[15] V. Lifschitz, Between circumscription and autoepistemic logic, in: *Proceedings of the First International Conference on Knowledge Representation and Reasoning*, Toronto, Ontario (1989).

[16] V. Lifschitz, On open defaults, in: *Proceedings of the Symposium on Computational Logic*, Brussels (1990).

[17] W. Marek and V. S. Subrahmanian, The relationship between stable, supported, default and auto-epsitemic semantics for general logic programs, Technical Report 128-88, Department of Computer Science, University of Kentucky (1988).

[18] J. McCarthy, Circumscription — a form of nonmonotonic reasoning, *Artificial Intelligence* **13** (1-2) (1980).

[19] J. McCarthy, Applications of circumscription to formalizing commonsense knowledge, *Artificial Intelligence* **28** (1986).

[20] R. C. Moore, Reasoning about knowledge and action, PhD thesis, Massachusetts Institute of Technology, Cambridge, MA (1980).

[21] R. C. Moore, Semantical considerations on nonmonotonic logic, *Artificial Intelligence* **25** (1) (1985).

[22] D. Poole, Fixed predicates in default reasoning, in: *Proceedings of the International Joint Conference on Artificial Intelligence*, Milan, Italy (1987) 905-908.

[23] P. K. Rathmann and M. Winslett, Circumscribing equality, in: *Proceedings of the International Joint Conference on Artificial Intelligence*, Detroit, MI (1989) 468-473.

[24] R. Reiter, Equality and domain closure in first-order data bases, *Journal of the ACM* **27** (1980).

[25] R. Reiter, A logic for default reasoning, *Artificial Intelligence* **13** (1-2) (1980).

[26] J. S. Schlipf, How uncomputable is general circumscription, in: *Proceedings of the Symposium on Logic in Computer Science*, Cambridge, MA (1986) 92-95.

[27] R. C. Stalnaker, A note on nonmonotonic modal logic, Department of Philosophy, Cornell University, (1980).

Fundamenta Informaticae XV (1991) 301-323
IOS Press

CONDITIONALS AND ARTIFICIAL INTELLIGENCE

John F. HORTY
Institute for Advanced Computer Studies, University of Maryland, College Park, MD 20742, USA

Richmond H. THOMASON
Intelligent Systems Program, University of Pittsburgh, Pittsburgh, PA 15260, USA

1 Introduction

The problem of conditionals, like many of the larger and more difficult problems in logic and AI, is hard to explain to outsiders. People reason effortlessly with 'if', and it is surprising that there should be a mystery about something that seems so straightforward in practice. Puzzles about what 'if' means arise only in a theoretical context, in which an account is sought of valid reasoning with 'if' that deals with the phenomena, and that meets the theoretical standards that apply in logic.

These standards are only as recent as mathematical logic—so the logical problem of conditionals appeared only after it was discovered that the theory of 'if' that was developed in logical accounts of mathematical reasoning did not apply well to other domains. According to the so-called theory of the "material conditional" that emerged from work in Nineteenth Century logic, and that is found in Frege's work and *Principia Mathematica*, a conditional statement *If A then B* is true if and only if the antecedent *A* is false or the conclusion *B* is true. But it is a matter of common sense that some conditionals with false antecedents are true while others are false, and there is an abundance of simple counterexamples to the material condition for the truth of sentences involving 'if'. Consider, for example, the following true-false examination.

> John's salary this year is \$21,000. Last year he received a 5% raise. Are the following statements true or false?
>
> 1. If he had received a 10% raise last year, his salary would be \$22,000.
> 2. If he had received a 10% raise last year, his salary would be \$21,000.

The first statement is clearly true, the second false.

The most radical reactions to this discrepancy between the theory and common sense have proposed solutions that would affect the truth conditions for conditionals in mathematics, as well as those of contingent conditionals. This line of development tends to stress the idea of a connection between antecedent and consequent. The most important work in this area belongs to the tradition of "entailment", or "relevance logic," set out in [2].

Another tradition, inspired to a large extent by considerations in the philosophy of science, and from a desire to extend the scope of symbolic logic from mathematical domains to the empirical sciences, has concentrated on the problem of contingently

false antecedents, or "contrary to fact conditionals." The logical accounts that derive from this tradition, which share the idea that the antecedent of a conditional induces some sort of a small change in a state, are in general conservative in their treatment of mathematical conditionals; the best known theories are equivalent to the material conditional in cases where the antecedent is necessarily true or necessarily false.

Though these two traditions emerged from similar origins, by now they have diverged into quite separate areas of research. In this paper, we will concentrate on the second approach, relating it to issues in AI.

We believe that an examination of these issues in the context of the larger problems of conditional reasoning will be timely and useful for both philosophical logic and AI. Formalists in AI have discovered the relevance of the earlier theories of the conditional to nonmonotonic reasoning and have rediscovered or imported many of the earlier results. Placing conditional logics in the larger setting of nonmonotonic reasoning is a major logical advance; we will argue that this new setting provides a promising approach to fundamental problems that the earlier approaches were not able to solve. On the other hand, we believe that a fresh look at the traditional problem of counterfactuals will be helpful in the present setting. With the proliferation of formal work in this area, it is easy to forget the larger problems with which the formalisms are meant to deal.

2 Attempts at an analysis

Unlike the relevance logics, which were stimulated in part by applications of logic within mathematics, the study of counterfactual conditionals grew out of concerns in the philosophy of science. One of the most important of these concerns, stressed especially by Carnap in [9], was the treatment of *dispositionals*—predicates like 'soluble' or 'flammable', that describe how things should be expected to behave under certain conditions. It is natural to expect the word 'if' to figure in the definitions of these predicates. A soluble substance will dissolve *if* it is immersed in water; a flammable object will burn *if* it is heated sufficiently. But definitions of this kind yield incorrect results if they are based on the material theory of the conditional; for example, this theory must classify any object that is never heated sufficiently as flammable. Because of this problem, Carnap abandoned the idea of explicitly defining these dispositional predicates, and instead suggested in [9] that they should be introduced into a scientific language by postulates that amount to only partial definitions.

Another approach, of course, is to continue with the project of definition, but to try to overcome the problems due to the material conditional by developing an account of counterfactual implication. The earliest attempts at developing such a notion—exemplified by the work of Chisholm [11], for example—focused on ways of defining the counterfactual conditional using the resources of ordinary classical logic. But this line of research was undermined with the publication of Goodman [19], which, without offering anything as strong as a proof, set out a compelling case for thinking that the task of defining counterfactual implication within ordinary logic would be unexpectedly difficult, and perhaps impossible. It is worth considering this argument here, because it raises issues that are still, as we will see, alive today in recent attempts to deploy theories of conditionals within AI.

Most theorists interested in analyzing counterfactual conditionals within ordinary logic proceed by attempting to define a certain kind of connection between the antecedent and the consequent of a true conditional. Such a conditional is supposed to be true if its antecedent, together with some set of laws and some appropriate subset of the true background facts, logically entails its conclusion. A standard example is the counterfactual

If this match were struck, it would light,

said of a match that is not struck, and so does not light. In this case, the relevant background facts might include the fact that the match is dry, and that oxygen is present; the physical laws might include the statement that any dry match that is struck when oxygen is present will light. The conditional is then supposed to be true because, taken together with these background facts and laws, its antecedent does yield its consequent as a matter or ordinary logic.

Goodman associates two problems with this kind of analysis.[1] The first—which we will not consider in any detail here—is that the analysis seems to require some criterion for distinguishing genuine causal laws from mere accidental generalizations. Statements of both kinds are truths of the form $\forall t x_1 \ldots x_n (A \supset B)$, where t ranges over times. But though the former do support counterfactuals, the latter do not. Even if it is a true generalization, for example, that all the coins in my pocket this week are dimes, we would not want an account of counterfactual reasoning to support the conclusion that some penny currently in my hand would be a dime if I were to place it in my pocket. Because of their different role in supporting counterfactuals, an analysis of this kind must make the distinction between accidental and lawlike generalizations. But it isn't easy to see how to do this, without using counterfactuals; Goodman considers a number of more or less principled ways of formulating the distinction, but is able to show that they all are problematic. Though Goodman regards this as the harder of his two problems, it is perhaps not so disturbing in the context of AI, where domain dependent solutions can be accepted; we might attempt to axiomatize the causal knowledge pertaining to a domain, and define the laws as the logical consequences of these axioms. As long as we can combine axiomatizations of different domains, we can hope to make incremental progress on the formalization of causal knowledge.

But the second of Goodman's two problems is closely related to issues that are current and problematic in AI; this is the problem of identifying the relevant background facts for the evaluation of a counterfactual—the particular set of truths that, taken together with the laws and the antecedent of a true counterfactual, are supposed to imply its consequent.

Evidently, the set of background facts relevant for evaluating a counterfactual conditional cannot include the entire set of truths. Because the antecedent of such a conditional is false this idea would yield an inconsistency; any other statement could then be derived from this set, and so any counterfactual at all would have to be judged true. Nor can we remove all the truths B that are inconsistent with the antecedent A, because the resulting set may still be inconsistent. (For instance, A could be $\neg(B_1 \wedge B_2)$,

[1] In what follows, we are trying to capture what we think is most important about Goodman's argument in terms that will make the difficulties clear to readers who have some familiarity with more recent ideas from conditional logic, and from related work in theory revision, such as [14]. We have changed the emphasis and order of exposition, and have introduced some anachronisms into the argument. Readers who want an accurate historical picture of Goodman's work should read [19] and [20].

where B_1 and B_2 are truths.) We are left with the idea that the relevant set of background facts for evaluating a counterfactual should be some subset of the facts that are, in some sense, compatible with its antecedent. But to specify the right subset, we must characterize the sort of compatibility that is wanted here, and explain how the subset is chosen.

Mere logical compatibility is clearly too weak, since the negation of any nontrivial conditional's consequent will be logically compatible with its antecedent. Goodman suggests that compatibility with physical laws is what is wanted. If this is is correct, compatibility could be explicated as consistency relative to a theory consisting of classical logic and an axiomatization of the underlying causal knowledge, and so would depend on a solution to the first problem.

But what subset of compatible truths is needed? The empty set is evidently much too small. A large number of totally irrelevant truths are automatically preserved by contrary to fact conditionals. If my shirt is red, the conditional

If this match were struck, my shirt would (still) be red

is true (though its truth is so trivial that it is hard to think of any natural reason to assert it). And to enable the consequents of intuitively true conditionals to be logically deduced from the set of background facts together with the laws, we will need to preserve the truth of the facts that need to be used in the deduction; in the case of the match, for instance, we will need to have the fact that the match is dry in our set of relevant background facts.

There is a general way to ensure that a large set of truths will be preserved: require that the set of background facts for evaluating a conditional with antecedent A be a *maximal* subset of the set of truths that are logically compatible with Γ, where Γ is the appropriate set of causal laws. A familiar logical construction guarantees the existence of such a maximal subset. But uniqueness is not guaranteed, and in general there may turn out to be many such maximal subsets. Goodman does not in fact explore this idea, which is a later theme developed in work such as [14]. The variations that he does consider do not address the problem of maximality adequately, and are unsatisfactory in various ways.

However, Goodman goes on to note a problem that is much larger than these matters of detail, and that would also apply to proposals involving maximality. As it turns out, there are far too many maximal sets that are compatible with the antecedent and with physical laws, and it is hard to see how to eliminate these alternatives without appealing to the notions we are trying to explicate.

The easiest way to see this to return to the match example. We assume a situation in which oxygen is present and the match is dry, but in which it is not struck and so does not light. The initial set of truths can thus be represented as the collection of statements:

$$O,\ \neg S,\ D,\ \neg L.$$

And we assume as our sole physical law that the match will light if it is dry and struck in the presence of oxygen:

$$(D \wedge O \wedge S) \supset L.$$

Now suppose we want to know what would happen if the match were struck; what background set of facts should we consider? One maximal subset of the initial facts that is consistent both with the physical laws and our conditional hypothesis is $\{O, D\}$. Together with these laws and our hypothesis, this background set does logically imply L, supporting the intuitively correct conclusion that the match would light if it were struck. However, as Goodman points out, the simple compatibility criterion that we have been working with also allows us to consider pathological sets of background facts, such as $\{D, \neg L\}$ and $\{O, \neg L\}$. The former supports the conclusion that, if the match were struck, there would no longer be oxygen present. The latter supports the conclusion that the match would no longer be dry.[2]

Evidently, we need some way of distinguishing the preferred background set $\{O, D\}$ from the others, but the present criterion does not do that: each of these candidates is a maximal subset of the original truths compatible with the physical laws and our antecedent hypothesis. At this point Goodman gives up, and describes this problem by saying that, although the pathological candidates satisfy the conditions of compatibility set out so far, they nevertheless represent background conditions that would not hold *if* the antecedent of our conditional were true, would not hold *if* the match were struck.

This formulation does make it clear what is needed; we can say that a sentence B is supposed to be *cotenable* with an antecedent A if and only if it is not the case that B would be false if A were true. Goodman's ultimate position, then, is that the relevant set of background facts for evaluating a conditional should be some maximal subset of the truths cotenable with the antecedent of that conditional. This may be correct, but of course, as Goodman points out, to rely so heavily on the notion of cotenability, which itself involves counterfactuals, is to abandon entirely the prospect of providing an analysis of the counterfactual conditional within ordinary logic.

Goodman's work is important for contemporary AI because it constrains the enterprise of representing common sense knowledge, and indicates some very general problems that are, we think, sure to apply to any attempt to account for our ability to know an enormous number of conditionals and to reason with them. The problem is central because many fundamental common sense notions, like dispositionals, are definable in terms of conditionals, and because conditionals figure pervasively in reasoning processes such as diagnosis and abduction. For this reason, it is worthwhile to extract some general conclusions from Goodman's work on counterfactuals.

(1) It is knowledge of relevance that enables us to recognize what contingent facts will remain unchanged when a counterfactual hypotheses is made. If a match is dry, it will remain dry if I strike it. If I am in the living room and my hat is in the living room on a peg, then my hat would still be in the living room if I went from the living room into the next room to answer the telephone. But if my hat is on my head, it would not remain in the living room.

Goodman provided a preliminary exploration of the the problem of relevance, but no positive solution. The examples we have used to introduce Goodman's problem of relevance, however, should make it clear that we are dealing with a counterfactual version of the frame problem; how do we specify what factual knowledge is unchanged by a contrary to fact condition? Research in AI on the frame problem should help to

[2]Readers familiar with recent work in nonmonotonic reasoning will note a resemblance between this problem and the problem of "unwanted extensions" which has arisen in various forms in the development of nonmonotonic logic. The resemblance is not accidental.

convince us that Goodman's problem is genuine, timely, and difficult. The counterfactual frame problem is more general than the frame problem in AI, because it requires frame axioms for conditions (where any formula could express a condition), rather than just for actions. However, because recent work in AI has produced new theories that illuminate the frame problem, and the problem is regarded as solvable, at least for special domains, there is still some reason to hope that new methods not known to the earlier conditional theorists may provide insights into Goodman's counterfactual frame problem.

(2) Natural laws should remain invariant under reasonable conditional hypotheses, and so could be used to infer conclusions from the antecedent and from the factual information that is left unchanged by the condition. In retrospect, it would be more accurate to talk about common sense knowledge than natural laws; much of the general, "lawlike" information that is used in conditional reasoning is not at all scientific in character, belonging rather to areas like common sense physics and folk psychology.[3] In any case, the problem of representing general causal knowledge of the sort that is used in conditionals is also genuine, and also is closely related to timely themes in AI.

3 Analysis versus logic

Logical and philosophical work inspired by the problem of contrary to fact conditionals separates into two quite distinct phases. As we have seen, the earlier research, exemplified by the work of philosophers like Carnap, Chisholm, and Goodman, did not seek important modifications of the logical theory; it was generally assumed that what was wanted was an *analysis* of counterfactuals, and that the problems affecting the material conditional could be solved by adding some condition requiring a lawlike connection between the antecedent and the consequent. Since the problem of characterizing the nature of the required connection proved to be difficult, the best of this early work, such as Goodman's, consists primarily of negative results.

The later research—beginning with Stalnaker's theory, presented in [30] and [33]—is motivated by goals that are quite different. Here the idea is to give an account of general inferential characteristics of conditionals, using a model theoretic approach (possible worlds semantics) that treats the antecedent and consequent of a conditional as expressing a set of possible states.[4] David Lewis's independent work in [26] is inspired by much the same motives, though as we will see, the logical goals are less ambitious.

The difference between an analysis and the theories that Stalnaker, Lewis, and the other possible-worlds semanticists have presented is worth noting. The idea of an *analysis* was based on Bertrand Russell's account of definite descriptions. Russell provided a theory of the definite article that was not only logically precise, but that was very conservative in the resources it used (employing only truth functional connectives, the existential quantifier, and identity). Indeed, this conservatism was an intentional part

[3] After all, we are not likely to find definitions of terms like 'match' and 'dry' in any physics or chemistry book. And it is hard to see how the fact that dry matches light if they are struck when oxygen is present could be deduced from any list of physical laws. This is one of the reasons why common sense physics has emerged as a separate area of inquiry from physical science.

[4] In this paper, we will use 'world' and the less philosophical term 'state' more or less interchangeably. It is important to realize that in AI applications, worlds may be relatively small-scale affairs, involving fairly small sets of independent variables. We use the less portentous term 'state' as a reminder that we have in mind small-scale applications of this sort.

of Russell's theory, and an important factor in its influence on philosophers. This theory seemed to explain away a potential source of philosophical confusion (involving nonexistent beings) by eliminating problematic constructions in terms of others that were relatively well understood, and that demonstrably made no appeals to nonexistent beings. Philosophers sought to deliver such an analysis for 'if'.

In developing or evaluating any theory of 'if', it is important to be aware of the differences between "subjunctive" and "indicative" conditionals. There is a famous example pair, due to Ernest Adams, that illustrates the difference.

(1) If Oswald hadn't shot Kennedy, then someone else would have.

(2) If Oswald didn't shoot Kennedy, then someone else did.

The former conditionals are generally used to express causal claims; the latter usually express conditional beliefs. Thus, we would expect someone who asserted **(1)** to have a theory of the assassination; on asking "Why?", we would expect to hear a story about how Kennedy's Presidency had made an assassination certain. Someone who asserts **(2)**, however, needs no theory, and is relying only on the fact that Kennedy was shot. On the other hand, the person who asserts **(2)** is committed to changing his beliefs; **(2)** expresses a willingness to conclude that someone else shot Kennedy on learning that Oswald didn't. To see that the subjunctive conditional doesn't carry such a commitment, it is best to change the example slightly.

(3) If Oswald hadn't shot Kennedy, then Kennedy would be alive today.

(4) If Oswald didn't shoot Kennedy, then Kennedy is alive today.

We could not hope to convince someone who asserts **(3)** that Kennedy is alive today by arguing that Oswald didn't shoot him. What makes **(4)** peculiar is that, unlike **(3)**, it does seem to express a willingness to conclude that Kennedy is alive.

Philosophers like Carnap and Goodman, who were interested in scientific reasoning, were mainly concerned with subjunctive conditionals. Since they tended to assume that the material theory was more or less correct but that the notion of a connection between the antecedent and consequent of a conditional figures in scientific and common sense subjunctives, they did not need to be much concerned about the generality of their analyses.

Stalnaker and David Lewis, however, thought of themselves as providing a better account of the inferential properties of 'if' by changing the underlying logic of the conditional. Here, the theory's scope becomes an issue. Lewis assumed, in effect, that 'if' is ambiguous between the subjunctive and indicative uses, and claimed to be giving a logical theory only of the former. Stalnaker, on the other hand, presented his theory as a logic underlying both the indicative and the subjunctive uses of the conditional, and has tried to account for the differences between the two constructions in terms of pragmatic effects having to do with presuppositions. For more references on this issue, see [22], especially Harper's introduction and the papers in Part 5.

Issues about the scope of the logic of conditionals can be important in determining which applications in AI are appropriate. Indicative conditionals figure mainly in epistemic operations such as truth maintenance, while subjunctives figure in reasoning tasks such as diagnosis. Thus, anyone who applies the logic of conditionals to the former

reasoning tasks is relying implicitly on arguments like Stalnaker's for the generality of the logical theory.

4 Semantic theories of conditionals

By now, a great deal is known about the semantics of conditionals and its axiomatization. This is a brief introduction to some of the main alternatives.

The following classification provides a useful guide to theories of counterfactuals.[5]

	Material	Strict
Absolute	Russell	C.I. Lewis
Relative	Stalnaker	D. Lewis

An absolute theory of the conditional assumes a fixed semantic relation between the antecedent and consequent, which is unaffected by the antecedent; a relative theory allows the "connection" between antecedent and consequent of a conditional to depend systematically on the antecedent. A material theory takes conditionals to be statements of contingency or matter of fact; a strict theory takes them to be statements of necessity. We will refer to this classification in comparing various theories of the conditional and explaining their motivation.

The family of conditional theories deriving from Stalnaker's and Lewis's work are all generalizations of the theories of modal logics that were first proved complete by Saul Kripke. In Kripke's semantics for modal logics, formulas are represented as sets of states, or "possible worlds," and necessity is construed as an operator *Nec* that takes a set of states into a set of states. In the simplest case, that of the modal logic **S5**, $Nec(P) = W$ if $P \neq W$, and otherwise $Nec(P) = \emptyset$, where W is the set of all states.

Where necessity is represented as a one-place operator, the conditional must correspond to a two-place operator. The intuition is that *If A then B* says that B is somehow necessary relative to the antecedent A. Thus, $Nec^*(P)(Q)$ says that Q is necessary relative to some restricted set of states depending on P. This means that we can think of Nec^* as a function taking sets of states into an operator of the same type as *Nec*. To specify a model theory for the conditional, along with a notion of validity, we need to define such an operator. The absolute operators are the simplest cases: the material absolute operator $Nec^*_{\supset}$ and the strict absolute operator $Nec^*_{\prec}$ (for the modal system **S5**) are defined as follows.

$$Nec^*_{\supset}(P)(Q) = \{w : w \notin P \text{ or } w \in Q\}$$
$$Nec^*_{\prec}(P)(Q) = W \text{ if } P \subseteq Q, \text{ otherwise } Nec^*_{\prec}(P)(Q) = \emptyset.$$

What makes these operators absolute is the fact that they can both be defined in terms of a one-place modal operator of some sort and boolean operations; they both have the form $O(-P \cup Q)$, where O is a necessity operator. For the material case, $O(P) = P$; in the strict case, O is **S5** necessity.

C.I. Lewis proposed the strict absolute operator as a solution to his "paradoxes" of the material absolute operator of Frege and Russell; and this proposal does invalidate some unwanted formulas that are valid on the absolute material theory, such as

[5]Taken from [21]. The idea of the classification is Anil Gupta's; it elaborates an earlier classification of David Lewis's.

$$\neg A \supset (A > B).$$

However, absolute theories also validate other patterns that are intuitively invalid in conditional reasoning; the most noteworthy and characteristic of these is *antecedent monotonicity*. The absolute strict conditional validates the inference from **(5)** to **(6)**—an inference that the variable conditional logics are designed to avoid.

(5) If I remove a block from this pile, there will be two blocks remaining.

(6) If I remove a block from this pile and someone adds two blocks to it, there will be two blocks remaining.

The variable theories all make use of a set $s(P, w)$ of states depending on P and w; usually, we are supposed to think of these as the states somehow "closest" to w in which P is true. A variable operator $Nec^*_{>}$ can then be defined in terms of $s(P, w)$ as follows:

$$Nec^*_{>}(P)(Q) = \{w : s(P, w) \subseteq Q\}.$$

Though the propositional operator approach is best for putting the interpretation of conditional logics in perspective, the idea can also be presented as a satisfaction condition for conditional statements $A > B$. Where W is a set, let $\mathcal{P}(W)$ be its power set. A *model structure* for conditional logic is a pair (W, s), where W is a nonempty set of states and s is a function from $\mathcal{P}(W) \times W$ to $\mathcal{P}(W)$. The satisfaction condition for the conditional is then as follows.

(If-Sat) $\quad [\![A > B]\!]_w = \text{T}$ iff $s([\![A]\!], w) \subseteq [\![B]\!]$

According to Condition **(If-Sat)**, a consequent is true in a state w, relative to an antecedent proposition P, in case it is true in all states belonging to $s(P, w)$. This formulation makes it clear how the conditional is a sort of relative necessity.

This theory, with no restrictions whatever placed on the function s, corresponds to a very general, basic conditional logic; specialized conditional logics can then be obtained by adding constraints on the function s. This basic conditional logic, referred to here as **Basic**, can be axiomatized by supplementing ordinary truth functional logic with the following two rules.

(R1) From $A \equiv B$ to infer $(A > C) \equiv (B > C)$.

(R2) From $(C_1 \wedge \ldots \wedge C_n) \supset B$ to infer $((A > C_1) \wedge \ldots \wedge (A > C_n)) \supset (A > B)$.

Intuitions about reasoning with 'if' may differ, but (as in the case of modal logic), the theoretical situation is simplified by a more or less modular relation between axioms and plausible conditions on models, so that a variety of proposals can be axiomatized and proved complete.[6]

Consider, for example,

[6]We won't provide references to proofs of all of the completeness results that are mentioned below. All the ones that are not cited can be proved complete by methods like the ones used in the proofs in [10], [33], [26], [8], and [34]. The last two of these references contain completeness proofs for a variety of conditional logics.

(Id) $A > A$.

This formula is certainly valid intuitively for conditionals; but it does not hold in the basic system.[7] To validate $A > A$, we require $s(P, w)$ to select a set of states in which P holds:

(Supp) $s(P, w) \subseteq P$.

In fact, we can prove completeness, as well as soundness: the system **Basic + Id** is complete with respect to models satisfying **Supp**.

To have even a minimal logic of the conditional, we need to add *modus ponens:* a genuine conditional is false if its antecedent is true and its consequent false.[8] The semantic condition on s corresponding to this inference is a generalization of reflexivity, saying that a state is related to itself, relative to an antecedent, whenever it can be—i.e., whenever the antecedent is true in the state.

(Relative Reflexivity) If $w \in P$ then $w \in s(P, w)$.

Conditions **(Supp)** and **(Relative Reflexivity)** provide a kind of minimal model theoretic semantics for the conditional, which is axiomatized by the system **Basic + Id + MP**.

(MP) $(A > B) > (A \supset B)$

A good deal of the logical work on conditionals consists of extensions of completeness results like the one for the basic system: finding plausible axioms, working out corresponding constraints on models, and proving completeness.

Since many of these further axioms are motivated by the idea that the states in $s(P, w)$ are "close to" w, this logical work has some relevance to attempts in AI to solve the "counterfactual frame problem." Any solution to this problem will, of course, include much domain information; but domain independent, logical constraints on closeness could help in formalizing this information.

The simplest conditions on s motivated by closeness require the actual state to be maximally close to itself, and the "inconsistent state" to be maximally far. More precisely:

(Impossible is Furthest) If $s(P, w) = \emptyset$ then $P = \emptyset$;
(Actual is Close) If $w \in P$ then $w \in s(P, w)$.

The following axioms capture these constraints.

[7] To invalidate $p > p$, choose a model with just two states w_1 and w_2, let p be true in w_1 but not in w_2, and let $s(\{w_1\}, w_1) = \{w_2\}$. Then $p > p$ is false in w_1.

[8] The matter is complicated, however, by the fact that *modus ponens* does not hold of many *conditional modalities* such as conditional obligation. Since the logics of these modalities resemble the logics that have been proposed for 'if', they are often called 'conditional logics'. The situation here resembles the logics for knowledge, which validate $\Box A > B$, and the similar logics for belief, which do not validate this formula.

(MOD) $(A > \neg A) > (B > \neg A)$
(TA) $(A \wedge B) > \neg(A > \neg B)$

Thus, for instance, constraints **Supp**, **Relative Reflexivity**, **Impossible is Furthest**, and **Actual is Close** provide a kind of minimal logic for the conditional, which is axiomatized by the system **Basic + Id + MP + MOD + TA**.

Actual is Close only requires that no world be closer to the actual world than it is to itself; it is natural to strengthen **(Actual is Close)** by requiring that no world can be *as close* to the actual world as it is to itself.

(Actual is Closest) If $w \in P$ then $s(P, w) = \{w\}$.

The axiom corresponding to **Actual is Closest** is

(CS) $(A \wedge B) > (A > B)$.

In the presence of condition **Supp**, **Actual is Closest** is equivalent to a stronger and better known constraint called *Centering.*

(Centering) $w \in P$ if and only if $s(P, w) = \{w\}$.

To say that the actual world is maximally close and that absurdity is maximally far may be a beginning in explicating intuitions about closeness, but does not go very far. The account that is recommended in [26] as the official logic of counterfactual 'if' is a strict relative theory that is stronger in a number of ways than this minimal theory. Lewis imagines a nested system of sets of states (or "spheres") surrounding the actual world. These spheres provide a measure of closeness among states: w_1 is closer than w_2 to w if w_1 belongs to some sphere to which w_2 doesn't belong.

(Spheres) Add to the model structure $\langle W, s \rangle$ a function taking $w \in W$ into a family $\mathcal{S}_w$ of sets of members of W such that (i) for all $P, Q \in \mathcal{S}_w$, either $P \subseteq Q$ or $Q \subseteq P$, (ii) $\mathcal{S}_w$ is closed under unions, (iii) $\mathcal{S}_w$ is closed under nonempty intersections, (iv) $\{w\} \in \mathcal{S}_w$.
Let

$$s(P, w) = \bigcap \{P \cap Q : Q \in \mathcal{S} \text{ and } P \cap Q \neq \emptyset\}.$$

Unlike Lewis, impose the *limit assumption* on models: for all A,

$$\bigcap \{[\![A]\!] \cap Q : Q \in \mathcal{S} \text{ and } [\![A]\!] \cap Q \neq \emptyset\} \neq \emptyset.$$

Lewis recommends a semantic package consisting of the constraints **Supp**, **Spheres**, **Impossible is Furthest**, and **Centering**. This is axiomatized by a system **VC**, one of the best known systems of conditional logic. A number of versions of **VC** have been presented; one of these (from [27]) consists of **Id + MP + MOD + CS**, together with the following two additional axioms.

(CSO) $((A > B) \wedge (B > A)) \supset ((A > C) \equiv (B > C))$
(CV) $((A > B) \wedge \neg(A > \neg C)) \supset ((A \wedge C) > B)$

Lewis's sphere condition is one of many possible constraints on the selection function that relate it to some sort of preference among states. Recent research in "preference semantics" has disclosed a number of other preference conditions, and investigated their relation to axiomatic theories.[9].

A particularly natural condition of this sort is related to uses of *model preference semantics* in nonmonotonic logic.[10] This condition requires $s(P, w)$ to choose the set of states in P that are minimal with respect to a well-founded partial order $\leq_w$. (A partial order having w as its least element is well-founded if it allows no infinite "descending chains" of the form $\ldots w_3 \leq_w w_2 \leq_w w_1$.)

(Partial Order) There is a well-founded partial order $\leq_w$ with least element w such that for all P, $s(P, w) = \{w : w \text{ is } \leq_w\text{-minimal in } P \text{ with respect to } \leq_w\}$.

The reader can check that this condition validates **Id**, **MP**, and **CS**. This condition is axiomatized by these axioms, together with the following two additions.[11]

(AD) $((A > C) \wedge (B > C)) \supset ((A \vee B) > C)$
(ASC) $((A > B) \wedge (A > C)) \supset ((A \wedge B) > C)$

These completeness results are extended in [5] and [23].

In many ways, it seems implausible to assume that, when an ordering relation underlies the selection of close counterfactual states, the order is total. In view of pairs of conditionals like

If Bizet and Verdi had been compatriots, Bizet would have been Italian
If Bizet and Verdi had been compatriots, Verdi would have been French

(an example of Quine's) it certainly seems that many details are simply left unresolved in counterfactual selection. Though on theoretical grounds, it is hard to see how to justify a selection function so refined that it chooses just one state, evidence based on common sense reasoning with conditionals does tend to support Stalnaker's theory,[12] and in fact the first modern counterfactual logics to be proposed (in [30] and [33]) involved total, well-founded orderings. In particular, the condition on models is this.

(Total Order) There is a well-founded total order $\leq_w$ with least element w such that for all P, $s(P, w) = \{w : w \text{ is } \leq_w\text{-minimal in } P \text{ with respect to } \leq_w\}$.

Theories invoking total orderings validate a principle of "Conditional Excluded Middle"—that is, they make the formula

[9] See, for instance, [5] and [24].
[10] See [28] and [29], for instance.
[11] See [34] and [4].
[12] See [31].

(CEX) $(A > B) \vee (A > \neg B)$

valid. It follows from the main result of [33] that a system like our version of **VC**, but with axiom **CV** replaced by **CEX**, axiomatizes this notion of validity.

The main intuition behind the total ordering condition is that counterfactuals are not statements of necessity, but rather are *contingent, factual statements* about contrary to fact states. However, these states are dependent on the antecedent, and (according to Stalnaker) are highly sensitive to details of the context of utterance. Thus, the **Total Order** condition corresponds to the last position in the classification of conditional theories with which we began; it is a *variable material* theory.

Both Stalnaker and Lewis pay much attention to certain *conditional fallacies*—reasoning patterns that are valid in mathematical reasoning, and so are associated in the minds of logicians with the conditional, but which not only are invalid on all the selection-based semantic theories we have mentioned, but would trivialize the axiomatic theories if added to them. That is, the conditional will collapse to the material conditional, in the sense that $(A > B) \equiv (A \supset B)$ could then be proved, if any of the following three patterns is added to any of the theories we have discussed.

(Transitivity) $((A > B) \wedge (B > C)) \supset (A > C)$
(Contraposition) $(A > B) \supset (\neg B > \neg A)$
(Antececent-NM) $(A > B) \supset ((A \wedge C) > B)$

Of these fallacies, the last is probably the most interesting from the standpoint of theoretical AI, since it addresses the issue of nonmonotonicity in ways that relate quite closely to recent theories of nonmonotonic reasoning, even though the *consequence relation* of these conditional logics remains monotonic. In fact, the connection to the model preference theories of nonmonotonic consequence should be readily apparent. Indeed, if we think of $s(P, w)$ as the set of most preferred states, relative to w, in which P is true, the satisfaction condition for the conditional is an almost word-for-word paraphrase of the definition of logical consequence relation in works such as [28].

The underlying motivation for the selection function account of contrary to fact conditionals is also very similar, if not identical, to the motivation of the various approaches to nonmonotonic reasoning. If A is false at w, but true in some other states, the idea in distinguishing the selected A worlds from the others is that the selected worlds should make A true in a non-farfetched way that respects general principles about what is normally true, as well as conditions about what truths from w should be preserved, even if A is supposed. Both of these issues are crucial in thinking about nonmonotonic consequence relations.

5 Reifying possible worlds

The most direct and straightforward way to apply results from conditional logic in AI would be to develop approaches to problems—such as planning, causal reasoning, or diagnosis—that seem to rely heavily on reasoning with conditionals, and to look for either semantic solutions that involve closeness relations between worlds, or proof theoretic approaches relying on theorem proving in a conditional logic like the ones described in the previous section. Since the model theory of conditionals presupposes

both possible worlds and similarity relations among them, a semantic approach in AI based on this work would have to reify the worlds as more concrete symbolic structures, and then seek to define similarity relations among these structures in such a way that a reasonable account of counterfactual reasoning can be obtained.

This line of research was initiated by Matthew Ginsberg, who develops in [16] an account of counterfactual conditionals loosely modeled on David Lewis's variably strict theory.

In Ginsberg's account, worlds are reified as consistent sets of formulas, denoted here by capital Greek letters. Truth at a world of a formula not involving conditionals is represented using ordinary entailment; the formula A is true at the world Γ just in case $\Gamma \vdash A$.[13] For a conditional formula of the form $A > B$, Ginsberg follows Lewis's lead, evaluating the conditional as true at Γ just in case B is true at all the worlds nearest to Γ at which A holds. However, rather than taking the notion of nearness involved here for granted, as an unanalyzed background feature, he attempts to define it explicitly.

The intuitive idea is that the nearest A-worlds to Γ are those that make A true, and which, in addition, retain as much as possible of the original information from Γ, consistent with A's being true. This leads to following definition of a selection function:

$$s(\llbracket A \rrbracket, \Gamma) = \{\Delta \cup \{A\} : \Delta \subseteq \Gamma \ \& \ \Delta \nvdash \neg A \ \& \ (\Delta \subset \Theta \subseteq \Gamma \Rightarrow \Theta \vdash \neg A)\}.$$

And a conditional of the form $A > B$ is then defined as true at a world Γ just in case B is true at each member of $s(\llbracket A \rrbracket, \Gamma)$. If, in the present context, we define $\llbracket A \rrbracket = \{\Gamma : \Gamma \vdash A\}$, this truth definition takes on the familiar form: $A > B$ is true at Γ just in case

$$s(\llbracket A \rrbracket, \Gamma) \subseteq \llbracket B \rrbracket.$$

This simple analysis actually yields a surprisingly robust theory of counterfactual entailment. Of course, it applies only to first degree conditionals (in fact, to those with the conditional as their main connective, and with no nested occurrences of the conditional); but Ginsberg is able to show that, if we define the valid statements as those true at each possible world or formula set, then the set of validities according to his analysis coincides with the first degree fragment of a logic very close to Lewis's. In addition, a system based on this analysis for counterfactual reasoning in a syntactically restricted language has been implemented. Ginsberg describes the implementation in [16], along with an application to diagnostic problem solving; and together with David Smith in [17] and [18], he describes ways in which this analysis of counterfactuals can be applied in the treatment of certain planning problems, in particular the frame and qualification problems.

Unfortunately, though, because of its method of defining similarity among worlds, Ginsberg's theory of counterfactuals does not always yield intuitively correct conclusions. The simple syntactic measure of nearness among worlds is too naive; the measure treats all formulas as equally important, but in fact, some are more important than others, and some way is needed to avoid abandoning these more important worlds in favor of ones that are intuitively farther away. This is especially evident in the case of laws, which should certainly take precedence over factual background conditions, and should

[13] Of course, this yields a notion of truth that, in general, is not bivalent; it is not the case that any formula or its negation will be true at any world. However, for the kind of syntactically restricted languages common in AI, bivalence can often be restored through the closed world assumption or one of its generalizations.

not be deleted to preserve consistency when background facts can be deleted instead. Of course, Ginsberg is aware of this problem concerning laws, and it is actually rather easy to solve; he simply rules out of consideration those formula sets from which laws have been deleted. Formally, he introduces a "badworld" predicate, and then complicates the selection function defined above to guarantee that it never picks badworlds; but it is just as easy, given some particular set of laws that must be preserved, to define the possible worlds as the consistent formula sets that contain those laws.

Even once the laws are protected, however, the treatment of nearness underlying Ginsberg's analysis still leads to incorrect results—and in fact, it turns out that the problems that arise in this account are identical to those faced by Goodman forty years earlier.

The best way to see this is to recast Goodman's match example within Ginsberg's framework; so let us suppose again that the actual world is governed by the law that a dry match struck in the presence of oxygen lights. Also suppose that in fact oxygen is present and some particular match is dry, but that it is not struck and so does not light. This gives as our actual world the formula set

$$\begin{aligned} \Gamma \ = \ & \{O, \neg S, D, \neg L \\ & (D \wedge O \wedge S) \supset L\}. \end{aligned}$$

When we want to consider what would happen if the match were struck, though, the selection function, even once it is modified to preserve laws, gives us $s(\llbracket S \rrbracket, \Gamma) = \{\Delta_1, \Delta_2, \Delta_3\}$, where

$$\begin{aligned} \Delta_1 &= (\Gamma \cup \{S\}) - \{\neg S, \neg L\}, \\ \Delta_2 &= (\Gamma \cup \{S\}) - \{\neg S, O\}, \\ \Delta_3 &= (\Gamma \cup \{S\}) - \{\neg S, D\}. \end{aligned}$$

Evidently, then, we cannot conclude that $S > L$—telling us that the match would light if it were struck—is true at the original world Γ; we are prevented from drawing this conclusion by interference from the worlds Δ_2 and Δ_3, which are inappropriately characterized as no further away from the original world than Δ_1.

Though Ginsberg does not mention the connection with Goodman, he provides a different example that illustrates the same difficulty.[14] The solution he suggests is to replace his simple syntactic measure of similarity, which treats all facts true in the original world as equally important, with a measure that also reflects some information about the relative importance of those facts. Formally, he introduces a partial ordering $\preceq$ on worlds which extends the ordering based on set-theoretic inclusion, but which is supposed to reflect information about the relative importance of various facts as well. A new selection function is then based on this new ordering:

$$s'(\llbracket A \rrbracket, \Gamma) = \{\Delta \cup \{A\} : \Delta \subseteq \Gamma \ \& \ \Delta \nvdash \neg A \ \& \ (\Delta \prec \Theta \preceq \Gamma \Rightarrow \Theta \vdash \neg A)\}.$$

In the case of Goodman's match example, we would have

$$\Delta_1 \prec \Delta_2, \Delta_3;$$

this gives us $s'(\llbracket S \rrbracket, \Gamma) = \{\Delta_1\}$, so that $S > L$ is then true at Γ, as desired.

[14] The existence of such problems would not surprise anyone familiar with the history of nonmonotonic logic, since experience with a variety of formalisms in that area have made clear the need for ways of incorporating "preferences" among worlds, or extensions.

Although this maneuver does, in a sense, handle this local problem, it is hard to make much of it as a general strategy for solving this kind of difficulty. It should be obvious, for example, that Ginsberg's introduction of the $\prec$ ordering to select the most plausible among those maximal formula sets consistent with some counterfactual antecedent is equivalent to Goodman's use, discussed earlier, of the notion of cotenability to solve the same problem. Thus, this solution is exposed to many of the same objections. It cannot be said to provide an analysis of counterfactual reasoning, since it relies on an ordering among states that already reflects counterfactual ideas. This more philosophical problem is related to the approach's lack of generality as a solution to problems of AI. It is unclear how in general the ordering is to be constructed; it seems to represent only *ad hoc,* domain dependent information. Certainly this seems to be true in the match example; without a general principle for preferring Δ_1 over Δ_2 and Δ_3, or a general methodology for constructing preferences in other domains, we can't be sure that this *ad hoc* ordering on formula sets will generalize to other cases or scale up, even though it may correctly reflect our intuitions about the true counterfactuals in this particular example.

Because Ginsberg is not attempting to provide an analysis of counterfactual reasoning, but instead, to develop a theory motivated by the semantic accounts of conditionals that can be applied in AI, his program has a line of continuation that was not available to Goodman. By concentrating on limited and well-structured domains, perhaps Ginsberg can find ways of characterizing the plausibility ordering that are not *ad hoc.* One of the domains that Ginsberg considers, for example, is that of circuit diagnosis. In this domain, it is reasonable to prefer explanations of anomalous behavior that involve the failure of minimal numbers of components. If the formula sets represent statements to the effect that various components are working properly, this preference criterion may allow us to extend the simple subset ordering to a richer ordering in a principled way: not only do we prefer maximal subsets of the original formula set, but even among maximal subsets, we prefer those that postulate a greater number of working components.[15]

Some AI applications, then, seem to allow an approach much like Goodman's to be carried through without circularity.[16] Although counterfactuals are analyzed in terms of a notion like cotenability, in these limited and well-structured domains, cotenability can be defined independently. On the other hand, many of the application areas that Ginsberg envisages for his theory, such as general planning problems, are not so well-structured; and here, the kinds of difficulties that Goodman raised about grounding the theory of counterfactuals in a notion of cotenability without being able to provide an independent analysis of this notion seem to confront Ginsberg's proposals as well.

6 Conditionals and nonmonotonicity

Connections between conditional logics and AI like those described in the previous section are based on the use of conditionals in describing counterfactual, or subjunctive, reasoning. The other primary application of conditional logics in AI focuses on the use of

[15]If some components are more likely to fail than others, we can use statistics on component reliability to construct a weighted estimate that is better than number of components; Ginsberg explores this possibility in his paper.

[16]We would expect these to be the domains in which a Bayesian analysis would also be feasible.

these logics as a vehicle for nonmonotonic reasoning; and here the emphasis is on indicative conditionals. In our discussion of conditional fallacies we have already seen some analogies between conditional logics and familiar patterns of nonmonotonic reasoning. As in nonmonotonic logics, both transitivity and strengthening in the antecedent fail for conditionals; contraposition, which is at least debatable in the nonmonotonic context, fails also for conditionals.

The idea that these analogies could be exploited systematically, and that conditional logic could be used as a foundation for nonmonotonic reasoning, is due originally to Stalnaker's well-known but unpublished [32]. The general proposal set out in that paper is that common sense defaults can be represented by universally quantified conditionals from one of the standard theories such as Lewis's or Stalnaker's. For example, the defaults that birds fly and penguins don't might be represented as (i) $\forall x(Bx > Fx)$ and (ii) $\forall x(Px > \neg Fx)$.

This representation has certain advantages. Even if we accept also a statement like (iii) $\forall x \Box(Px \supset Bx)$, telling us that all penguins must be birds, our set of formulas is still consistent. We are not forced to conclude $\forall x(Px > Fx)$, because of the failure of transitivity, even when one of the premises is strict.

But, as Stalnaker notes, this idea suffers from glaring deficiencies: together, our three formulas logically imply that there are no penguins. Any standard conditional logic validates **MP**, so that from (i) and (ii) we can infer $\forall x(Bx \supset Fx)$ and $\forall x(Px \supset \neg Fx)$; classical logic then yields $\neg\exists x(Bx \wedge Px)$. In his unpublished paper, Stalnaker sketches an approach to solving this problem about reasoning with individuals.[17]

Since *modus ponens* plays a central part in deriving these unwelcome consequences, it is natural to represent defaults in a weaker conditional logic, in which **Relative Reflexivity** (and hence, **Centering**) is absent and *modus ponens* fails. Of course, this surrenders the idea that the 'if' of a default is a genuine conditional. But it would allow us to say that defaults involve a conditional modality (perhaps a conditional form of belief, or of normalcy), with logical properties very similar to that of counterfactual 'if'. To mark the difference between the "alethic" conditionals that satisfy *modus ponens* and the "deontic" ones that do not, we will use $>\!\!\circ$ for the latter.

James Delgrande's [12] presents the first sustained development of this approach. Unlike Ginsberg, Delgrande does not attempt to reify the worlds as more concrete structures. But, unlike the developers of the standard conditional logics, he does not base his analysis on a relation of similarity among worlds. Instead, he introduces a relation of relative normality. A conditional is not evaluated with respect to the worlds most similar to the home world, but instead, with respect to the worlds in which things are as normal, or typical or unexceptional, as possible. This difference is not just terminological, since it explains why **Relative Reflexivity** no longer holds; if w is abnormal in some gratuitous respects, w will not be as normal as possible, although it certainly will be as like itself as it can be.

[17]The suggestion is to abandon certain conditionals, and perhaps replace them with others, once instantiation is used to infer conclusions about individuals from conditionals. In the our example, the information that $Pt \wedge \neg Ft$ would lead to the replacement of (i) by $\forall x((Bx \wedge \neg Px) > Fx))$. However, the details of this approach are only hinted at in Stalnaker's paper; the general strategy for deciding what conditionals to replace and what to replace them with is never clarified. We believe that Stalnaker's suggestion is plausible if it is interpreted as saying that the universal quantifier in generic statements like 'Birds fly' is somehow restricted by context. But as far as we know, the idea has never been spelled out or formalized in detail.

Delgrande's analysis, then, is that a conditional $A \circ\!> B$ should be true at a particular world just in case B holds at all of the most normal worlds (from the standpoint of the original) at which A holds. Where R is the relation of relative normality (so that $w_1 R w_2$ means that the world w_2 is at least as normal as w_1), Delgrande gives this idea formal sense by defining a selection function

$$s([\![A]\!], w) = \{w_1 : wRw_1 \ \& \ w_1 \in [\![A]\!] \ \& \ (w_1 R w_2 \ \& \ w_2 \in [\![A]\!] \Rightarrow w_2 R w_1)\}$$

that picks out those most normal worlds, relative to the original, in which A holds. The truth conditions for a conditional can then be given in the standard way; $A \circ\!> B$ is true at the world w just in case

$$s([\![A]\!], w) \subseteq [\![B]\!].$$

Of course, in order to get a plausible logic, Delgrande has to impose some conditions on the relation R of relative normality. Reflexivity and transitivity both seem natural; but in addition, he imposes the requirement of forward connectedness: if $w_1 R w_2$ and $w_1 R w_3$, then either $w_2 R w_3$ or $w_3 R w_2$. This requirement seems rather unintuitive when one simply thinks in the abstract about a relation of relative normality.[18] This abstract way of thinking about semantic constraints is not the best way to evaluate a logic; it is better to look at the plausibility of the valid formulas generated by the set of constraints, and here we can find some justification for the requirement of forward connectedness. Without it, it is possible to satisfy the formula $\Diamond A \wedge (A \circ\!> B) \wedge (A \circ\!> \neg B)$, but this formula clashes with our intuitive understanding of defaults; as long as there are birds, we cannot accept both the statements 'Birds fly' and 'Birds don't fly.' The condition of forward connectedness guarantees that the formula is not satisfiable.

Delgrande provides in [12] an axiomatization of this logic that takes the conditional connective as basic. A somewhat neater presentation, however, is given by Craig Boutilier in [7]. The modal logic whose accessibility relation is governed by reflexivity, transitivity, and forward connectedness is a familiar system (known as S4.3), with a number of well-known axiomatizations. Boutilier shows that, in fact, any of these axiomatizations will suffice because, unlike the standard conditional connectives surveyed earlier, which are known not to be definable in terms of the usual modal connectives, Delgrande's conditional can be defined, as follows:

$$A \circ\!> B \ =_{df} \ \Box(\Box\neg A \vee \Diamond(A \wedge \Box(A \supset B))).$$

In addition, Boutilier establishes in this paper the rather surprising result that the first degree fragment of this logic (S4.3 with a conditional as defined above) actually coincides with the logic R of Kraus et al. [24], another conditional logic of nonmonotonicity which is limited to first degree formulas.

The use of these logics for representing defaults has a number of advantages, which are explored in the work of Delgrande, Boutilier, Kraus et al., and others. In particular, they do not succumb to the inconsistency problems that faced Stalnaker's proposal. The following set of facts, for example—representing the usual information about Tweety, birds, penguins, and flying—is consistent if the conditional is Delgrande's: (i) $Bt \circ\!> Ft$, (ii) $Pt \circ\!> \neg Ft$, (iii) $\Box(Pt \supset Bt)$, (iv) Pt. In addition, these systems handle at least certain examples of property inheritance in an intuitively correct way. From these formulas, for instance, it follows in Delgrande's logic that $(Bt \wedge Pt) \circ\!> \neg Ft$.

[18] For example, suppose we recognize the norms that birds fly and that dogs bark. Imagine that the world w_1 is populated by ten birds only eight of which fly, and ten dogs only eight of which bark. Let w_2 be like w_1 except that nine of the birds fly; let w_3 be like w_1 except that nine of the dogs bark. Although both of these other worlds seem more normal than w_1, it is hard to find any reason for thinking that either should be comparable in normality to the other.

Unfortunately, these systems also suffer from a problem that is the dual of Stalnaker's. The reason the set consisting of (i) through (iv) above is consistent in Delgrande's system, of course, is that it lacks *modus ponens* as a generally applicable rule. However, although this protects the system from inconsistency in situations like the penguin example, it limits its applicability by preventing typical conclusions to be drawn from conditionals. If it were provided with (i) and (ii), for example, as our only information, any intuitively acceptable default reasoner would have to draw the conclusion $\neg Ft$. But Delgrande's logic does not yield $\neg Ft$ as a consequence of (i) and (ii); blocking the validity of this inference in general was in fact the motivation for the logic.

Another problem, which applies to alethic conditional systems as well as the deontic ones, affects conditionals in much the same way that the qualification problem affects generalizations in a monotonic logic.[19] With a conditional that is antecedent-monotonic (either the material or C.I. Lewis's strict conditional), rules that have many exceptions are difficult to state, because all the exceptions have to be built into the antecedent. The impossibility of stating exceptions separately leads to an unpleasant lack of modularity; and in the case of open-ended generalizations, we can be sure that our formulations are false. With a conditional for which **Antecedent Nonmonotonicity** is invalid both these problems can be avoided. But a moment's thought shows that any such attempt is just as quixotic as the corresponding explicit attempt to solve the qualification problem for antecedent-monotonic conditionals. Though we would be working in the opposite direction, by adding generalizations rather than removing exceptions, the cases that would have to be enumerated explicitly are simply too overwhelming.

Because of these limitations, it is clear that these weak conditional logics must be supplemented by some mechanism for allowing a certain amount of *modus ponens* and antecedent monotonicity. The most natural solution to both problems is to explore conditional logics that are based on a nonmonotonic consequence relation, and which are engineered somehow to make **MP** and **Antecedent Monotonicity** hold as defaults.

This is an idea that has occurred in various forms to many researchers during the last few years. Recently a variety of options have been suggested and explored. Delgrande in [13] suggests two (equivalent) techniques of introducing defaults into the original system he developed in [12]; the techniques are consistency based, and somewhat like those involved in default and autoepistemic logic. Boutilier in [6] supplements the same system with a model preference criterion for default reasoning, and applies the resulting logic to inheritance reasoning. Lehmann and Magidor describe in [25] a way of introducing a certain mechanism for default reasoning into the logic of [24]. In [3], Nicholas Asher and Michael Morreau develop a semantic approach to the problem that brings in novel elements—in particular, a sophisticated treatment of ignorance and of normality-governed knowledge update.

Since these new approaches involve *both* a nonmonotonic consequence relation, which could either be based on a familiar idea (such as autoepistemic logic) or on an entirely new formalism, and a logic of the conditional, they can't be seen as providing a general

[19] See, for instance, [29], pp. 16-17.

logical foundation for nonmonotonic inference in terms of conditionals. In place of the foundational approach that Stalnaker proposed and that many of the earlier works on conditionals in AI recommended, conditionals and nonmonotonicity would be treated as logically independent phenomena, which nevertheless can interact strongly.

In fact, the interactions between the two are complex, and we do not yet have a very clear idea of what they are or how to account for them. This makes it difficult to see what is expected of the new theories mentioned above, and hard to tell whether the formal complexities that they all suffer from in one way or another are forced on us by the problem, or can be removed by further logical work. Since the theories are so new, however, it is reasonable to hope for progress in this important area.

7 Conclusion

The search for fundamental extensions to expert systems technology has been a chief focus of research in AI for the last dozen years. This work can be seen as grappling, in one way or another, with the problem of how to represent conditional knowledge and control conditional reasoning. The discovery that conditionals (in the form of rules governing expertise) could be extracted from experts and used to build systems that could simulate expert reasoning in limited tasks was vital in stimulating interest in rule-based formalisms, and the limitations of the resulting technology inspired a variety of responses, including the theoretical work on declarative representations that led, among other things, to a revival in AI of conditional logic.

The course of development of these theories, as we have traced it here, may have reached a point where it can make contact with some of the original larger, motivating problems in AI. We stressed, in our brief account of theories of the conditional based on an underlying nonmonotonic logic, that there remains much theoretical research to do in developing these relatively new logics. More challenging, though, is the need to show that these theories can be used successfully to represent domain knowledge and carry out expert reasoning, at least in domains comparable in size to those in which expert systems have been successfully deployed.

The original promise of conditional logics for AI consisted partly in the fact that much knowledge is naturally encoded in conditionals. To show that these logics are successful in the arena of applications, it needs to be demonstrated that the knowledge they require can be acquired and represented, and that implementations based on the theoretical work can carry out the relevant reasoning tasks. Perhaps, for instance, domain knowledge could be acquired in much the same way that it is obtained for a rule-based system, and could then be compiled into a rule interpreter by a program that is sound with respect to a suitable conditional logic, or that at least is sound under general conditions.

It is vital to keep the goal of a system such as this in sight, because the weakest point of conditional logic as a knowledge representation tool is its potential for making contact with workable methods of knowledge acquisition and feasible reasoning. Without a large scale attempt to deploy and test conditional logics in systems that carry out domain reasoning, we will not be able to tell whether the new combinations of nonmonotonic and conditional logic can generalize rules adequately. Will plausible, acquirable domain axioms allow us to infer from the rule that a match will light if it is struck, that *this* kitchen match, about to be struck against my doorstep, will light? Using rule-based technology, we know that the right amount of generalization can be obtained by *ad hoc* procedural techniques, which however are brittle and do not extend easily even to closely related domains. Whether we can obtain the right amount of generalization in a system based on a conditional logic, and whether this will improve the flexibility and extensibility of knowledge bases and reasoning systems, is something that can ultimately only be established by a large program of experimentation.

Formalizing domains and developing implementations may also help to work out the relationship of the more qualitative tradition of conditional logic to the Bayesian, probability-based tradition. We know, from Pearl and Geffner's extensions, described in [15], of older results due to Adams, that there are deep relationships between Bayesian theories on the one hand, and conditional and nonmonotonic logics on the other. The hope of the conditional logic community is that the logical approach will be more general, and in particular will apply to cases in which probabilities are not available. However, as we saw in discussing Ginsberg's account of circuit troubleshooting, a thorough formalization based on conditional logic of a domain in which probabilities are at work may force us to reintroduce these probabilities in some form. To show that conditional logic gives us something more, and to advance the debate between the probabilists and the conditionalists, we will need to see whether the former approach can be made to produce results in domains in which no one is able to use probabilities successfully.

Acknowledgments

This material is based on work supported by the National Science Foundation under grant IRI-9003165.

References

[1] E. Adams. *The Logic of Conditionals.* D. Reidel Publishing Co., Dordrecht, 1975.

[2] A. R. Anderson and N. D. Belnap, Jr. *Entailment,* vol. 1. Princeton University Press, Princeton NJ, 1980.

[3] N. Asher and M. Morreau. Commonsense entailment: a model of non-monotonic reasoning. In J. Mylopoulos and R. Reiter, eds., *IJCAI 1991: Proceedings of the Twelfth International Conference on Artificial Intelligence,* Morgan Kaufmann, San Mateo CA, 1991, pp. 387–392.

[4] J. Bell. The logic of nonmonotonicity. *Artificial Intelligence,* vol. 41 (1990), pp. 365–374.

[5] J. Bell. Pragmatic logics. In J. Allen, R. Fikes and E. Sandewall eds., *KR'91: Principles of Knowledge Representation and Reasoning,* Morgan Kaufmann, San Mateo CA, 1991, pp. 50–60.

[6] C. Boutilier. On the semantics of stable inheritance reasoning. Technical Report KRR-TR-89-11, Department of Computer Science, University of Toronto, 1989. An abbreviated version appears in T. Dietterich and W. Swartout, eds., *AAAI 1990: Proceedings of the Eighth National Conference on Artificial Intelligence,* Morgan Kaufmann, San Mateo CA, 1989. pp. 594–599

[7] C. Boutilier. Viewing conditional logics of normality as extensions of the modal system S4. Technical Report KRR-TR-90-4, Department of Computer Science, University of Toronto, 1990. An abbreviated version appears in *Proceedings of AAAI-90*, The MIT Press, 1990.

[8] J. Burgess. Quick completeness proofs for some logics of conditionals. *Notre Dame Journal of Formal Logic*, vol. 22 (1981), pp. 76–84.

[9] R. Carnap. Testability and meaning. *Philosophy of Science*, vol. 3 (1936), pp. 419–471, and vol. 4 (1937), pp. 1–40.

[10] B. Chellas. Basic conditional logic. *Journal of Philosophical Logic*, vol. 4 (1975), pp. 133–153.

[11] R. Chisholm. The contrary-to-fact conditional. *Mind*, vol. 55 (1946), pp. 289–307.

[12] J. Delgrande. A first-order conditional logic for prototypical properties. *Artificial Intelligence*, vol. 33 (1987), pp. 105–130.

[13] J. Delgrande. An approach to default reasoning based on a first-order conditional logic: revised report. *Artificial Intelligence*, vol. 36 (1988), pp. 63–90.

[14] P. Gärdenfors. *Knowledge in Flux*. The M.I.T. Press, Cambridge MA, 1988.

[15] H. Geffner. *Default Reasoning: Causal and Conditional Theories*. Ph.D. Dissertation, Computer Science Department, UCLA, Los Angeles, 1989. Available as Technical Report 137, Cognitive Systems Laboratory, UCLA, Los Angeles, 1989.

[16] M. Ginsberg. Counterfactuals. *Artificial Intelligence*, vol. 30 (1986), pp. 35–79.

[17] M. Ginsberg and D. Smith. Reasoning about action I: a possible worlds approach. *Artificial Intelligence*, vol. 35 (1988), pp. 165–195.

[18] M. Ginsberg and D. Smith. Reasoning about action II: the qualification problem. *Artificial Intelligence*, vol. 35 (1988), pp. 311–342.

[19] N. Goodman. The problem of counterfactual conditionals. *Journal of Philosophy*, vol. 44 (1947), pp. 113–128. Reprinted in [20], pp. 3–27.

[20] N. Goodman. *Fact, Fiction and Forecast*. Harvard University Press, Cambridge MA, 1955.

[21] R. Thomason and A. Gupta. A theory of conditionals in the context of branching time. *Philosophical Review*, vol. 88 (1980), pp. 65–90. Reprinted in [22], pp. 299–322.

[22] W. Harper, R. Stalnaker, and G. Pearce, eds. *Ifs: Conditionals, Belief, Decision, Chance, and Time*. D. Reidel Publishing Company, Dordrecht, 1981.

[23] H. Katsuno and K. Satoh. A unified view of the consequence relation, belief revision and conditional logic. In J. Mylopoulos and R. Reiter, eds., *Proceedings of the Twelfth International Joint Conference on Artificial Intelligence*, Morgan Kaufmann, San Mateo CA, pp. 406–412.

[24] S. Kraus, D. Lehman, and M. Magidor. Nonmonotonic reasoning, preferential models, and cumulative logics. *Artificial Intelligence*, vol. 44 (1990), pp. 167–207.

[25] D. Lehmann and M. Magidor. What does a conditional knowledge base entail? *Artificial Intelligence*, forthcoming.

[26] D. Lewis. *Counterfactuals*. Oxford University Press, Oxford, 1973.

[27] D. Nute. Conditional logic. In D. Gabbay and F. Guenthner, eds., *The Handbook of Philosophical Logic*, vol. 2, D. Reidel Publishing Co., Dordrecht, 1984, pp. 392–439.

[28] Y. Shoham. A Semantical Approach to Nonmonotonic Logics. In J. McDermott, ed., *Proceedings of the Tenth International Joint Conference on Artificial Intelligence*, Morgan Kaufmann, San Mateo CA, 1987, pp. 388–392.

[29] Y. Shoham. *Reasoning about Change*. The MIT Press, Cambridge MA, 1988.

[30] R. Stalnaker. A theory of conditionals. In *Studies in Logical Theory*, American Philosophical Quarterly Monograph Series, No. 2, Basil Blackwell, 1968, pp. 98–112. Reprinted in [22], pp. 41–55.

[31] R. Stalnaker. A defense of conditional excluded middle. In [22], pp. 87–104.

[32] R. Stalnaker. A note on nonmonotonic modal logic. Manuscript, Department of Linguistics and Philosophy, MIT, 1980.

[33] R. Stalnaker and R. Thomason. A semantic analysis of conditional logic. *Theoria*, vol. 36 (1970), pp. 23–47.

[34] F. Veltman. *Logics for Conditionals*. Ph.D. Thesis, University of Amsterdam, Amsterdam, 1985.