What is a Default Priority?

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

Abstract

The notion of default priority has played a central role in default reasoning research. We show that Pearl's Z-ranking of default rules need not always correspond to priorities. System Z can still be used for priorities, but perhaps not in the obvious manner. Rather than the Z-ranking of rules, we show that the Z-rankings of the *negations of the material counterparts of rules* correspond naturally to priorities. We also show that the priorities of default rules can be explained in terms of belief revision by appeal to the epistemic entrenchment of the material counterpart of a rule in a *theory of expectations* in the Theorist sense. Furthermore, Brewka's notion of preferred subtheories provides a means of improving on Pearl's 1-entailment, given this connection. These results are demonstrated within the modal logic CO*, a unifying framework for various types of default reasoning and belief revision.

1 Introduction

The notion of default priority has played a central role in default reasoning research. Default rules can lead to conflicting conclusions based on certain evidence, but some rules seem to naturally take priority over others. Thus, conflicts are resolved by permitting the violation of lower priority rules in order to satisfy higher priority rules, those that are deemed more important or seen as somehow providing more information.

Many mechanisms have been proposed for representing priorities in systems like default logic and circumscription. The use of semi-normal defaults has been proposed for asserting priorities in default logic (Reiter and Criscuolo 1981). McCarthy's (1986) simple abnormality theories also embody this notion through the introduction of *cancellation of inheritance axioms*, while prioritized circumscription allows the explicit expression of priorities (McCarthy 1986; Lifschitz 1985).

While these systems allow users to express priorities, nothing about these systems explains what a priority *is*, or *why* certain rules should have higher priority. In default logic and circumscription, one merely asserts the priorities of rules, and nothing constrains these rankings to respect any intuitions. Neither of these systems provides an account of naturally occurring priorities.

Various explanations of priorities rely on the notion of *specificity* (Poole 1985). Suppose we have two default rules "birds fly" and "penguins don't." If something is known to be both a bird and a penguin, the conclusions sanctioned by these rules conflict. Most accounts claim that the rule "penguins don't fly" should be applied (or has a higher priority) because penguins are a specific subclass of birds and we prefer conclusions based on more specific information. In a probabilistic setting, this corresponds to making inferences based on the narrowest *reference class* (Bacchus 1990).[1]

While specificity seems to be an appropriate criterion for deciding priority, it wasn't until conditional theories of default reasoning were developed that specificity was put on a firm semantic basis. In particular, Pearl's (1990) System Z is a natural and compelling method of assigning "priorities" to default rules, possible worlds, and arbitrary formulae. It has commonly been understood that the Z-ranking of rules corresponds to the priorities of those rules. In this paper, we will show that this is not always the case. Rather, the priorities of rules are the Z-rankings of certain formulae, the negations of the *material counterparts* of these rules.

We can explain default priorities in terms of belief revision as well. When revising a theory or set of beliefs to include some new fact (say, some new information that has been learned), we are often forced to give up some of these original beliefs if the new information is inconsistent with the theory. The *epistemic entrenchment* of beliefs in a theory determines which of these should be given up and which should be kept when conflict arises (Gärdenfors 1988). We prefer to hold on to more entrenched beliefs. Default reasoning can be viewed as the

[1] However, other considerations may be involved in choosing the *appropriate* reference class (Kyburg 1983).

revision of a *theory of expectations* to accommodate the known facts (Gärdenfors and Makinson 1994; **?**). Adopting this perspective, we show that a default priority, as defined above, is nothing more than the *degree of entrenchment of the material counterpart of a rule in the theory of expectations*. We use the correspondence between defaults and expectations to propose an extension of Pearl's notion of 1-entailment, based on Brewka's (1989) *preferred subtheories*, that corrects certain deficiencies in its behavior with inheritance and independent conditionals. This approximates the idea of counting "weighted rule violations" prescribed by the maximum entropy principle (Goldszmidt, Morris and Pearl 1990), and is very similar to *conditional entailment* (Geffner and Pearl 1992).

In this paper, we will develop these connections within the unifying framework of the bimodal logic CO*. In (Boutilier 1991) this logic was shown to have the power to express *normative conditionals*, statements much like default rules, and capture solutions to the problem of irrelevance. In (Boutilier 1992c) we show that CO* is the first "classical" logic of *AGM belief revision* (Gärdenfors 1988). The results of this paper indicate that default reasoning can be viewed as a form of belief revision, a connection developed much further in (**?**). In Section 2 we very briefly review the logic CO* and define the normative conditional $\Rightarrow$, reading $A \Rightarrow B$ as "A normally implies B." In Section 3, we discuss belief revision and epistemic entrenchment, recall some results showing how CO* can represent these concepts, and show how it is related to default reasoning. In Section 4, we discuss System Z and some results showing that Pearl's system of ε-semantics is equivalent to a fragment of CO*. In Section 5, we demonstrate that the Z-ranking of a default rule is not always equivalent to its priority, but that the Z-ranking of a certain formula is. This turns out to be precisely the degree of entrenchment of this formula in a theory of expectations. This allows us to relate 1-entailment, a form of inference based on Z-ranking, to revision. In Section 6 we again use Z-ranking to determine priorities, but use these priorities in Brewka's (1989) model of Theorist to determine a more reasonable notion of consequence extending 1-entailment.

2 The Logic CO*

In (Boutilier 1990; Boutilier 1991), we presented modal logics in which we defined a conditional connective $\Rightarrow$, reading $A \Rightarrow B$ as "A normally implies B." While the original logics are standard modal systems, the logics CO and CO* of (Boutilier 1991) are bimodal logics with considerable expressive power. They can be used to axiomatize solutions to the problem of irrelevance that have typically required extra-logical machinery. These logics can also be used to characterize the classic AGM model of revision (Boutilier 1992c) described in the next section. We recall several definitions here, but refer the reader to these papers for further motivation and details.

We now review the Kripkean possible worlds semantics for logics of normality. The bimodal logic CO is based on a standard propositional modal language (over variables $\mathbf{P}$) augmented with an additional modal operator $\overleftarrow{\Box}$. The sentence $\overleftarrow{\Box}\alpha$ is read "α is true at all *inaccessible* worlds" (in contrast to the usual $\Box\alpha$ that refers to truth at accessible worlds).

Definition 1 A *CO-model* is a triple $M = \langle W, R, \varphi\rangle$, where W is a set (of possible worlds), R is a transitive, connected[2] binary relation on W (the accessibility relation), and φ maps $\mathbf{P}$ into 2^W ($\varphi(A)$ is the set of worlds where A is true).

Satisfaction is defined in the usual way, with the truth of a modal formula at a world defined as:

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

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 $\overleftrightarrow{\Diamond}\alpha \equiv_{\text{df}} \Diamond\alpha \vee \overleftarrow{\Diamond}\alpha$. 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 accessible (inaccessible) world; $\overleftrightarrow{\Box}\alpha$ ($\overleftrightarrow{\Diamond}\alpha$) holds iff α holds at all (some) worlds. The logic CO is based on the following axioms and inference rules, and is complete for the class of CO-models:

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

K′ $\overleftarrow{\Box}(A \supset B) \supset (\overleftarrow{\Box}A \supset \overleftarrow{\Box}B)$

T $\Box A \supset A$

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

S $A \supset \overleftarrow{\Box}\Diamond A$

H $\overleftrightarrow{\Diamond}(\Box A \wedge \overleftarrow{\Box}B) \supset \overleftrightarrow{\Box}(A \vee B)$

Nes From A infer $\overleftrightarrow{\Box}A$.

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

We often want to insist that all logically possible worlds be contained in our set of worlds W. This gives us the extension of CO called CO*.

Definition 2 A *CO*-model* is any $M = \langle W, R, \varphi\rangle$, such that M is a CO-model and $\{f : f \text{ maps } \mathbf{P} \text{ into } \{0,1\}\} \subseteq \{w^* : w \in W\}$.[3]

This class of models is characterized by the logic CO*, the smallest extension of CO containing the following:

[2] R is (totally) *connected* if wRv or vRw for any $v, w \in W$ (this implies reflexivity).

[3] For all $w \in W$, w^* is defined as the map from $\mathbf{P}$ into $\{0,1\}$ such that $w^*(A) = 1$ iff $w \in \varphi(A)$; in other words, w^* is the valuation associated with w.

LP $\overset{\leftrightarrow}{\Diamond}\alpha$ for all satisfiable propositional α.

In order to define a normative conditional, we impose the following interpretation on the accessibility relation R: world v is accessible to w (wRv) iff v is *at least as normal* as w. Thus, R is an ordering of situations respecting the degree to which an agent judges them to be "normal" or unexceptional. The truth conditions for $A \Rightarrow B$ can be stated as "In the most normal situations in which A holds, B holds as well."[4] This condition can be expressed as

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

3 Revision and Entrenchment

In this section we give a sparse description of the main ideas behind belief revision, referring readers to Gärdenfors (1988) for a comprehensive presentation of, and motivation for, work in the area.

Most work on belief revision models belief sets as deductively closed sets of sentences. We will use K to denote arbitrary belief sets, and if $K = Cn(KB)$ for some finite set KB, we will usually refer to the revision of K as the revision of its base set KB. Revising a belief set K is required when new information is learned and must be accommodated with these beliefs. If $K \not\models \neg A$, learning A is relatively unproblematic. More troublesome is revision when $K \models \neg A$ as some beliefs in K must be given up. The problem is in determining which part of K to give up, as there are a multitude of choices. Choosing which of these alternative revisions is acceptable depends largely on context. Fortunately, there are some logical criteria for reducing this set of possibilities.

The main criterion for discarding some revisions in deference to others is that of *minimal change*. Informational economy dictates that as few beliefs as possible from K be discarded in order to facilitate belief in A (Gärdenfors 1988). While pragmatic considerations will often enter into these deliberations, the main emphasis of the work of Alchourrón, Gärdenfors and Makinson (1985) is in logically delimiting the scope of acceptable revisions. They propose a set of eight postulates maintained to hold for any reasonable notion of revision (see e.g. (Gärdenfors 1988)). A revision function $*$ maps a belief set K and a proposition A to another belief set K^*_A, the result of revising K by A.

While these postulates describe logical constraints on revision, it is often the case when revising K we are more willing to give up certain sentences than others. This is referred to as *epistemic entrenchment*, and we say A is no more entrenched than B ($A \leq_E B$) if we are at least as willing to give up A as B when revising theory K. In (Gärdenfors 1988) postulates for guiding the revision of K are presented that any reasonable notion of epistemic entrenchment should satisfy. He also shows that a revision operator satisfies the eight AGM postulates iff it respects the following postulates for entrenchment:

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

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

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

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

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

The revision operator $*$ and the entrenchment ordering are related by the identity

$$B \in K^*_A \quad \text{iff} \quad A \supset \neg B <_E A \supset B$$

Gärdenfors and Makinson (1994) have also shown how belief revision can determine a nonmonotonic consequence relation. We describe a specific instance of this construction applied to a default theory. We can think of default rules as corresponding to *expectations* about the world. This view is adopted in Poole's (1988) Theorist framework, where default inference is effected by considering maximal subsets of such expectations (or "hypotheses") that are consistent with the known facts. Let T be a set of default rules or expectations; for instance,

$$T = \{\texttt{penguin} \supset \texttt{bird}, \texttt{bird} \supset \texttt{fly}, \texttt{penguin} \supset \neg\texttt{fly}\}.$$

If we took this theory at face value, we could never add `penguin` on threat of inconsistency. However, revising T by `penguin` allows us to give up some of the "default rules", which allow exceptions by their very nature. In general, we say that B is a nonmonotonic consequence of A (given default theory T) if $B \in T^*_A$. This is shown to be a meaningful notion (Makinson and Gärdenfors 1990), and bears a close relationship to Theorist. Unfortunately, no notion of priority or specificity is sanctioned by these considerations alone. We examine this idea that default reasoning is the process of revising such a theory of "expectations" in Section 5, and how priorities can be determined naturally.

CO* can also be used to represent AGM revision as discussed in (Boutilier 1992c). We omit details here, but note that we can define a subjunctive conditional in CO* that captures precisely the AGM revision postulates. The connection to revision is made via the *Ramsey test*, whereby a subjunctive $A > B$ is true with respect to a given state of belief K just when revising K by A results in a belief in B. Not surprisingly, the subjunctive and normative conditionals are identical in CO* and in (**?**) we pursue this connection, further establishing the correlation between default reasoning and belief revision that is our concern here.[5]

[4] This is only a rough formulation, for it presupposes the *Limit Assumption*, which is not a property required by our definition. There need not be a set of *most* normal worlds satisfying A. The conditional $A \Rightarrow B$ is still meaningful in this circumstance (Boutilier 1991; Boutilier 1992c).

[5] For the interested reader, we also describe how the Gärdenfors *triviality result* is avoided in (**?**).

Naturally, we can also capture the entrenchment of beliefs in CO*. Assuming a particular theory K is determined by a CO*-model M (see (Boutilier 1992c) for details), we can define $\leq_{EM}$, the *entrenchment ordering determined by* M, as

$$B \leq_{EM} A \text{ iff } M \models \ddot{\Box}(\neg A \supset \Diamond \neg B).$$

In (Boutilier 1992b) we show this ordering respects the postulates (E1) through (E5), and that any entrenchment ordering is representable in a CO*-model.

4 System Z

One problem that has plagued default reasoning is that of priorities, as manifested in the above example. Revising T by P (using P, B, and F as the obvious abbreviations) provides no guidance as to which of the two possible resulting theories should be preferred. Intuitively, we ought to give up $B \supset F$, since P is logically stronger, or *more specific* than B. That is, default $B \supset F$ has *lower priority*, or is more readily violated, than the others. Priorities have a long tradition in default reasoning (e.g., (McCarthy 1986)), but the most natural account seems to be that of Pearl.

Pearl (1990) describes a natural ordering on default rules named the *Z-ordering*, and uses this to define a nonmonotonic entailment relation, 1-entailment, put forth as an extension of ε-semantics (Pearl 1988). The default rules r of (Pearl 1990) have the form $\alpha \rightarrow \beta$, where α and β are propositional. We say a valuation (possible world) w *verifies* the rule $\alpha \rightarrow \beta$ iff $w \models \alpha \wedge \beta$, *falsifies* the rule iff $w \models \alpha \wedge \neg\beta$, and *satisfies* the rule iff $w \models \alpha \supset \beta$. For any rule $r = \alpha \rightarrow \beta$, we define r^* to be its *material counterpart* $\alpha \supset \beta$. We assume that T is a finite set of such rules (and when the context is clear, we take T to refer also to the set of material counterparts).

Definition 3 (Pearl 1990) T *tolerates* $\alpha \rightarrow \beta$ iff there is some world that verifies $\alpha \rightarrow \beta$, and falsifies no rule in T; that is, $\{\alpha \wedge \beta\} \cup \{\gamma \supset \delta : \gamma \rightarrow \delta \in T\}$ is satisfiable.

Toleration can be used to define a natural ordering on default rules by partitioning T as follows:

Definition 4 (Pearl 1990) For all $i \geq 0$ we define $T_i = \{r : r \text{ is tolerated by } T - T_0 - T_1 - \cdots T_{i-1}\}$

Assuming T is ε-consistent (see below), this results in an ordered partition $T = T_0 \cup T_1 \cup \cdots T_n$. Now to each rule $r \in T$ we assign a rank (the *Z-ranking*), $Z(r) = i$ whenever $r \in T_i$. Roughly, but not precisely (see below), the idea is that lower ranked rules are more general, or have lower priority. Given this ranking, we can rank worlds according to the highest ranked rule they falsify:

$$Z(w) = \min\{n : w \text{ satisfies } r, \text{ for all } r \in T \text{ such that } Z(r) \geq n\}.$$

Again, lower ranked worlds are to be considered more normal. Now any propositional α can be ranked according to the lowest ranked world that satisfies it; that is

$$Z(\alpha) = \min\{Z(w) : w \models \alpha\}.$$

Given that lower ranked worlds are considered more normal, we can say that a default rule $\alpha \rightarrow \beta$ should hold iff the rank of $\alpha \wedge \beta$ is lower than that of $\alpha \wedge \neg\beta$. This leads to the following definition:

Definition 5 (Pearl 1990) Formula β is *1-entailed* by α with respect to T (written $\alpha \vdash_1 \beta$) iff $Z(\alpha \wedge \beta) < Z(\alpha \wedge \neg\beta)$ (where Z is determined by T).

We will see some examples of the types of conclusions sanctioned by 1-entailment in Section 6, but refer to (Pearl 1990) for further details.

The default rules ordered by Z-ranking are usually assumed to be statements of high probability, based on Pearl's (1988) ε-semantics. A set of rules T is *ε-consistent* if each rule can be consistently assigned a conditional probability no less than $1 - \varepsilon$ for arbitrary ε approaching 0. The theory T *ε-entails* a rule $\alpha \rightarrow \beta$ if the conditional probability of β given α can be made arbitrarily high merely by making the probabilities of each rule in T achieve some threshold. We can show that this logic of default rules, based on arbitrarily high probabilities, corresponds exactly to the fragment of CO* consisting of simple conditional sentences of the form $\alpha \Rightarrow \beta$. Assume T is a finite set of simple default rules.[6]

Theorem 1 *T is CO*-consistent iff T is ε-consistent.*

Theorem 2 *$T \models_{CO*} A \Rightarrow B$ iff T ε-entails $A \rightarrow B$.*

Semantically, the equivalence of normative conditional inference in CO* and ε-semantics can be seen by examining Pearl's (1990) construction used to determine the ε-consistency of a set of rules (see also Adams (1975)) and equating more probable worlds with more normal worlds in the sense of CO*.

As shown in (Boutilier 1991), the notion of 1-entailment can be axiomatized in CO*. As discussed there, for any theory T there exists a unique CO*-model Z_T corresponding precisely to the Z-ranking of worlds according to T (Z_T is defined by the identity vRw iff $Z(v) \geq Z(w)$). We recall that $Z_T \models A \Rightarrow B$ iff $A \vdash_1 B$, for any given T. Thus, any semantic or syntactic results regarding Z-ranking and 1-entailment can be applied to simple conditional theories in CO*. In what follows, we take default rules to be normative conditionals in CO*.

[6] We take a rule to be either $\alpha \rightarrow \beta$ or $\alpha \Rightarrow \beta$, depending on context. We also assume each antecedent α is satisfiable, for simplicity, but these results can be restated for the more general case (Boutilier 1992a). Proofs may also be found in (Boutilier 1992a).

5 Priorities as Entrenchment

We saw that revision in its simplest form could not account for priorities on conflicting defaults. However, it seems clear that some notion of epistemic entrenchment could characterize this feature. Let T be a set of expectations. As in the Theorist framework, or the Gärdenfors-Makinson revision model of nonmonotonic logic, facts (nonmonotonically) derivable from premise A are those sentences (classically) derivable from $\{A\} \cup T'$, where $T' \subseteq T$ is some maximal subset of T consistent with A. This corresponds to having an initial belief set $K = Cn(T)$ and revising K with new facts A, keeping as much of K (or more precisely, T) as possible. As we saw earlier, there can be several choices for T', but some are preferable to others. In default reasoning, these preferences are expressed as priorities on default rules: certain rules should not be applied in deference to others in the case of conflict. In revision, preferences are represented by epistemic entrenchment: certain sentences (in this case defaults) are more likely to be given up when revising than others.

Given this parallel, the question remains: do priorities correspond to entrenchments? The most obvious proposal is to associate priorities of default rules with their Z-rankings. In most naturally occurring sets of defaults (or rather, most naturally occurring "examples") this proposal is adequate. However, the following example quickly shows that the Z-ranking of rules, $Z(r)$, cannot be viewed as entrenchment of the corresponding material conditionals r^* in a default theory.

Theorem 3 *Let T be a default theory with $r_1, r_2 \in T$. Let $r_1^* \leq_E r_2^*$ iff $Z(r_1) \leq Z(r_2)$. Then $\leq_E$ will not, in general, satisfy (E1)–(E5).*

Proof As a counterexample, consider T consisting of the following four rules:

r_1 : $(p \wedge q) \Rightarrow x$
r_2 : $c \Rightarrow (\neg p \vee \neg q \vee x)$
r_3 : $p \Rightarrow (\neg c \vee q)$
r_4 : $p \Rightarrow \neg x$

It is easy to show that all rules have rank 0, except r_1, which has rank 1. A verifying assignment for r_3, r_4 is $\{p, \neg q, \neg x, \neg c\}$, and for r_2 is $\{\neg p, c\}$, while any assignment verifying r_1 must falsify r_4. On this definition of entrenchment, $r_2^* <_E r_1^*$. However, $r_1^* \vdash r_2^*$, violating postulate (E2). ■

While Z-ranking of rules is not a coherent notion of entrenchment for our theory of expectations, we observe that the Z-ranking of *formulae* does in fact satisfy postulates corresponding to the notion of *plausibility*, the dual of entrenchment (Grove 1988; Gärdenfors 1988), and leads us to the following definition. We assume a background set of rules T and propositional A, B.

Definition 6 Let T be a default theory. We say A is *no more entrenched* (with respect to T) than B (written $A \leq_{ET} B$) iff $Z(\neg A) \leq Z(\neg B)$.

Recall that Z_T is the unique CO*-model respecting the Z-ranking of worlds determined by T.

Theorem 4 *Let $\leq_{EM}$ be the entrenchment ordering determined by Z_T. Then $\leq_{ET}$ is identical to $\leq_{EM}$.*

Corollary 5 *The relation $\leq_{ET}$ satisfies (E1)–(E5).*

What does this say about priorities on default rules? Clearly a formula is less entrenched than another if the Z-ranking of its negation is less than that of the other. This means we are prepared to violate default rules according to the following definition of priorities:

Definition 7 Let T be a default theory, $r_1, r_2 \in T$. We say r_1 has *no priority over* r_2 ($r_1 \preceq r_2$) iff $Z(\neg r_1^*) \leq Z(\neg r_2^*)$. If $r_1 \preceq r_2$ and not $r_2 \preceq r_1$, then r_1 has *lower priority* than r_2 ($r_1 \prec r_2$).

This notion of priority is consistent with the view that we will satisfy defaults with higher priorities in the case of conflict when determining the consequence relation of 1-entailment, as substantiated by the following theorem.

Theorem 6 *Let T be a default theory and let $*$ be the revision function determined by the ordering of entrenchment $\leq_{ET}$. Then $A \vdash_1 B$ iff $B \in T_A^*$.*

Corollary 7 *$B \in T_A^*$ iff $A \supset \neg B <_{ET} A \supset B$*

Example Let T contain the following conditionals:

$$P \Rightarrow B,\ B \Rightarrow F,\ P \Rightarrow \neg F,\ B \Rightarrow W$$

where we read P, B, F, W and G as "penguin," "bird," "fly," "has-wings," and "green" respectively. Using CO* alone we can derive

$$B \wedge P \Rightarrow \neg F,\ F \Rightarrow \neg P,\ B \Rightarrow \neg P.$$

Using 1-entailment we can derive further:

$$\neg B \Rightarrow \neg P,\ \neg F \Rightarrow \neg B,\ G \wedge B \Rightarrow F,\ P \wedge \neg W \Rightarrow B.$$

In our theory of expectations we have, for instance, both $P \supset F$ and $P \supset \neg F$ (since $\neg P$ is also an expectation). Since $P \supset \neg F$ is more entrenched (under Z-ranking) than $P \supset F$, the latter is given up when revising our expectations to include P, or equivalently, asking for the default consequences of P. Hence, $P \vdash_1 \neg F$.

Unfortunately, certain intuitively desirable conclusions cannot be reached through 1-entailment, in particular $P \Rightarrow W$. We turn to such difficulties in the next section.

Thus, we see that 1-entailment can be modeled using a revision function that satisfies entrenchment of formulae respecting the Z-ordering. That the Z-ordering of the negations of material counterparts of rules determines natural priorities (as specified by the definition given above), rather than the Z-ranking of the rules themselves, should be obvious given the following corollary.

r	$B \Rightarrow F$	$P \Rightarrow \neg F$	$P \Rightarrow B$
$Z(r)$	0	1	1
$Z(\neg r^*)$	1	2	2

Figure 1: "Common" example.

r	r_1	r_2	r_3	r_4
$Z(r)$	1	0	0	0
$Z(\neg r^*)$	2	2	1	1

Figure 2: "Uncommon" example.

Corollary 8 *For any i, if $\{A\} \cup \{r^* : Z(r^*) \geq i\}$ is consistent, then $A \vdash_1 r^*$ for each r^* such that $Z(r^*) \geq i$.*

This shows that if the set of rules above a certain priority threshold is consistent with some set of premises A, each of these rules is "applied" when 1-entailment is used. Thus rules are satisfied according to the priority determined by the Z-ranking of their negated material counterparts; in other words, the priority determined by the degree of entrenchment of their counterparts in the theory of expectations

$$\{A \supset B : A \Rightarrow B \in T\}$$

To the extent that these priorities are representative of default priorities in general, we can state that default priorities correspond accurately to the epistemic entrenchment of the corresponding material conditionals within a default theory. Why the Z-ranking of rules doesn't correspond to priorities is demonstrated, using our previous examples, in Figures 1 and 2. In the first, we see that "common" examples of sets of rules often satisfy the property $Z(r) = Z(\neg r^*) - 1$. Whenever we ask for the consequences of α using 1-entailment, we must inspect the set of minimal (in Z-rank) α-worlds. Given this relationship between $Z(r)$ and $Z(\neg r^*)$ we notice that default rules are "given up" in the order of their Z-ranking. However, this connection does not always hold, as evidenced by Figure 2. Using the example from Theorem 3 we see that $Z(r_2) = Z(\neg r_2^*) - 2$; so, while rule r_1 has a higher Z-ranking than r_2, r_2 cannot have lower priority than r_1, since it cannot be given up (i.e., falsified) without giving up r_1. Whenever we apply r_1, rule r_2 automatically "follows". We have determined that while the Z-ranking of default rules is a natural ordering, it cannot, in general, be viewed as a priority ranking. Instead, it induces priorities, by associating ranks with formulae, priorities corresponding to the Z-ranking of the negation of the material counterparts of rules. This view of priorities is also in concordance with the notion of epistemic entrenchment of rules within a default theory, or theory of expectations.

6 Preferred Subtheories

The Z-ranking of rules provides a very natural and compelling method of determining default priorities. The preference for more specific default rules is put on a firm semantic basis in a conditional framework and Z-ranking reflects this. To take our standard set of three default rules ($P \Rightarrow B$, $B \Rightarrow F$, $P \Rightarrow \neg F$), given the knowledge $P \wedge B$, we could potentially choose to use either the second rule or the third, but not both. In CO* (indeed, in most conditional logics for default reasoning) $P \wedge B \Rightarrow \neg F$ is derivable from this rule set, indicating a preference for the third rule. This can be explained as follows: at the most normal P-worlds, B and $\neg F$ must both hold. Since the most normal $P \wedge B$-worlds cannot be *more* normal than these P-worlds, this set is also the set of most normal $P \wedge B$-worlds. Since $\neg F$ holds at each of these, the conditional holds. Of course, these cannot be the most normal B-worlds due to the constraint $B \Rightarrow F$. Thus, the most normal B-worlds must be strictly more normal than these P-worlds (hence $B \Rightarrow \neg P$ is derivable as well). The Z-ranking of a rule indicates the degree of normality of the most normal worlds that can *confirm* that rule. The rules with P in the antecedent necessarily have a higher ranking (are confirmed by less normal worlds) than those with B as a head.

The notion of 1-entailment is determined by the Z-ranking of rules, and is based on the intuition that worlds should be considered as normal as possible subject to the constraints imposed by the rules. The Z-ranking of rules induces a ranking of worlds (or ordering of normality) through the definition of $Z(w)$. While this seems to determine a reasonable notion of consequence, certain classes of examples are not treated appropriately using 1-entailment. This problem has been identified in (Goldszmidt, Morris and Pearl 1990; Geffner and Pearl 1992). One problem is with the default inheritance of properties from superclasses. The example in the last section illustrates this phenomenon. The conclusion W (wings) is not derivable given P (penguin) even though we should expect (default) transitivity through the class B (bird). This can be explained by observing that the most normal P-worlds must violate the rule $B \Rightarrow F$ and must be given a Z-rank of 1 rather than 0 (most normal). However, any world w_1 satisfying $P \wedge B \wedge \neg F \wedge \neg W$ is given the same rank as a world w_2 satisfying $P \wedge B \wedge \neg F \wedge W$. While w_1 violates both $B \Rightarrow F$ and $B \Rightarrow W$, the Z-ranking of w_1 is determined by the *maximum* rank of the set of rules it violates. Once the rank 1 rule $B \Rightarrow F$ is violated (as in w_2), violation of a further rank 1 rule $B \Rightarrow W$ (as in w_1) incurs no additional penalty. In terms of entrenchment in the induced default theory T, the expectations $P \supset W$ and $P \supset \neg W$ are equally entrenched. A related class of examples are those containing *independent conditionals*.

Example Consider two independent defaults $R \Rightarrow W$ (if it's raining I walk to school) and $F \Rightarrow M$ (if

it's Friday I have a meeting). From the knowledge $R \wedge \neg W \wedge F$ one cannot conclude via 1-entailment that M is true, even though the violation of the first default, $R \wedge \neg W$, should not prevent application of the second. Since both rules have rank 0, 1-entailment assumes that violating both rules makes a world no less normal than violating one rule.

The counterintuitive results provided by 1-entailment in each of these examples is due to its insistence on making worlds as normal as possible. This ensures that a world violating many rules of a certain rank is no less normal than a world violating a single rule of that rank, as reflected in the definition of $Z(w)$. Forcing such worlds to be as normal as their less objectionable counterparts (those violating single rules) will not effect the satisfaction of each default rule: violating one rule already ensures that a world is considered abnormal and cannot be used to confirm rules of that rank. While 1-entailment considers only the "quality" of violated default rules in its ranking of worlds, it seems natural to consider also the number of violated rules. This has been suggested by Goldszmidt, Morris and Pearl (1990) in their *maximum entropy* proposal, and in Pearl and Geffner's (1992) *conditional entailment*.

In a priority-free setting Poole's (1988) Theorist framework can be viewed as counting rule violations. Given a set of default expectations D, the default conclusions based on a set of facts F are determined by adding to F some maximal subset S of D where $F \cup S$ is consistent. As described earlier, this can seen as the revision of D to include F, but unfortunately does not allow for priorities on default rules, or the entrenchment of expectations in D. Brewka (1989) has presented a simple generalization of Theorist in which the set of defaults D is partitioned to reflect priorities. Specifically

$$D = D_0 \cup D_1 \cup \cdots D_n$$

where each D_i is a set of propositional formulae (default expectations) such that the defaults in D_i are preferred to, or have priority over, those in D_j just when $i < j$. In particular, we take D_0 to be a consistent set of *premises* which will not be violated.[7] Just as Theorist takes maximal consistent subsets of D, Brewka proposes *preferred subtheories* of D:

$$S = D_0 \cup S_1 \cup \cdots S_n$$

is a preferred subtheory of D iff $D = D_0 \cup S_1 \cup \cdots S_k$ is a maximal consistent subset of $D = D_0 \cup D_1 \cup \cdots D_k$ for $1 \leq k \leq n$. In other words, we add to D_0 as many formulae from D_1 as possible without forcing inconsistency, then add to these defaults from D_2, and so on.

[7]Brewka "prohibits" D_0, maintaining that by allowing the most reliable set of formulae (premises) to be inconsistent Theorist is generalized. However, if consistent premises are not required, this is easily captured by postulating an empty set D_0 (or F in the case of Theorist).

The problematic aspect of Brewka's theory is that little indication of the source of these priorities (the partitioning of D) is given (although he does provide a syntactic mechanism for determining specificity). In general, any partitioning in permitted even though some of these are effectively useless. For instance, placing $B \supset F$ in D_1 and $B \wedge P \supset \neg F$ in D_2 will ensure that the second default is never applied to derive $\neg F$. System Z provides a natural ranking of rules that would seem to determine just the "priorities" needed for Brewka's partitioning. However, Brewka's notion of preferred subtheory automatically accounts for the intuitive preference that as many default rules as possible of a certain priority level be applied. This is due to the maximality condition on the subsets S_i, and stands in sharp contrast with 1-entailment. It should be a simple exercise to combine the two notions.

Let T be a consistent set of default rules or conditionals in CO*, such that T is partitioned as $T = T_0 \cup T_1 \cup \cdots T_n$. Thus, the highest ranked rules have a rank of n. Roughly, these rules should have higher priority than the others, followed by rank $n-1$ rules, and so on. The corresponding set of expectations D (the material counterparts of T) should be partitioned similarly.

Definition 8 Let T be a consistent set of simple conditionals in CO* with a maximum Z-rank of n. The *Brewka theory* D_B of expectations corresponding to T is given by $D_B = D_1 \cup \cdots \cup D_{n+1}$ where

$$D_i = \{\alpha \supset \beta : \alpha \Rightarrow \beta \in T_{n+1-i}\}$$

So, e.g., D_1 consists of the counterparts of the rank n rules, while D_{n+1} corresponds to rank 0 rules.

We can also define a skeptical consequence relation, called *B-entailment*, using the Brewka theory determined by T simply by considering what holds in all preferred subtheories of D_B. If α is a premise, it should be added to D_B in the role of D_0.

Definition 9 Let T be a set of conditionals and $D_B = D_1 \cup \cdots \cup D_{n+1}$ the corresponding Brewka theory. We say α *B-entails* β with respect to T (written $\alpha \vdash_B \beta$) iff β is entailed by all preferred subtheories of (setting $D_0 = \{\alpha\}$)

$$D_{B+\alpha} = \{\alpha\} \cup D_1 \cup \cdots \cup D_{n+1}.$$

It should be clear that asking for the consequences of the theory D_B will not correspond to 1-entailment.

Example Consider theory T from our earlier example. We saw that $P \vdash_1 W$ is not sanctioned by T. The Brewka theory D_B contains two partitions: $D_1 = \{P \supset B, P \supset \neg F\}$ and $D_2 = \{B \supset F, B \supset W\}$. The unique preferred subtheory of $D_{B+\alpha}$ where $\alpha = P$ will contain all expectations except $B \supset F$. In particular, even though the expectations in D_1

cause the violation of $B \supset F$, the other expectation in D_2 is consistent and will be satisfied, unlike 1-entailment. Hence, $P \vdash_B W$ is sanctioned by T.

A similar analysis shows that B-entailment provides more intuitive results on our example containing independent conditionals.

The Z-ranking of rules can be used to determine an ordering on possible worlds that captures 1-entailment. The definition of $Z(w)$ induced by the ranking of rules is characterized by the unique CO*-model Z_T, which in turn satisfies a conditional $\alpha \Rightarrow \beta$ just when $\alpha \vdash_1 \beta$. However, it is not an intrinsic property of the Z-ranking of rules that causes the problems in 1-entailment we examined above. Rather it is the induced ranking of worlds and the model Z_T. Indeed, Z-ranking can be used to determine different orderings on worlds, or different CO*-models. In particular, we can define a ranking of worlds, or CO*-model, that satisfies the conditionals corresponding to B-entailment, thus capturing some notion of counting rule violations in CO*. We must first introduce some terminology. We assume a fixed consistent set of conditionals T throughout, partitioned as usual.

Definition 10 For any valuation (world) w, the set of rules of rank i falsified by w is denoted

$$V_w^i = \{r \in T_i : w \text{ falsifies } r\}$$

Definition 11 For valuations w, v, let

$$\max(v, w) = \max\{i : V_w^i \subset V_v^i \text{ or } V_v^i \subset V_w^i\}$$

If the set above is empty, we let $\max(v, w) = -1$.

Thus, $\max(v, w)$ denotes the highest rule ranking such that the set of rules of this rank violated by w and v are such that one set is strictly contained in the other. We will use this quantity to rank worlds. If a world v violates a rule ranked higher than any rule violated by v, then v will be considered more normal. However, if this highest rank is the same for each world, v can be considered more normal is it violates fewer (with respect to set inclusion) rules of that rank than w.

Definition 12 The *Brewka model* of T, denoted $Z_T^B = \langle W, R, \varphi \rangle$, is defined as follows: we let W and φ be as usual, capturing the set of valuations appropriate for the our propositional language; we define R as

1. If $\max(v, w) = -1$ then vRw and wRv
2. If $V_w^{\max(v,w)} \subset V_v^{\max(v,w)}$, vRw but not wRv

Proposition 9 *Z_T^B is a CO*-model.*

Theorem 10 *$Z_T^B \models \alpha \Rightarrow \beta$ iff $\alpha \vdash_B \beta$.*

Thus, the Z-ranking of rules can be used to determine a notion of entailment in CO* that differs from 1-entailment and captures the idea that as many defaults as possible should be applied within a given priority threshold, even if certain rules cannot be applied. Naturally, the results of Section 5 can be applied directly to B-entailment as they were to 1-entailment. In particular, we can define an entrenchment ordering and revision function that corresponds to the notion of revising our expectations to effect default prediction.

Definition 13 Let T be a default theory. The entrenchment ordering for the purposes of B-entailment $\leq_{EB}$ is the entrenchment ordering $\leq_{EM}$ determined by the CO*-model Z_T^B.

Proposition 11 *The relation $\leq_{EB}$ satisfies (E1)–(E5).*

Theorem 12 *Let $*$ be the revision function induced by $\leq_{EB}$, and let D be the expectation set determined by conditionals T. Then $A \vdash_B B$ iff $B \in D_A^*$ iff $A \supset \neg B <_{EB} A \supset B$.*

Naturally, the priorities of default rules reflect the entrenchment of the corresponding material counterparts or expectations. Just as Theorem 3 shows that the Z-ranking of rules does not capture priorities precisely within the context of 1-entailment, the counterexample there also shows this to be the case for B-entailment. So while the "priority levels" of Brewka seem compelling, they do not provide a guarantee that rules (or more precisely, expectations) will be given up in the order specified by the partition. In particular, the ordering specified by the partition $D_1 \cup \cdots D_n$ will not, in general, be an entrenchment ordering; but the adopting the view that priorities correspond to entrenchment of expectations is justified, as it is quite easy to show that the analog of Corollary 8 holds for B-entailment (where Z-ranking of formulae is replaced by entrenchment using $\leq_{EB}$).

7 Concluding Remarks

We have shown that Z-ranking is a useful way of ranking rules, but that these ranks cannot generally be interpreted as priorities. Rather these induce entrenchment orderings on a theory of expectations, and revision of this theory corresponds to default prediction. Furthermore, Z-ranking need not be tied to 1-entailment, but can be used to induce priorities for other forms of entailment. We have presented one such notion, appealing to Brewka's preferred subtheories, and demonstrating its applicability on certain examples on which 1-entailment fails to behave appropriately. Brewka's model reflects many of the same intuitions as prioritized circumscription (McCarthy 1986) and our B-entailment bears a remarkable similarity to Geffner and Pearl's (1992) conditional entailment. In particular, both of these notions of consequence have the goal of minimizing violations of defaults, but prefer to satisfy any higher priority default at the expense of lower priority defaults. In circumscription, however, priorities must be specified independently.

Although we have not done so here, it should be easy to see how the Brewka model Z_T^B can be axiomatized in CO*, just as the model Z_T is axiomatized in (Boutilier 1991). However, the Theorist-style formulation suggests a straightforward conceptual and computational approach to B-entailment. While the notion of counting rule violations within a priority level is captured by B-entailment, this notion is somewhat different from the implicit "sum of weighted rule violations" of the maximum entropy formalism (Goldszmidt, Morris and Pearl 1990). Both maximum entropy and conditional entailment address certain difficulties with the priorities induced by System Z. An investigation of the differences with B-entailment should prove enlightening. The revision model of defaults might also be applied to these systems as well, 1-entailment and B-entailment simply being two examples of the use of priorities as entrenchment. This may illuminate important similarities and distinctions among these systems.

Acknowledgements

I'd like to thank Moisés Goldszmidt, Judea Pearl and David Poole for their helpful comments.

References

Adams, E. W. 1975. *The Logic of Conditionals*. D.Reidel, Dordrecht.

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

Bacchus, F. 1990. *Representing and Reasoning with Probabilistic Knowledge*. MIT Press, Cambridge.

Boutilier, C. 1990. Conditional logics of normality as modal systems. In *Proceedings of the Eighth National Conference on Artificial Intelligence (AAAI-90)*, pages 594–599, Boston.

Boutilier, C. 1991. Inaccessible worlds and irrelevance: Preliminary report. In *Proceedings of the Twelfth International Joint Conference on Artificial Intelligence (IJCAI-91)*, 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. Epistemic entrenchment in autoepistemic logic. *Fundamenta Informaticae*, 17(1–2):5–30.

Boutilier, C. 1992c. A logic for revision and subjunctive queries. In *Proceedings of the Tenth National Conference on Artificial Intelligence (AAAI-92)*, pages 609–615, San Jose.

Brewka, G. 1989. Preferred subtheories: An extended logical framework for default reasoning. In *Proceedings of the Eleventh International Joint Conference on Artificial Intelligence (IJCAI-89)*, pages 1043–1048, Detroit.

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

Gärdenfors, P. and Makinson, D. 1994. Nonmonotonic inference based on expectations. *Artificial Intelligence*, 65:197–245.

Geffner, H. and Pearl, J. 1992. Conditional entailment: Bridging two approaches to default reasoning. *Artificial Intelligence*, 53:209–244.

Goldszmidt, M., Morris, P., and Pearl, J. 1990. A maximum entropy approach to nonmonotonic reasoning. In *Proceedings of the Eighth National Conference on Artificial Intelligence (AAAI-90)*, pages 646–652, Boston.

Grove, A. 1988. Two modellings for theory change. *Journal of Philosophical Logic*, 17:157–170.

Kyburg, Jr., H. E. 1983. The reference class. *Philosophy of Science*, 50(3):374–397.

Lifschitz, V. 1985. Computing circumscription. In *Proceedings of the Ninth International Joint Conference on Artificial Intelligence (IJCAI-85)*, pages 121–127, Los Angeles.

Makinson, D. and Gärdenfors, P. 1990. Relations between the logic of theory change and nonmonotonic logic. In Fuhrmann, A. and Morreau, M., editors, *The Logic of Theory Change*, pages 185–205. Springer-Verlag, Berlin.

McCarthy, J. 1986. Applications of circumscription to formalizing commonsense reasoning. *Artificial Intelligence*, 28:89–116.

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

Pearl, J. 1990. System Z: A natural ordering of defaults with tractable applications to default reasoning. In Vardi, M., editor, *Proceedings of Theoretical Aspects of Reasoning about Knowledge*, pages 121–135. Morgan Kaufmann, San Mateo.

Poole, D. 1985. On the comparison of theories: Preferring the most specific explanation. In *Proceedings of the Ninth International Joint Conference on Artificial Intelligence (IJCAI-85)*, pages 144–147, Los Angeles.

Poole, D. 1988. A logical framework for default reasoning. *Artificial Intelligence*, 36:27–47.

Reiter, R. and Criscuolo, G. 1981. On interacting defaults. In *Proceedings of the Seventh International Joint Conference on Artificial Intelligence (IJCAI-81)*, pages 270–276, Vancouver.