Abduction As Belief Revision: A Model of Preferred Explanations

Craig Boutilier and Verónica Becher
Department of Computer Science
University of British Columbia
Vancouver, British Columbia
CANADA, V6T 1Z2
email: cebly,becher@cs.ubc.ca

Abstract

We propose a natural model of abduction based on the revision of the epistemic state of an agent. We require that explanations be sufficient to induce belief in an observation in a manner that adequately accounts for factual and hypothetical observations. 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. We reconstruct the Theorist system in our framework, and show how it can be extended to accommodate our predictive explanations and semantic preferences on explanations.

1 Introduction

A number of different approaches to abduction have been proposed in the AI literature that model the concept of abduction as some sort of deductive relation between an explanation and the explanandum, the "observation" it purports to explain (e.g., Hempel's (1966) *deductive-nomological* explanations). Theories of this type are, unfortunately, bound to the unrelenting nature of deductive inference. There are two 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. Second, while there may be many competing explanations for a particular observation, certain of these may be relatively implausible. Thus we require some notion of preference to chose among these potential explanations.

Both of these problems can be addressed using, for example, probabilistic information (Hempel 1966; de Kleer and Williams 1987; Poole 1991; Pearl 1988): we might simply require that an explanation render the observation sufficiently probably and that most likely explanations be preferred. Explanations might thus *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). There have been proposals to address these issues in a more qualitative manner using "logic-based" frameworks also. Peirce (see Rescher (1978)) discusses the "plausibility" of explanations, as do Quine and Ullian (1970). Consistency-based diagnosis (Reiter 1987; de Kleer, Mackworth and Reiter 1990) uses abnormality assumptions to capture the context dependence of explanations; and preferred explanations are those that minimize abnormalities. Poole's (1989) assumption-based framework captures some of these ideas by explicitly introducing a set of default assumptions to account for the nonmonotonicity of explanations.

We propose a semantic framework for abduction that captures the spirit of probabilistic proposals, but in a qualitative fashion, and in such a way that existing logic-based proposals can be represented as well. 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. 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. Semantically, such conditionals are interpreted relative to an ordering of plausibility or normality over worlds. Our conditional logic, described in earlier work as a representation of belief revision and default reasoning (Boutilier 1991; 1992b; 1992c), has the desired nonmonotonicity and induces a natural preference ordering on sentences (hence explanations). In the next section we describe our conditional logics and the necessary logical preliminaries. In Section 3, we discuss the concept of explanation, its epistemic nature, and its definition 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 Theorist framework (and Brewka's (1989) extension) 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 explanations. These two abilities have until now

been unexplored in this canonical abductive framework. We conclude by describing directions for future research, and how consistency-based diagnosis also fits in our system.

2 Conditionals and Belief Revision

The problem of revising a knowledge base or belief set when new information is learned has been well-studied in AI. One of the most influential theories of belief revision is the *AGM theory* (Alchourrón, Gärdenfors and Makinson 1985; Gärdenfors 1988). If we take an agent to have a (deductively closed) belief set K, adding new information A to K is problematic if $K \vdash \neg A$. Intuitively, certain beliefs in K must be retracted before A can be accepted. The AGM theory provides a set of constraints on acceptable belief revision functions $*$. Roughly, using K_A^* to denote the belief set resulting when K is revised by A, the theory maintains that the least "entrenched" beliefs in K should be given up and then A added to this *contracted* belief set.

Semantically, this process can be captured by considering a *plausibility ordering* over possible worlds. As described in (Boutilier 1992b; Boutilier 1992a), we can use a family of logics to capture the AGM theory of revision. The modal logic CO is based on a propositional language (over variables P) augmented with two modal operators $\Box$ and $\overleftarrow{\Box}$. $\mathbf{L}_{CPL}$ denotes the propositional sublanguage of this bimodal language $\mathbf{L}_B$. The sentence $\Box\alpha$ is read as usual as "α is true at all *equally or more plausible* worlds." In contrast, $\overleftarrow{\Box}\alpha$ is read "α is true at all *less plausible* worlds."

A CO-model is a triple $M = \langle W, \leq, \varphi \rangle$, where W is a set of worlds with valuation function φ and $\leq$ is a plausibility ordering over W. If $w \leq v$ the w is at least as plausible as v. We insist that $\leq$ be transitive and connected (that is, either $w \leq v$ or $v \leq w$ for all w, v). CO-structures consist of a totally-ordered set of *clusters* of worlds, where a cluster is simply a maximal set of worlds $C \subseteq W$ such that $w \leq v$ for each $w, v \in C$ (that is, no extension of C enjoys this property). This is evident in Figure 1(b), where each large circle represents a cluster of equally plausible worlds. Satisfaction of a modal formula at w is given by:

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

We define several new connectives as follows: $\Diamond\alpha \equiv_{\text{df}} \neg\Box\neg\alpha$; $\overleftarrow{\Diamond}\alpha \equiv_{\text{df}} \neg\overleftarrow{\Box}\neg\alpha$; $\overleftrightarrow{\Box}\alpha \equiv_{\text{df}} \Box\alpha \wedge \overleftarrow{\Box}\alpha$; and $\overleftrightarrow{\Diamond}\alpha \equiv_{\text{df}} \neg\overleftrightarrow{\Box}\neg\alpha$. It is easy to verify that these connectives have the following truth conditions: $\Diamond\alpha$ ($\overleftarrow{\Diamond}\alpha$) is true at a world if α holds at some more plausible (less plausible) world; $\overleftrightarrow{\Box}\alpha$ ($\overleftrightarrow{\Diamond}\alpha$) holds iff α holds at all (some) worlds, whether more or less plausible.

The modal logic CT4O is a weaker version of CO, where we weaken the condition of connectedness to be simple reflexivity. This logic is based on models whose structure is that of a partially-ordered set of clusters (see Figure 1(a)). Both logics can be extended by requiring that the set of worlds in a model include every propositional valuation over

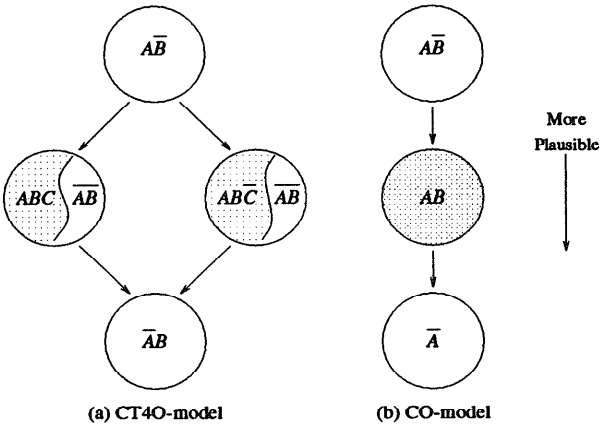

Figure 1: CT4O and CO models

P (so that every logically possible state of affairs is possible). The corresponding logics are denoted CO* and CT4O*. Axiomatizations for all logics may be found in (Boutilier 1992b; Boutilier 1992a). For a given model, we define the following notions. We let $||\alpha||$ denote the set of worlds satisfying formula α (and also use this notion for sets of formulae K). We use $min(\alpha)$ to denote the set of *most plausible* α-worlds:[1]

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

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 $||K|| = min(\top)$; that is, $||K||$ forms the (unique) minimal cluster in the model. This reflects the intuition that all and only K-worlds are most plausible (Boutilier 1992b). The CT4O-model in Figure 1(a) is a K-revision model for $K = Cn(\neg A, B)$, while the CO-model in Figure 1(b) is suitable for $K = Cn(\neg A)$.

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 $||K_A^*|| = min(A)$; 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:

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

Both models in Figure 1 satisfy $A \Rightarrow B$, since B holds at each world in the shaded regions, $min(A)$, of the models. Using the *Ramsey test* for acceptance of conditionals (Stalnaker 1968), we equate $B \in K_A^*$ with $M \models A \Rightarrow B$. Indeed, for both models we have that $K_A^* = Cn(A, B)$. If the model in question is a CO*-model then this characterization of revision is equivalent to the AGM model (Boutilier

[1] 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 (Boutilier 1992b).

1992b). Simply using CT4O*, the model satisfies all AGM postulates (Gärdenfors 1988) but the eighth. Properties of this conditional logic are described in Boutilier (1990; 1991).

We briefly describe the *contraction* of K by $\neg A$ in this semantic framework. To retract belief in $\neg A$, we adopt the belief state determined by the set of worlds $||K|| \cup min(A)$. The belief set $K^-_{\neg A}$ does not contain $\neg A$, and this operation captures the AGM model of contraction. In Figure 1(a) $K^-_{\neg A} = Cn(B)$, while in Figure 1(b) $K^-_{\neg A} = Cn(A \supset B)$.

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 $Pl(A)$, so that $min(A) = \cup Pl(A)$.[2] Taking each such subset 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 at *some* element of $Pl(A)$, C holds:

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

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.

Finally, we define the *plausibility* of a proposition. 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 $\vec{\Box}(B \supset \Diamond A)$. 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 $A \wedge \neg B$.

3 Epistemic Explanations

Often explanations are postulated relative to some background theory, which together with the explanation entails the observation. Our notion of explanation will be somewhat different than the usual ones. We define an explanation relative to the epistemic state of some agent (or program). An agent's beliefs *and* judgements of plausibility will be crucial in its evaluation of what counts as a valid explanation (see Gärdenfors (1988)). We assume a deductively closed belief set K along with some set of conditionals that represent the revision policies of the agent. These conditionals may represent statements of normality or simply subjunctives (below).

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 most 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," to justify its reasoning. The explanation should clearly be given in terms of *other* (perhaps more fundamental) beliefs held by the program. This applies to advice-systems, intelligent databases, tutorial systems, or a robot that must explain its actions.

A second type of explanation is *hypothetical*. Even if β is not believed, we may want a hypothetical 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 diagnosis tasks (see below), planning, and so on (Ginsberg 1986). For example, if A explains B in this sense, it may be that ensuring A will bring about B. If α is to count as an explanation of β in this case, we must insist that α is also not believed. If it were, it would hardly make sense as a predictive explanation, for the agent has already adopted belief in α without committing to β. This leads us to the following condition on epistemic explanations: if α is an explanation for β then α and β must have the same epistemic status for the agent. In other words, $\alpha \in K$ iff $\beta \in K$ and $\neg\alpha \in K$ iff $\neg\beta \in K$.[3]

Since our explanations are to be predictive, there has to be some sense in which α is sufficient to cause acceptance of β. On our interpretation of conditionals (using the Ramsey test), this is the case just when the agent believes the conditional $\alpha \Rightarrow \beta$. So for α to count as an explanation of β (in this predictive sense, at least) this conditional relation must hold.[4] In other words, if the explanation were believed, so too would the observation.

Unfortunately, this conditional is vacuously satisfied when β is believed, once we adopt the requirement that α be believed too. Any $\alpha \in K$ is such that $\alpha \Rightarrow \beta$; but surely arbitrary beliefs cannot count as explanations. To determine an explanation for some $\beta \in K$, we want to (hypothetically) suspend belief in β and, *relative to this new belief state*, eval-

[2] $Pl(A) = \{min(A) \cap \mathcal{C} : \mathcal{C} \text{ is a cluster}\}$.

[3] This is at odds with one prevailing view of explanation, which takes only non-beliefs to be valid explanations: 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, β is unexpected. An explanation relieves this dissonance when it is accepted (Gärdenfors 1988). 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. *Factual* explanations are retrospective in the sense that they (should) describe "historically" what explanation was *actually* adopted for a certain belief.

In (Becher and Boutilier 1993) we explore a weakening of this condition on epistemic status. Preferences on explanations (see below) then play a large role in ruling out any explanation whose epistemic status differs from that of the observation.

[4] See the below for a discussion of non-predictive explanations.

uate the conditional $\alpha \Rightarrow \beta$. This hypothetical belief state should simply be the *contraction* of K by β. The contracted belief set K_β^- is constructed as described in the last section. We can think of it as the set of beliefs held by the agent before it came to accept β.[5] In general, the conditionals an agent accepts relative to the contracted set need not bear a strong relation to those in the original set. Fortunately, we are only interested in those conditionals $\alpha \Rightarrow \beta$ where $\alpha \in K$. The AGM contraction operation ensures that $\neg\alpha \notin K_\beta^-$. This means that we can determine the truth of $\alpha \Rightarrow \beta$ relative to K_β^- by examining conditionals in the original belief set. We simply need to check if $\neg\beta \Rightarrow \neg\alpha$ relative to K. This is our final criterion for explanation. If the observation had been absent, so too would the explanation.

We assume, for now, the existence of a model M that captures an agent's objective belief set K and its revision policies (e.g., M completely determines K_A^*, K_A^- and accepted conditionals $A \Rightarrow B$). When we mention a belief set K, we have in mind also the appropriate model M. All conditionals are evaluated with respect to K unless otherwise indicated. We can summarize the considerations above:

Definition A *predictive explanation* of $\beta \in \mathbf{L}_{CPL}$ relative to belief set K is any $\alpha \in \mathbf{L}_{CPL}$ such that: (1) $\alpha \in K$ iff $\beta \in K$ and $\neg\alpha \in K$ iff $\neg\beta \in K$; (2) $\alpha \Rightarrow \beta$; and (3) $\neg\beta \Rightarrow \neg\alpha$.

As a consequence of this definition, we can have the following property of factual explanations:

Proposition 1 *If $\alpha, \beta \in K$ then α explains β iff $\alpha \Rightarrow \beta$ is accepted in K_β^-.*

Thus factual explanations satisfy our desideratum regarding contraction by β. Furthermore, for both factual and hypothetical explanations, only one of conditions (2) or (3) needs to be tested, the other being superfluous:

Proposition 2 *(i) If $\alpha, \beta \in K$ then α explains β iff $\neg\beta \Rightarrow \neg\alpha$; (ii) If $\alpha, \beta \notin K$ then α explains β iff $\alpha \Rightarrow \beta$.*

Figure 2 illustrates both factual and hypothetical explanations. In the first model, wet grass (W) is explained by rain (R), since $R \Rightarrow W$ holds in that model. Similarly, sprinkler S 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 (the lawn is covered) 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 (or otherwise), 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

[5] We do not require that this must *actually* be the case.

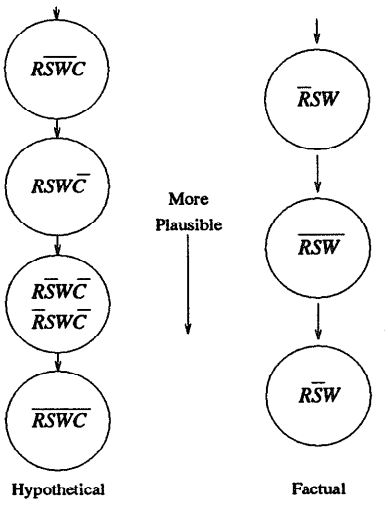

Figure 2: Explanations for "Wet Grass"

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. Notice that if we relax the condition on epistemic status, we might accept S as a hypothetical explanation for factual belief R. This is explored in (Becher and Boutilier 1993).

Semantic Preferences: Predictive explanations are very general, for any α that induces belief in β satisfies our conditions. Of course, some such explanations should be ruled out on grounds of implausibility (e.g., a tanker truck exploding in front of my house explains wet grass). 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. Preferences can be easily accommodated within our framework. We assume that the β to be explained is not (yet) believed and rank possible explanations for β.[6] 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 described in the last section.

Definition If α and α' both explain β then α is *at least as preferred as* α' (written $\alpha \leq_P \alpha'$) iff $M \models \overleftrightarrow{\Box}(\alpha' \supset \Diamond\alpha)$. The *preferred explanations* of β are those α such that not $\alpha' <_P \alpha$ for all explanations 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 while R, S and $R \wedge S$ each explain W, R and S are preferred to $R \wedge S$ (I may not know whether my sprinkler was

[6] We adopt the view that 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. In fact, the only explanations in K can be those that are preferred in K_β^-.

on or it rained, but it's unlikely that my sprinkler was on in the rain). If we want to represent the fact, say, that the failure of fewer components is more plausible than more failures, we simply rank worlds accordingly. Preferred explanations of β are those that predict β and presume as few faults as possible.[7] We can characterize preferred explanations by appealing to their "believability" given β:

Proposition 3 *α is a preferred explanation for β iff* $M \models \neg(\beta \rightarrow \neg\alpha)$.

In the next section, we discuss the role of $\rightarrow$ further.

This approach to preferred explanations is very general, and is completely determined by the conditionals (or defaults) held by an agent.[8] We needn't restrict the ordering to, say, counting component failures. It can be used to represent any notion of typicality, normality or plausibility required. For instance, we might use this model of abduction in scene interpretation to "explain" the occurrence of various image objects by the presence of actual scene objects (Reiter and Mackworth 1989). Preferred explanations are those that match the data best. However, we can also introduce an extra level of preference to capture preferred interpretations, those scenes that are *most likely* in a given domain among those with the best fit.

We should point out that we do not require a complete semantic model M to determine explanations. For a given incomplete theory, one can simply use the derivable conditionals to determine derivable explanations and preferences. This paper simply concentrates on the semantics of this process. All conditions on explanations can be tested as object-level queries on an incomplete *KB*. However, should one have in mind a complete ordering of plausibility (as in the next section), these can usually be represented as a compact object-level theory as well (Boutilier 1991).

Other issues arise with this semantic notion of explanation. Consider the wet grass example, and the following conditionals: $R \Rightarrow W$, $S \Rightarrow W$ and $S \wedge R \Rightarrow W$ (note that the third does not follow from the others). We may be in a situation where rain is preferred to sprinkler as an explanation for wet grass (it is more likely). But we might be in a situation where R and S are equally plausible explanations.[9] We might then have $W \Rightarrow (S \equiv \neg R)$. That is, S and R are the *only* plausible "causes" for W (and are mutually exclusive). Notice that $S \equiv \neg R$ is a preferred explanation for W, as is $S \vee R$. We say α is a *covering explanation* for β iff α is a preferred explanation such that $\beta \Rightarrow \alpha$. Such an α represents all preferred explanations for β.[10]

[7] In consistency-based systems, explanations usually do not *predict* an observation without adequate fault models (more on this in the concluding section).

[8] Direct statements of belief, relative plausibility, integrity constraints, etc. in $\mathbf{L}_B$ may also be in an agent's *KB*.

[9] We can ensure that $R \wedge S$ is less likely, e.g., by asserting $S \Rightarrow \neg R$ and $R \Rightarrow \neg S$.

[10] Space limitations preclude a full discussion (see (Becher and Boutilier 1993)), but we might think of a covering explanation as the disjunction of all likely causes of β in a causal network (Pearl 1988). We are currently investigating *causal explanations* in our conditional framework and how a theory might be used to derive causal influences (Lewis 1973; Goldszmidt and Pearl 1992).

Pragmatics: We note that β is always an explanation for itself. Indeed, semantically β is as good as any other explanation, for if one is convinced of this *trivial* explanation, one is surely convinced of the proposition to be explained. There are many circumstances in which such an explanation is reasonable (for instance, explaining the value of a root node in a causal network); otherwise we would require infinite regress or circular explanations.

The undesirability of such trivial explanations, in certain circumstances, is not due to a lack of predictive power or plausibility, but rather its *uninformative* nature. We think it might be useful to rule out trivial explanations as a matter of the *pragmatics* of explanation rather than semantics, much like Gricean maxims (but see also Levesque (1989)). But, we note, that in many cases, trivial (or overly specific) explanations may be desirable. We discuss this and other pragmatic issues (e.g., irrelevance) in the full paper (Becher and Boutilier 1993). We note that in typical approaches to diagnosis this problem does not arise. Diagnoses are usually selected from a pre-determined set of conjectures or component failures. This can be seen as simply another form of pragmatic filtering, and can be applied to our model of abduction (see below).

4 Reconstructing Theorist

Poole's (1989) Theorist system is an assumption-based model of explanation and prediction where observations are explained (or predicted) by adopting certain hypotheses that, together with known facts, entail these observations. We illustrate the naturalness and generality of our abductive framework by recasting Theorist in our model. It shows why Theorist explanations are paraconsistent and non-predictive, how they can be made predictive, and how a natural account of preferred explanation can be introduced to Theorist (and Brewka's (1989) extension of it). Our presentation of Theorist will be somewhat more general than that found in (Poole 1989), but unchanged in essential detail.

We assume the existence of a set $\mathcal{D}$ of *defaults*, a set of propositional formulae taken to be "expectations," or facts that normally hold (Boutilier 1992c). We assume $\mathcal{D}$ is consistent.[11] Given a fixed set of defaults, we are interested in what follows from a given (known) finite set of facts $\mathcal{F}$; we use F to denote its conjunction. A *scenario* for $\mathcal{F}$ is any subset D of $\mathcal{D}$ such that $\mathcal{F} \cup D$ is consistent. An *extension* of $\mathcal{F}$ is any maximal scenario. An *explanation* of β given $\mathcal{F}$ is any α such that $\{\alpha\} \cup \mathcal{F} \cup D \models \beta$ for some scenario D of $\{\alpha\} \cup \mathcal{F}$.[12] Finally, β is *predicted* given $\mathcal{F}$ iff $\mathcal{F} \cup D \models \beta$ for each extension D of $\mathcal{F}$.

In the definition of prediction in Theorist, we find an implicit notion of plausibility: we expect some maximal subset of defaults, consistent with $\mathcal{F}$, to hold. Worlds that violate

[11] Nothing crucial depends on this however.

[12] Theorist explanations are usually drawn from a given set of conjectures, but this is not crucial.

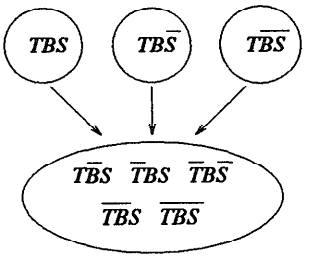

Figure 3: A Theorist Model

more defaults are thus less plausible than those that violate fewer. We define a CT4O*-model that reflects this.

Definition For a fixed set of defaults $\mathcal{D}$, and a possible world (valuation) w, the *violation set* for w is defined as $V(w) = \{d \in \mathcal{D} : w \models \neg d\}$. The *Theorist model* for $\mathcal{D}$ is $M_{\mathcal{D}} = \langle W, \leq, \varphi \rangle$ where W and φ are as usual, and $\leq$ is an ordering of plausibility such that $v \leq w$ iff $V(v) \subseteq V(w)$.

Thus, $M_{\mathcal{D}}$ ranks worlds according to the sets of defaults they violate. We note that $M_{\mathcal{D}}$ is a CT4O*-model, and if $\mathcal{D}$ is consistent, $M_{\mathcal{D}}$ has a unique minimal cluster consisting of those worlds that satisfy each default. It should be clear that worlds w, v are equally plausible iff $V(w) = V(v)$, so that each cluster in $M_{\mathcal{D}}$ is the set of worlds that violate a particular subset $D \subseteq \mathcal{D}$. The α-worlds minimal in $M_{\mathcal{D}}$ are just those that satisfy some maximal subset of defaults consistent with α.

Theorem 4 *β is* predicted *given $\mathcal{F}$ iff $M_{\mathcal{D}} \models F \Rightarrow \beta$.*

Thus, predictions based on $\mathcal{F}$ correspond to the belief set obtained when $\mathcal{D}$ is revised to incorporate $\mathcal{F}$. This is the view of default prediction discussed in (Boutilier 1992c).

We now turn our attention to explanations. Theorist explanations are quite weak, for α explains β whenever there exists *any* set of defaults that, together with α, entails β. This means that α might explain both β and $\neg\beta$. Such explanations are in a sense paraconsistent, for α cannot usually be used to explain the conjunction $\beta \wedge \neg\beta$. Furthermore, such explanations are not predictive: if α explains contradictory sentences, how can it be thought to predict either? Consider a set of defaults in Theorist

$$\mathcal{D} = \{T \supset S, T \wedge B \supset \neg S\}$$

which assert that my car will start (S) when I turn the key (T), unless my battery is dead (B). The Theorist model $M_{\mathcal{D}}$ is shown in Figure 3. Suppose our set of facts $\mathcal{F}$ has a single element B. When asked to explain S, Theorist will offer T. When asked to explain $\neg S$, Theorist will again offer T. If I want my car to start I should turn the key, and if I do not want my car to start I should turn the key. There is certainly something unsatisfying about such a notion of explanation. Such explanations do, however, correspond precisely to *weak explanations* in CT4O using $\rightarrow$.

Theorem 5 *α is a Theorist explanation of β given $\mathcal{F}$ iff $M_{\mathcal{D}} \models \alpha \wedge F \rightarrow \beta$.*

This illustrates the conditional and defeasible semantic underpinnings of Theorist's weak (paraconsistent) explanations in the conditional framework.

In our model, the notion of predictive explanation seems much more natural. In the Theorist model above, there is a possibility that $T \wedge B$ gives S and a possibility that $T \wedge B$ gives $\neg S$. Therefore, T (given B) *explains* neither possibility. One cannot use the explanation to ensure belief in the "observation" S. We can use our notion of predictive explanation to extend Theorist with this capability. Clearly, predictive explanations in the Theorist model $M_{\mathcal{D}}$ give us:

Definition α is a *predictive explanation* for β given $\mathcal{F}$ iff β is predicted (in the Theorist sense) given $\mathcal{F} \cup \{\alpha\}$.

Theorem 6 *α is a predictive explanation for β given $\mathcal{F}$ iff $M_{\mathcal{D}} \models \alpha \wedge F \Rightarrow \beta$ (i.e., iff $\mathcal{F} \cup D \cup \{\alpha\} \models \beta$ for each extension D).*

Taking those α-worlds that satisfy as many defaults as possible to be the most plausible or typical α-worlds, it is clear that revising by α should result in acceptance of those situations, and thus α should (predictively) explain β iff β holds in each such situation. Such explanations are often more useful than weak explanations for they suggest *sufficient* conditions α that *will* (defeasibly) lead to a desired belief β. Weak explanations of the type originally defined in Theorist, in contrast, merely suggest conditions that *might* lead to β.

Naturally, given the implicit notion of plausibility determined by $\mathcal{D}$, we can characterize *preferred* explanations in Theorist. These turn out to be exactly those explanations that force the violation of as few defaults as possible.

Definition Let α, α' be predictive explanations for β given $\mathcal{F}$. α is *at least as preferred as* α' (written $\alpha \leq_{\mathcal{F}} \alpha'$) iff each extension of $\mathcal{F} \cup \{\alpha'\}$ is contained in some extension of $\mathcal{F} \cup \{\alpha\}$.

Theorem 7 *$\alpha \leq_{\mathcal{F}} \alpha'$ iff $M_{\mathcal{D}} \models \vec{\Box}((\alpha' \wedge F) \supset \Diamond(\alpha \wedge F))$.*

So the notion of preference defined for our concept of epistemic explanations induces a preference in Theorist for predictive explanations that are consistent with the greatest subsets of defaults; that is, those explanations that are most plausible or most normal (see Konolige (1992) who proposes a similar notion).

This embedding into CT4O provides a compelling semantic account of Theorist in terms of plausibility and belief revision. But it also shows directions in which Theorist can be naturally extended, in particular, with predictive explanations and with preferences on semantic explanations, notions that have largely been ignored in assumption-based explanation.

In (Becher and Boutilier 1993) we show how these ideas apply to Brewka's (1989) prioritized extension of Theorist by ordering worlds in such a way that the prioritization relation among defaults is accounted for. If we have a prioritized default theory $D = D_1 \cup \cdots D_n$, we still cluster worlds according to the defaults they violate; but should w violate fewer high priority defaults than v, even if it violates more low priority defaults, w is considered more plausible than v.

This too results in a CT4O*-model; and prediction, (weak and predictive) explanation, and preference on explanations are all definable in the same fashion as with Theorist. We also show that priorities on defaults, as proposed by Brewka, simply prune away certain weak explanations and make others preferred (possibly adding predictive explanations). For instance, the counterintuitive explanation T above, for S given B, is pruned away if we require that the default $T \supset S$ be given lower priority than the default $T \wedge B \supset \neg S$. A model for such a prioritized theory simply makes the world TBS less plausible than $TB\overline{S}$. We note, however, that such priorities need not be provided explicitly if the Theorist model is abandoned and defaults are expressed directly as conditionals. This preference is derivable in CT4O from the conditionals $T \Rightarrow S$ and $T \wedge B \Rightarrow \neg S$ automatically.

5 Concluding Remarks

We have proposed a notion of epistemic explanation based on belief revision, and preferences over these explanations using the concept of plausibility. We have shown how Theorist can be captured in this framework. In (Becher and Boutilier 1993), we show how this model can be axiomatized. We can also capture consistency-based diagnosis in our framework, though it does not usually require that explanations be predictive in the sense we describe. Instead, consistency-based diagnosis is characterized in terms of "might" counterfactuals, or *excuses* that make an observation plausible, rather than likely (Becher and Boutilier 1993). Of course, fault models describing how failures are manifested in system behavior make explanations more predictive, in our strong sense. However, the key feature of this approach is not its ability to represent existing models of diagnosis, but its ability to infer explanations, whether factual or hypothetical, from existing conditional (or default) knowledge. We are also investigating the role of causal explanations in abduction, and how one might distinguish causal from non-causal explanations using only conditional information.

Acknowledgements: Thanks to David Poole for helpful comments. This research was supported by NSERC Research Grant OGP0121843.

References

Alchourrón, C., Gärdenfors, P., and Makinson, D. 1985. On the logic of theory change: Partial meet contraction and revision functions. *Journal of Symbolic Logic*, 50:510–530.

Becher, V. and Boutilier, C. 1993. Epistemic explanations. Technical report, University of British Columbia, Vancouver. forthcoming.

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

Boutilier, C. 1992a. Conditional logics for default reasoning and belief revision. Technical Report KRR-TR-92-1, University of Toronto, Toronto. Ph.D. thesis.

Boutilier, C. 1992b. A logic for revision and subjunctive queries. In *Proceedings of the Tenth National Conference on Artificial Intelligence*, pages 609–615, San Jose.

Boutilier, C. 1992c. Normative, subjunctive and autoepistemic defaults: Adopting the Ramsey test. In *Proceedings of the Third International Conference on Principles of Knowledge Representation and Reasoning*, pages 685–696, Cambridge.

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

de Kleer, J., Mackworth, A. K., and Reiter, R. 1990. Characterizing diagnoses. In *Proceedings of the Eighth National Conference on Artificial Intelligence*, pages 324–330, Boston.

de Kleer, J. and Williams, B. C. 1987. Diagnosing multiple faults. *Artificial Intelligence*, 32:97–130.

Gärdenfors, P. 1988. *Knowledge in Flux: Modeling the Dynamics of Epistemic States*. MIT Press, Cambridge.

Ginsberg, M. L. 1986. Counterfactuals. *Artificial Intelligence*, 30(1):35–79.

Goldszmidt, M. and Pearl, J. 1992. 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.

Hempel, C. G. 1966. *Philosophy of Natural Science*. Prentice-Hall, Englewood Cliffs, NJ.

Konolige, K. 1992. 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.

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

Lewis, D. 1973. Causation. *Journal of Philosophy*, 70:556–567.

Pearl, J. 1988. *Probabilistic Reasoning in Intelligent Systems: Networks of Plausible Inference*. Morgan Kaufmann, San Mateo.

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

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

Quine, W. and Ullian, J. 1970. *The Web of Belief*. Random House, New York.

Reiter, R. 1987. A theory of diagnosis from first principles. *Artificial Intelligence*, 32:57–95.

Reiter, R. and Mackworth, A. K. 1989. A logical framework for depiction and image interpretation. *Artificial Intelligence*, 41:125–155.

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

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