mathcal{D} \rangle$ iff $M_\mathcal{D} \models \mathcal{F} \Rightarrow A$.*

Proof We have that A is predicted iff A is in all extensions of $\mathcal{F}$. By Proposition 4.5, this is the case iff $S \subseteq \|A\|$ for all $S \in Pl(\mathcal{F})$. Since *min*$(\mathcal{F}) = \cup Pl(\mathcal{F})$, this holds iff $\mathcal{F} \Rightarrow A$. ■

Theorem 4.8 *Let $C \subseteq \mathcal{C}$. Then C is a Theorist explanation for β iff $M_{\mathcal{D}} \models (\mathcal{F} \wedge C) \rightarrow \beta$ and $\mathcal{F} \wedge C$ is consistent.*

Proof C is a Theorist explanation for β iff $\mathcal{F} \cup D \cup C \models \beta$ for some $D \subseteq \mathcal{D}$ and $\mathcal{F} \cup D \cup C$ is consistent. This is equivalent to β belonging to some extension of the (consistent) set $\mathcal{F} \cup C$, which holds (by Proposition 4.5) iff $S \subseteq \|\beta\|$ for some $S \in Pl(\mathcal{F} \cup C)$ relative to $M_{\mathcal{D}}$ iff $M_{\mathcal{D}} \models (\mathcal{F} \wedge C) \rightarrow \beta$. ■

Theorem 4.9 *Let $C \subseteq \mathcal{C}$. Then C is a predictive Theorist explanation for β iff $M_{\mathcal{D}} \models (\mathcal{F} \wedge C) \Rightarrow \beta$ and $\mathcal{F} \wedge C$ is consistent.*

Proof This follows immediately from Definition 8 and Theorem 4.7. ■

Theorem 4.10 *Let $C, C' \subseteq \mathcal{C}$ be predictive Theorist explanations for β, relative to $\langle \mathcal{F}, \mathcal{D} \rangle$. Then $C \leq_{\mathcal{F}} C'$ iff $M_{\mathcal{D}} \models \overleftrightarrow{\Box}((C' \wedge \mathcal{F}) \supset \Diamond(C \wedge \mathcal{F}))$.*

Proof By definition of $\leq_{\mathcal{F}}$, C is preferred to C' iff each subset of defaults D' consistent with $C' \cup \mathcal{F}$ is contained in some subset of defaults D consistent with $C \cup \mathcal{F}$. By definition of $M_{\mathcal{D}}$ and Proposition 4.4, this is the case iff each world satisfying $C' \cup \mathcal{F}$ sees some world satisfying $C \cup \mathcal{F}$, iff $M_{\mathcal{D}} \models \overleftrightarrow{\Box}((C' \wedge \mathcal{F}) \supset \Diamond(C \wedge \mathcal{F}))$. ■

Proposition 4.19 *Let $M_{\mathcal{B}}$ be the Brewka model for $\langle \mathcal{D}_1, \cdots, \mathcal{D}_n \rangle$ and $M_{\mathcal{D}}$ the Theorist model for its reduction $\mathcal{D} = \mathcal{D}_1 \cup \ldots \cup \mathcal{D}_n$. Then $v \leq w$ in $M_{\mathcal{B}}$ whenever $v \leq w$ in $M_{\mathcal{D}}$.*

Proof If $v \leq w$ in $M_{\mathcal{D}}$, then $V(v) \subseteq V(w)$ relative to the flat set of defaults $\mathcal{D}$. Clearly then $V_i(v) \subseteq V_i(w)$ for each i relative to the prioritized set of defaults. By definition, $v \leq w$ in $M_{\mathcal{B}}$. ■

Theorem 4.20 *If $M_{\mathcal{D}} \models \alpha \Rightarrow \beta$ then $M_{\mathcal{B}} \models \alpha \Rightarrow \beta$.*

Proof By Proposition 4.19, it is clear that the set of minimal α-worlds in the Brewka model $M_{\mathcal{B}}$ is a subset of the minimal α-worlds in the Theorist model $M_{\mathcal{D}}$. Thus, if $M_{\mathcal{D}} \models \alpha \Rightarrow \beta$ then $M_{\mathcal{B}} \models \alpha \Rightarrow \beta$. ■

Theorem 5.3 *Let SD and COMP determine some system.* $D(\Delta)$ *is a CB-diagnosis for observation* β *iff* $AB(\Delta)$ *is an excuse for* β *relative to* M_{COMP}.

Proof By definition, $AB(\Delta)$ is an excuse for β iff $M_{COMP} \models AB(\Delta) \wedge SD \not\Rightarrow \neg\beta$. We note that this relation can hold only if $AB(\Delta) \wedge SD$ is consistent. Given this consistency and Proposition 5.2, we have that $AB(\Delta)$ is an excuse iff $M_{COMP} \models D(\Delta) \wedge SD \not\Rightarrow \neg\beta$ (this follows from the valid schematic entailment of $A \wedge B \not\Rightarrow C$ from $A \Rightarrow B$ and $A \not\Rightarrow C$). This holds iff $SD \cup \{\beta, D(\Delta)\}$ is consistent iff $D(\Delta)$ is a CB-diagnosis for β. ■

Theorem 5.4 $D(\Delta)$ *is a preferred diagnosis iff* $D(\Delta)$ *is a minimal diagnosis.*

Proof This follows immediately from the definition of M_{COMP}. ■

Theorem 5.6 *Assume that* $D(\Delta) \wedge SD$ *is consistent.* $D(\Delta)$ *is a predictive diagnosis for* β *iff*

$$M_{COMP} \models AB(\Delta) \wedge SD \Rightarrow \beta$$

Proof We observe that $min(D(\Delta) \wedge SD)$ consists of the set of all worlds satisfying $D(\Delta) \wedge SD$ by definition of M_{COMP}. Thus

$$M_{COMP} \models D(\Delta) \wedge SD \Rightarrow \beta$$

iff $SD \cup \{D(\Delta)\} \models \beta$, i.e., iff $D(\Delta)$ is a predictive diagnosis. By Proposition 5.2,

$$M_{COMP} \models D(\Delta) \wedge SD \Rightarrow \beta$$

iff

$$M_{COMP} \models AB(\Delta) \wedge SD \Rightarrow \beta$$

■

Proposition 5.8 *Let* $\Delta \subseteq COMP$. $M_{COMP} \models AB(\Delta) \rightarrow \beta$ *iff* $M_{COMP} \models AB(\Delta) \Rightarrow \beta$. *(Similarly for* $D(\Delta)$*.)*

Proof As observed above, the set of clusters in the model M_{COMP} are distinguished by the set of components they take to be normal and abnormal. This means that the sets $Pl(AB(\Delta))$ and $Pl(D(\Delta))$ are singletons consisting of a single cluster each, these clusters being exactly *min*$((AB(\Delta))$ and *min*$((D(\Delta))$ respectively. Thus, $AB(\Delta) \rightarrow \beta$ iff $AB(\Delta) \Rightarrow \beta$ and $D(\Delta) \rightarrow \beta$ iff $D(\Delta) \Rightarrow \beta$. ■

Theorem 5.9 *Let SD and COMP describe some system,* $\Delta \subseteq$ *COMP, and* $\mathcal{D}$ *be the set of defaults* $\{\neg ab(c) : c \in$ *COMP*$\}$. *Then* $D(\Delta)$ *is a CB-diagnosis for observation* β *iff* $\neg\beta \notin E$ *where* E *is the (only) Theorist extension of* $\langle$*SD* $\cup \{AB(\Delta)\}, \mathcal{D}\rangle$.

Proof We assume that *SD* $\cup \{D(\Delta)\}$ is consistent. By Theorem 5.3, $D(\Delta)$ is a CB-diagnosis for β iff

$$M_{COMP} \models AB(\Delta) \wedge \textit{SD} \not\Rightarrow \neg\beta$$

As indicated in the proof of Proposition 5.8, there is a unique minimal *SD* $\cup \{AB(\Delta)\}$-cluster in M_{COMP}; and as described in Section 4, this cluster determines the Theorist extension E of *SD* $\cup \{AB(\Delta)\}$. Thus,

$$M_{COMP} \models AB(\Delta) \wedge \textit{SD} \not\Rightarrow \neg\beta$$

iff $\neg\beta \notin E$. ■