Inaccessible Worlds and Irrelevance: Preliminary Report

Craig Boutilier
Department of Computer Science
University of Toronto
Toronto, Canada M5S 1A4
Email: cebly@ai.toronto.edu

Abstract

Recently, the relationship between several forms of default reasoning based on conditional defaults has been investigated. In particular, the systems based on ε-semantics, preferential models, and (fragments of) modally-defined conditional logics have been shown to be equivalent. These systems form a plausible core for default inference, but are too weak in general, failing to deal adequately with irrelevance. We propose an extension of the (modal) conditional logics in which one can express the truth of sentences at inaccessible possible worlds and show how this logic can be used to axiomatize a simple preference relation on the modal structures of this logic. This preferential semantics is shown to be equivalent to 1-entailment and rational closure. We suggest that many meta-logical systems of default inference can be axiomatized within this logic, using the notion of inaccessible worlds.

1 Introduction

Recently, the focus of much research on default reasoning has centered on the representation of default rules as conditional sentences in various conditional logics. While the motivation and underlying semantics for these logics often diverge, most allow "In the most normal situations in which A is true, B is true" as a profitable interpretation of the conditional $A \rightarrow B$. Each of these logics can be viewed as enforcing some type of normality ordering on states of affairs, or possible worlds. For instance, the logic of ε-semantics (Adams 1975; Pearl 1988) is based on a probability distributions over sets of worlds, and a conditional $A \rightarrow B$ is true iff B is true at the most probable (as ε approaches 0) A-worlds; hence, more probable worlds can be viewed as more normal under such an ordering. The preferential logics of (Kraus, Lehmann and Magidor 1990) embody an explicit normality ordering on situations, and the conditional logics of (Boutilier 1990) are similar, incorporating normality as a modal accessibility relation.

Given that these logics can be regarded in such a similar fashion, it is not surprising that they have been shown to be essentially equivalent (Boutilier 1990; Kraus, Lehmann and Magidor 1990), giving credence to their underlying motivation. Unfortunately, while each can be considered a "core" for default reasoning (Pearl 1989), they are too weak to sanction all of the patterns of default inference we would like. In particular, the conclusions authorized by such systems are often rendered invalid in the face of irrelevant information. For example, given that birds fly ($B \rightarrow F$) we cannot conclude that green birds fly ($G \wedge B \rightarrow F$).

To circumvent such difficulties, many schemes have been proposed which augment the logics under consideration with extra-logical machinery for deriving the appropriate conclusions (Delgrande 1988; Lehmann 1989; Pearl 1990; Goldszmidt, Morris and Pearl 1990). Goldszmidt and Pearl (1990) have shown that Pearl's 1-entailment corresponds to Lehmann's rational closure, demonstrating that these reasonable extensions (as well as the logical cores) of ε-semantics and preferential logics determine the same default conclusions.

We can view these logics as being able to express what is true at "more normal" possible worlds; hence, sentences can force worlds (which don't satisfy these constraints) to be less normal. The meta-logical extensions of these systems, conversely, attempt to force worlds to be more normal than is required. In this paper, we will show that the conditional logics of (Boutilier 1990) can also be extended in a manner which corresponds exactly to these systems. This extension is based on a simple preference relation over modal structures, one that prefers structures in which possible worlds are as normal as possible. Furthermore, we extend the logic itself such that we can make reference to inaccessible worlds in addition to accessible worlds; thus we can express truth at less normal worlds and force worlds to be more normal. With this capability, we can axiomatize a default theory making the derivable conclusions exactly those sanctioned by the preferential semantics. Just as the

second-order circumscription axiom conforms to truth in (predicate) minimal models (McCarthy 1986), so too does this closure correspond to preferred models. We suggest that, to the extent any meta-logical default system forces worlds to be more normal, it can be axiomatized within this extended conditional logic.

A more detailed presentation and proofs may be found in (Boutilier 1991c).

2 Inaccessible Worlds and the Logic CO

As in (Boutilier 1990), we will present a Kripke-style possible worlds semantics for a conditional logic capable of representing and reasoning with statements of normality or prototypicality. A sentence $A \Rightarrow B$ is intended to mean "A normally implies B". Following a suggestion of Delgrande (1988), we will interpret the truth conditions for such a statement roughly as "In the most normal situations in which A holds, B is true as well". The ordering of normality presupposed by such a reading will be represented as an accessibility relation on possible worlds; world v is accessible to w (wRv) iff v is *at least as normal as* w. In (Boutilier 1990), it is argued that such a relation should be, at a minimum, reflexive and transitive, and that a further restriction of total-connectedness[1] gives rise to a reasonable extension, namely the logic CT4D. It is also shown that CT4D is equivalent to the standard modal logic S4.3, in the sense that the conditional connective $\Rightarrow$ can be defined in terms of the modal operator $\Box$, and vice versa. In the sequel, we will take the modal connective to be basic and define the conditional within the modal language.

In (Boutilier 1990) it is shown that CT4D captures many of the properties we expect of normal implication, such as allowing exceptions to prototypical statements, and capturing rules such as cautious and rational monotony. It is also shown that (a fragment of) CT4D is equivalent to the logic of ranked preferential models of (Lehmann 1989).

The following approach to default reasoning using CT4D immediately presents itself. Let KB be (the conjunction of) a finite knowledge base of sentences of CT4D. It is reasonable to ask what is true at the most normal worlds in which all the facts in KB hold; that is a default reasoner could conclude α whenever $\vdash_{CT4D} KB \Rightarrow \alpha$ (see also Delgrande's (1988) *Assumption of Normality*). However, a serious problem arises when we consider certain classes of default inferences, specifically those involving irrelevant properties. If KB consists of two facts, $Bird$ and $Bird \Rightarrow Fly$, then $KB \Rightarrow Fly$ is derivable. However, if we add $Green$ to KB, then $KB \Rightarrow Fly$ is no longer a theorem, for the most normal $Green \wedge Bird$-worlds need not satisfy Fly, as long as these are not among the most normal $Bird$-worlds. In other words, greenness may be an exceptional property of birds (with respect to flying), as "penguinness" is.

The problem of irrelevance has been addressed using a number of extra-logical techniques, such as rational closure, assumptions of irrelevance (Delgrande 1988), and 1-entailment. We will approach the problem from a perspective which may in the future lead to a purely logical account of default inference, and, at present, provides some new insights into irrelevance.

Consider some normality ordering on all possible states of affairs, W, and refer to the most normal A-worlds in this ordering as $n(A)$. Then $A \Rightarrow C$ is true iff $n(A) \subseteq \|C\|$.[2] Of course, A must hold at all such worlds, so we can write this as $n(A) \subseteq \|A \wedge C\|$. Now, to say that B is irrelevant to the truth of this conditional is to say (roughly) that $A \wedge B \Rightarrow C$ is true. For this to be the case it is sufficient to insist that some B-world exist among those in $n(A)$, making this a most normal $A \wedge B$-world. In general, we want this to be true for arbitrary properties α (consistent with $A \wedge C$), so that $A \wedge \alpha \Rightarrow C$ holds. Hence, we need only insist that $\|A \wedge C\| \subseteq n(A)$. Together with the converse, this implies that $A \wedge \alpha \Rightarrow C$ for any such α. In other words, we would like to assume, if we know $A \Rightarrow C$, that the most normal A-worlds are *exactly* all $A \wedge C$-worlds.

These constraints are analogous to those used by Levesque (1990) to define the semantics of OL, the logic of "only knowing". In an entirely similar fashion, $A \Rightarrow C$ can be read as "at the most normal A-worlds, *at least* $A \wedge C$ is known".[3] Supposing a new connective $>$, we say $A > C$ is true iff $\|A \wedge C\| \subseteq n(A)$. We can read this, then, as "at the most normal A-worlds, *at most* $A \wedge C$ is known". Together, $A \Rightarrow C$ and $A > C$ tell us that exactly $A \wedge C$ is known at the most normal A-worlds, and allow us to conclude that all (consistent) properties are irrelevant. This captures the intuition that if some fact (other than A) were relevant to concluding C we would know this to be the case. Since we "haven't been told", we assume nothing else should affect our deliberations.

In order to formalize this discussion, we will provide a semantics and axiomatization for the connectives $\Rightarrow$ and $>$. As mentioned, we can define $\Rightarrow$ in terms of the standard modal operator $\Box$. However, in a similar manner we can define $>$ in terms of a new modal connective

[1] R is total-connected iff vRw or wRv for all v, w. In (Boutilier 1990) we use *forward-connectedness*, but this stronger condition results in an equivalent logic (Hughes and Cresswell 1984), and the distinction is important later.

[2] Formally, $n(A)$ need not exist, but the technical details which follow will not depend on this. We use this notation only informally, to illustrate the ideas which follow. $\|\alpha\|$ stands for the set of all possible worlds which satisfy α, and again, in our informal discussion, we take this to mean all logically possible worlds rather than those from the set W of some formal structure.

[3] We use "known" here in a much less technical sense than Levesque. More accurately, we could say if $n(A)$ were the only worlds an agent considered possible, then it would know at least $A \wedge C$.

$\overleftarrow{\Box}$. This connective corresponds to Levesque's N, and $\overleftarrow{\Box}\ \alpha$ will hold exactly when α is true at all inaccessible worlds.

Our language **L** will be formed from a denumerable set **P** of propositional variables, together with the connectives $\neg$, $\supset$, $\Box$ and $\overleftarrow{\Box}$. The connectives $\wedge$, $\vee$ and $\equiv$ are defined in terms of these in the usual way.

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

The truth of a formula α at w in M is defined in the usual inductive manner, with the interesting cases being:

$$M \models_w \Box\alpha \text{ iff for each } v \text{ such that } wRv,\ M \models_v \alpha.$$

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

We can 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 some world if α holds at all accessible (inaccessible) worlds; $\overleftrightarrow{\Box}\ \alpha$ ($\overleftrightarrow{\Diamond}\ \alpha$) holds iff α holds at all (some) worlds, whether accessible or inaccessible. Validity is defined in a straightforward manner, a sentence α being *CO-valid* ($\models_{CO}\ \alpha$) just when every CO-model M satisfies α. Finally, we define the connectives:

$$A \Rightarrow B \equiv_{\text{df}} \overleftrightarrow{\Box}\ \neg A \vee\ \overleftrightarrow{\Diamond}\ (A \wedge \Box(A \supset B)).^4$$

$$A > B \equiv_{\text{df}} \overleftrightarrow{\Box}\ (A \supset (\overleftarrow{\Box}\ (A \supset \neg B) \wedge \Diamond(A \wedge B))) \wedge\ \overleftrightarrow{\Diamond}\ A.$$

$A \Rightarrow B$ will be true vacuously if there is no world in the model at which A holds. Otherwise, it is true iff there is some world where A holds, and B holds at all more normal A-worlds The dual of this is $A > B$ which states that at most $A \wedge B$ is known at the most normal A-worlds. This is only true if A holds at some world (condition $\overleftrightarrow{\Diamond}\ A$), otherwise there would exist no such worlds and everything would be trivially satisfied ("known") by this (empty) set. Furthermore, $A > B$ can only hold if, at each A-world, $A \supset \neg B$ is true at every inaccessible world (otherwise this world would be strictly more normal than some $A \wedge B$-world); and if at each A-world, $A \wedge B$ is true at some more normal world (since such worlds are the most normal A-worlds, each A-world should "see" one). It is easy to verify that $A > B$ holds iff all $A \wedge B$-worlds are mutually accessible (equally normal) and no A-world is strictly more normal than these. In other words, by asserting that $\neg B$ holds at each inaccessible A-world, we force all $A \wedge B$-worlds to be accessible to, or at least as normal as, every other A-world.

We call the logic associated with this semantics CT4D-O, or CO for short, the extension of CT4D allowing conditionalized "only knowing". Completeness is proven using a technique of Humberstone (1983).

Definition The conditional logic CO is the smallest $S \subseteq$ **L** such that S contains classical propositional logic and the following axiom schemata, and is closed under the following rules of inference:

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

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

T $\Box A \supset A$

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

4′ $\overleftarrow{\Box}\ A \supset \overleftarrow{\Box}\overleftarrow{\Box}\ A$

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.

Theorem 1 *The system* CO *is characterized by the class of CO-models.*

That the connective $\Rightarrow$, as defined, captures a reasonable notion of normal implication has been discussed in (Boutilier 1990). It should also be clear that the connective $>$ captures the dual notion of (conditional) "knowing at most", and that together, $A \Rightarrow B$ and $A > B$ allow us to extend the conditional to include all irrelevant properties. Consider the following theorem of CO (for propositional A, B, and α):

$$(A \Rightarrow B \wedge A > B) \supset (\overleftrightarrow{\Diamond}\ (A \wedge B \wedge \alpha) \supset (A \wedge \alpha \Rightarrow B)).$$

This theorem states that if both $A \Rightarrow B$ and $A > B$ are true, then we can conclude that all properties α are irrelevant to the truth of the original conditional.

Notice that the extension of $A \Rightarrow B$ is conditional on the possibility of $A \wedge B \wedge \alpha$. If we insist that all logically possible worlds be contained in W, then we can derive $A \wedge \alpha \Rightarrow B$ directly (provided $A \wedge B \wedge \alpha$ is logically consistent). Levesque (1990) enforces a similar constraint. This gives rise to the following extension of CO:

Definition CO* is the smallest extension of CO closed under all rules of CO and containing the following:

[4] This definition of $A \Rightarrow B$ is different from that of (Boutilier 1990). It is more similar in spirit to the connective $\mid\!\sim$ of (Kraus, Lehmann and Magidor 1990), whereby if $A \Rightarrow B$ holds at any world in a model then it holds at all worlds. Previously, $A \Rightarrow B$ could hold "vacuously" if there were no *accessible* worlds at which A is true. While this is in accord with an epistemic reading of the relation R, it does not conform to our normative interpretation. It is entirely unreasonable to expect only more normal worlds to determine which normative statements we take to be true. Worlds which are exceptional should also play a role in such deliberations. One advantage of our approach is that the connective $\overleftarrow{\Box}$ allows us to define the truth conditions of $\Rightarrow$ at individual worlds, whereas the truth conditions of $\mid\!\sim$ can only be defined with respect to entire structures.

LP $\overleftrightarrow{\Diamond}\ \alpha$ for all satisfiable propositional α.[5]

Definition A *CO*-model* is any CO-model $M = \langle W, R, \varphi\rangle$, such that $\{w^* : w \in W\} \supseteq \{f : f \text{ maps } \mathbf{P} \text{ into } \{0,1\}\}$.[6]

Theorem 2 *The system* CO* *is characterized by the class of CO*-models.*

The logic CO* addresses the difficulty of having to conditionalize extensions of normative conditionals on the possibility of the antecedent (since all consistent such antecedents are possible). Hence a theorem of CO* (for satisfiable $A \wedge B \wedge \alpha$) is $(A \Rightarrow B \wedge A > B) \supset (A \wedge \alpha \Rightarrow B)$.

The notion of irrelevance sketched here is rather undermotivated. While it's clear in examples such as the case of green birds that *Green* should be irrelevant to *Fly*, the question remains: what do we mean by *irrelevance*? Space limitations preclude anything resembling a reasonable discussion of this point, but a few words are in order. Gärdenfors (1978) has presented and discussed a number of postulates which should be satisfied by any notion of relevance, motivated by the consideration that p is relevant to r (given evidence e), written $p\mathcal{R}_e r$, iff the conditional probability of r given $p \wedge e$ is different than that of r given e alone (i.e. $P(r|p \wedge e) \neq P(r|e)$). Postulates (R0) to (R4) are presented as reasonable restrictions on the relevance relation.[7] In (Boutilier 1991b) we define a notion of *statistical relevance* (*s-relevance*) which roughly states that $p\mathcal{R}_e r$ if learning the truth (or falsity) of p affects our judgement as to the truth of r. Assuming $e \Rightarrow r$ means we are willing to accept r based on evidence e, p is relevant to r if $p \wedge e \not\Rightarrow r$ or $\neg p \wedge e \not\Rightarrow r$ (similar definitional clauses apply when $e \Rightarrow \neg r$, or $e \not\Rightarrow r$ and $e \not\Rightarrow \neg r$). We show this definition to satisfy the postulates, and that asserting $e > r$ ensures that any sentence contingent on $e \wedge r$ is irrelevant to r in this sense.

We also define a weaker notion of *commonsense relevance* (*c-relevance*), violating the postulate (R2) which asserts that $p\mathcal{R}_e r$ iff $\neg p\mathcal{R}_e r$. Defined in terms of conditional independence, s-relevance must satisfy (R2). If learning p increases the probability of r, then learning $\neg p$ must decrease it. If the magnitudes of the changes are vastly different, this may seem a counterintuitive notion of relevance. For instance, if I'm about to cross a bridge a someone tells me to go ahead because there will be no earthquakes ($\neg Q$) in the next minute, I'm liable to dismiss my informant as a lunatic and discount $\neg Q$ as being irrelevant. However, if I am told there *will* be an earthquake (Q), I will surely consider this information to be relevant. Intuitively, $\neg Q$ is irrelevant because (statistically) it changes the probability of a safe crossing negligibly (assuming the prior probability of Q is very low), while Q changes the probability radically. We capture this potential failure of (R2) by saying that p is c-relevant to r if, e.g., $p \wedge e \not\Rightarrow r$ when $e \Rightarrow r$, and discounting the possible effect of learning $\neg p$. It often seems that p is only regarded as relevant if its *truth* can change the status of r as an accepted belief, not its falsehood.

Other reasons for disassociating conditional independence and irrelevance are mentioned in (Gärdenfors 1978) and we discuss our definition of irrelevance and how inaccessible worlds capture this concept in detail in (Boutilier 1991b).

While the logic CO* seems able to express the concept of irrelevance, it is not clear how a default reasoner should proceed given such a logic and a set of facts KB. A modest proposal is to simply assert $A > B$ for each $A \Rightarrow B$ in KB, so long as the result is consistent. This works on a wide variety of examples; for instance, if $KB = \{Bird \Rightarrow Fly\}$, then asserting $Bird > Fly$, we can derive conditionals such as $Bird \wedge Green \Rightarrow Fly$. If $Penguin \Rightarrow \neg Fly$ and $\Box(Penguin \supset Bird)$ are added to KB, $B > F$ is no longer consistent. However, $B \wedge \neg P \Rightarrow F$ and $B \wedge P \Rightarrow \neg F$ are both derivable and it is consistent to assert $B \wedge P > \neg F$ and $B \wedge \neg P > F$. Adding these to KB, we obtain the following theorems:

1. $(KB \wedge Bird) \Rightarrow Fly$
2. $(KB \wedge Bird \wedge Green) \Rightarrow Fly$
3. $(KB \wedge Penguin) \Rightarrow \neg Fly$
4. $(KB \wedge Penguin \wedge Green) \Rightarrow \neg Fly$

Such an approach, however, has limitations. Consider a KB of two independent conditionals $A \Rightarrow B$ and $C \Rightarrow D$. In this case, it is inconsistent to assert both $A > B$ and $C > D$, and it is not clear what "extendible" conditionals of interest are derivable from such a KB. Thus, the use of the connective $>$ for dealing with irrelevance requires further investigation. Another simple proposal, which adequately handles this KB, can be described as follows: since the material counterparts of these sentences, $A \supset B$ and $C \supset D$, must be normally true (that is $\top \Rightarrow (A \supset B \wedge C \supset D)$), it should be the case that $\top > (A \supset B \wedge C \supset D)$ holds as well. Extending this idea, we will show that the connective $>$ is capable of representing a certain form of default reasoning, namely 1-entailment or rational closure.

3 A Simple Preference Relation

A common approach to default reasoning is to use the notion of *preferred models* (Shoham 1986). Given a set of CO-models, we will suggest the preferred models are those in which possible worlds are as normal as possible. Consider again a KB containing only $Bird \Rightarrow Fly$. A

[5] Alternatively, we could use Levesque's schema: $\overleftarrow{\Box}\ \alpha \supset \neg\Box\alpha$ for all falsifiable α.

[6] 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.

[7] Gärdenfors also presents postulate (R5) — which leads to a triviality result — and two possible replacements, one of which (R7) is deemed acceptable. The notion of relevance defined below is a simple one, but can be extended easily to incorporate (R7) (see (Boutilier 1991b) for details).

model of KB will contain some $Bird \wedge Fly$-world which is more normal than any $Bird \wedge \neg Fly$-world. In general, we want to derive sentences like $Bird \wedge Green \Rightarrow Fly$, but the most normal worlds with green birds need not satisfy Fly. However, assuming that some $Green \wedge Bird \wedge Fly$-world is as normal as the most normal $Bird$-worlds violates no constraints imposed by KB. This assumption forces such a world to be more normal than we originally supposed, so if preferred models force worlds to be as normal as possible, $Bird \wedge Green \Rightarrow Fly$ will be derivable.

We must formulate the conditions under which one model will be more normal than another. Let $M_1 = \langle W, R_1, \varphi \rangle$ and $M_2 = \langle W, R_2, \varphi \rangle$ be CO-models.[8] To ensure that M_1 is at least as "normal" as M_2, each world in W should be as normal in M_1 as in M_2; so we will insist (in general) that each world "see" at least as many worlds in R_1 as in R_2. There are two cases to consider when a world w has fewer accessible worlds in R_1. First, some world v might be less normal in R_1 than in R_2, in which case it is inaccessible to (some) w in R_1 to which it was accessible in R_2. In such a case M_1 should not be preferred to M_2. However, in the second case, w may have become more normal in R_1, in which case it *should* see fewer worlds (since fewer will be more normal than it). In this circumstance, M_1 may well be preferable to M_2, and we relax the restriction that w see as many worlds in R_1. More formally, assume M_1 and M_2 are defined as above.

Definition $w \in W$ is *more normal in* R_1 than in R_2 (written $N(w, R_1, R_2)$) iff there is some $v \in W$ such that vR_1w, wR_1v, and not vR_2w.

Definition M_1 is *as preferable as* M_2 (written $M_1 \leq M_2$) iff for all $w \in W$, $N(w, R_1, R_2)$ is false only if $\{v : wR_2v\} \subseteq \{v : wR_1v\}$. M_1 is *preferred to* M_2 ($M_1 < M_2$) iff $M_1 \leq M_2$ and $M_2 \not\leq M_1$.

Definition Let $T \subseteq \mathbf{L}$ be a set of facts. M is a *minimal model of* T iff $M \models T$ and for all M' such that $M' \models T$, $M' \not\leq M$. α is a *default conclusion* based on T (written $T \models_{\leq} \alpha$) iff $M \models \alpha$ for each minimal model M of T.[9]

We examine the consequences of these definitions in the following section.

4 Equivalence to 1-entailment

Pearl (1990) describes a natural ordering on default rules named the *Z-ordering*, and uses this to define a nonmonotonic entailment relation, 1-entailment. While put forth as an extension of ε-semantics, this entailment relation is essentially based on using preferred models of a sort similar to those described in the previous section. In fact, we will show these notions correspond exactly, and that, while 1-entailment is defined in terms of a particular theory T and orders only models of T, it can be described in terms of our theory-independent preference criterion, whereby all logical interpretations are ordered.

In this section, we will assume a language with a finite set of propositional variables, and CO*-models only will be treated. The default rules r of (Pearl 1990) have the form $\alpha \rightarrow \beta$, where α and β are propositional formulae. These will correspond to our conditional sentences $\alpha \Rightarrow \beta$. 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 $w \models \alpha \supset \beta$. Let T be a finite set of such rules. From (Pearl 1990):

Definition T *tolerates* $\alpha \rightarrow \beta$ iff there is some world which 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.

This notion of toleration can be used to characterize probabilistic ε-consistency (Adams 1975; Pearl 1988) in a manner that also captures the CO-consistency of a set of rules.[10] Furthermore, toleration can be used to define a natural ordering on default rules (or conditionals) by partitioning T as follows (Pearl 1990):

Definition $T_i = \{r : r \text{ is tolerated by } T - T_0 - T_1 - \cdots T_{i-1}\}$, for $i \geq 0$.

Assuming T is ε-consistent, this results in an ordered partition $T = T_0 \cup T_1 \cup \cdots T_n$. Now to each rule $r \in T$ we assign a rank (the *Z-ranking*), $Z(r) = i$ whenever $r \in T_i$. The idea is that lower ranked rules are more general, or have lower priority. Given this ranking, we can rank possible worlds according to the highest ranked rule they falsify:

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

Lower ranked worlds are to be considered more normal. Now, a propositional formula α can be ranked according to the lowest ranked world which satisfies it: $Z(\alpha) = \min\{Z(w) : w \models \alpha\}$. Given that lower ranked worlds are considered more normal, we can say that a normative conditional $\alpha \Rightarrow \beta$ should hold iff the rank of $\alpha \wedge \beta$ is lower than that of $\alpha \wedge \neg\beta$. This leads to the following definition (Pearl 1990):

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

[8] We will only compare models which agree on possible worlds; however, it should be clear that the idea can be extended by taking preference relative to the subset of worlds two models have in common. See (Boutilier 1991c).

[9] Strictly speaking, T should consist only of conditional sentences and α should be conditional as well. See (Boutilier 1991c) for details on how to extend this relation.

[10] T is ε-consistent iff every non-empty subset of T contains some rule tolerated by all others. If T is the corresponding set of conditionals, this condition holds iff T is "non-vacuously" satisfiable; that is, if $T \cup \{\overset{\leftrightarrow}{\Diamond} \alpha : \alpha \Rightarrow \beta \in T\}$ is CO-consistent.

For details regarding the types of conclusions 1-entailment draws, see (Pearl 1990). It should be fairly clear that 1-entailment can be viewed as asserting a preference on models of theory T, namely that worlds should have their lowest possible rank (without violating the rules of T). In other words, worlds should be as normal as possible. Not surprisingly then 1-entailment corresponds to the preferential entailment relation of the preceding section. For a fixed theory T, we define the CO*-model Z_T as:

Definition $Z_T = \langle W, R, \varphi \rangle$ where wRv iff $Z(w^*) \geq Z(v^*)$.

Theorem 3 *$T \models_{\leq} \alpha \Rightarrow \beta$ iff $\alpha \vdash_1 \beta$ with respect to T.*

This means that the minimal Z-ranking of worlds corresponds to a theory-dependent instance of the more general preferential ranking of CO*-models. Furthermore, the explicit nature of this Z-ranking allows us to capture the exact nature of the (unique) preferred model Z_T. In particular, if T is ε-consistent and is partitioned as $T_0, \cdots, T_n$, then Z_T consists of $n+2$ "clusters" of mutually accessible (or equally normal) worlds; cluster 0 consists of all worlds of rank 0, cluster 1 consisting of all worlds of rank 1, and so on, with the most exceptional worlds being those of rank $n+1$.

Since the preferred model of T is unique, we can capture the exact structure of Z_T using sentences in the logic CO containing the connective $>$, since worlds in each cluster can be characterized by the rules they violate (see (Boutilier 1991c) for details).

Let T be a finite set of conditionals, partitioned as $T_0, T_1, \cdots, T_n$.

Definition Let $R^{\wedge}_{-1} \equiv_{\text{df}} \bot$. For $0 \leq i \leq n+1$, let $R^{\wedge}_i \equiv_{\text{df}} \bigwedge\{\alpha \supset \beta : \alpha \Rightarrow \beta \in T - T_0 - \cdots T_{i-1}\}$. We assume $\bigwedge \emptyset \equiv_{\text{df}} \top$ (hence $R^{\wedge}_{n+1} \equiv \top$).

Definition For theory T as above, the *closure* of T is $Cl(T) = T \cup \{\neg R^{\wedge}_i > R^{\wedge}_{i+1} : -1 \leq i \leq n\}$.

Theorem 4 *$Cl(T) \models_{CO*} \alpha \Rightarrow \beta$ iff $T \models_{\leq} \alpha \Rightarrow \beta$.*

Corollary 1 *$Cl(T) \models_{CO*} \alpha \Rightarrow \beta$ iff $\alpha \vdash_1 \beta$ with respect to T.*

Just as the (second-order) circumscriptive axiom applied to a theory T closes that theory to correspond to (predicate) minimal models, so too does this closure correspond to our notion of minimality. Theorem 4 shows that $Cl(T)$ can be regarded as an axiomatization of the notion of preference described in the previous section, and of the implicit preference ordering determined by System-Z. Hence, the types of conclusions sanctioned by 1-entailment (see (Pearl 1990)) are also determined by this form of closure. This implies, given the results of (Goldszmidt and Pearl 1990), that $Cl(T)$ determines the same consequence relation as that of rational closure (Lehmann 1989).

5 Concluding Remarks

We have presented a modal logic CO in which truth at inaccessible worlds is expressible. In this logic we can define not only the normative conditional $\Rightarrow$ of conditional "knowing at least", as in (Boutilier 1990), but also the dual connective $>$ of conditional "knowing at most". This provides us with a conditional version of Levesque's (1990) "only knowing". We discussed briefly the relationship of conditional only knowing to the problem of irrelevance in default reasoning, and have shown how a simple preference relation (corresponding to 1-entailment and rational closure) which deals with irrelevance can be axiomatized within this logic.

Much work remains to be done on the application of conditional only knowing to problems in default reasoning. Levesque's characterization is semantically very clear and elegant, but has the drawback of relying on an autoepistemic interpretation of default rules (see e.g. (Reiter 1987) for problems with this interpretation). Ultimately, we would like to push the "closure" of our default theories into the logic via some connective analogous to Levesque's O operator, thereby consolidating the semantic clarity of only knowing with the compelling conditional interpretation of default rules. We expect the expressive power gained by the use of inaccessible worlds makes this goal achievable.

In this connection, we have begun exploring the use of the logic CO to capture a number of other types of reasoning. In (Boutilier 1991a) we provide a logical calculus for belief revision within CO, and show how revision, subjunctive reasoning and default reasoning (including such varied approaches as autoepistemic logic, ε-semantics and normative conditionals) can be unified using a framework which exploits the power of inaccessible worlds.

Other avenues to pursue include the weakening of these logics, along the same lines suggested in (Boutilier 1990), providing other versions of "only knowing" (e.g. based on S4). A further task is to investigate how the logic CO can be used to capture other default reasoning systems, such as the maximum entropy formalism of (Goldszmidt, Morris and Pearl 1990), which makes finer-grained distinctions on the ordering of possible worlds. We suggest that most approaches to default reasoning which can be viewed as restricting the degree of abnormality of worlds may be axiomatized using some logic of inaccessible worlds, or conditional only knowing.

Acknowledgements

Thanks to Ray Reiter, Hector Levesque and especially Moisés Goldszmidt for helpful discussion and criticism. Financial support of NSERC and the University of Toronto is gratefully acknowledged.

References

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

Boutilier, C. 1990. Conditional logics of normality as modal systems. In *Proc. of AAAI*, pages 594–599, Boston.

Boutilier, C. 1991a. Belief revision as a modally defined conditional. Technical report, University of Toronto. forthcoming.

Boutilier, C. 1991b. *Conditional Logics for Default Reasoning and Belief Revision*. PhD thesis, University of Toronto. Forthcoming.

Boutilier, C. 1991c. Preliminary report on inaccessible worlds and irrelevance. Technical Report KRR-TR-91-1, University of Toronto.

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

Gärdenfors, P. 1978. On the logic of relevance. *Synthese*, 37(3):351–367.

Goldszmidt, M., Morris, P., and Pearl, J. 1990. A maximum entropy approach to nonmonotonic reasoning. In *Proc. of AAAI*, pages 646–652, Boston.

Goldszmidt, M. and Pearl, J. 1990. On the relation between rational closure and system Z. In *Nonmon. Reasoning Workshop*, pages 130–140, South Lake Tahoe.

Hughes, G. E. and Cresswell, M. J. 1984. *A Companion to Modal Logic*. Methuen, London.

Humberstone, I. L. 1983. Inaccessible worlds. *Notre Dame Journal of Formal Logic*, 24(3):346–352.

Kraus, S., Lehmann, D., and Magidor, M. 1990. Nonmonotonic reasoning, preferential models and cumulative logics. *Artificial Intelligence*, 44:167–207.

Lehmann, D. 1989. What does a conditional knowledge base entail? In *Proc. of KR'89*, pages 212–222, Toronto.

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

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. 1989. Probabilistic semantics for nonmonotonic reasoning: A survey. In *Proc. of KR'89*, pages 505–516, Toronto.

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

Reiter, R. 1987. Nonmonotonic reasoning. *Annual Reviews of Computer Science*, 2:147–186.

Shoham, Y. 1986. Reasoning about change: Time and causation from the standpoint of artificial intelligence. Technical Report YALEU/CSD/RR#507, Yale University, New Haven.