On the Semantics of Counterfactuals

Peter Jackson*
McDonnell Douglas Research Laboratories
Dept 225, Bldg 105/2, Mailcode 1065165
P.O. Box 516, St Louis, MO 63166, USA

Abstract

It is argued that Ginsberg's Possible Worlds Approach to counterfactual implication suffers from a number of defects which are the result of confusing proof theory and model theory. In particular, logically equivalent theories do not have identical counterfactual consequences, and monotonic theory revisions are not always preferred to nonmonotonic ones. This paper develops a situation semantics for counterfactual implication in which propositions are treated as operations on sets of possible worlds. Logically equivalent theories have identical consequences in the model theory, which always prefers monotonic revisions to nonmonotonic ones and validates all the axioms and derivation rules of counterfactual logic. The semantics is also contrasted with Winslett's Possible Models Approach.

1 Introduction

Counterfactuals are conditional statements in which the antecedent is deemed to be false, e.g. 'If Waldo were rich, he'd live in Las Vegas'. Considered as material conditionals, all such statements are trivially true. Thus 'If Waldo were rich, he'd live in Milwaukee' would also be a true statement, if Waldo were not rich. Yet the intention seems to be: 'if things were more or less as they are, except that Waldo were rich, he'd be living in Las Vegas', which rules out living in Milwaukee. It is customary to write a counterfactual conditional of the form 'if ψ then ϕ' as '$\psi > \phi$' rather than '$\psi \supset \phi$', in order to preserve this distinction.

Counterfactuals are not truth functional, because they can only be evaluated relative to (i) some theory of what the world is like, and (ii) some notions about what the world *might* be like if certain things were to change. Informally, we want $\psi > \phi$ to be true if and only if ϕ is true in all those plausible worlds where ψ holds which are most similar to the real world. From a formal point of view, the problem is to specify what we mean by 'plausible' and 'similar'.

*This work was supported by the McDonnell Douglas Independent Research and Development program.

Ginsberg's [1986] paper was perhaps the first to demonstrate the relevance of counterfactual reasoning to a range of AI applications. Its scope is considerable, including a review of the literature, an account of counterfactual implication, a discussion of implementation issues, and a survey of applications. In this paper, we shall be concerned only with the account of counterfactual implication, which we feel to be flawed in corrigible ways.

Our theses are the following: (i) that the account contains a confusion between proof and model theory; and (ii) that the theory revisions sanctioned are not always minimal.

Our method shall be: (i) to work though the main examples of the original paper, pointing out where difficulties lie; (ii) to present a semantics for counterfactual implication which alleviates these difficulties and compare it with that of Winslett [1988]; and (iii) to show that the model theory satisfies a well-known axiomatization of counterfactual logic.

2 Critique of Ginsberg's construction

We saw in Section 1 that counterfactual statements are not truth-functional, so we need some kind of *construction,* consisting of possible worlds other than the current world, which we can inspect in order to decide whether or not a counterfactual holds in the current world.

2.1 Counterfactual implication

Ginsberg's construction for counterfactuals has the following components: an initial set S of sentences; a predicate B on 2^S, the power set of S; and a partial order $<$ on 2^S which extends set inclusion and respects B. S is a theory of the world, B is a 'bad world' predicate which declares some worlds to be implausible, and $<$ is a comparator of possible worlds along the plausibility dimension. The definition of plausible, similar worlds is as follows.

Definition 1 A *possible world for ψ in S* is any subset T of S such that $T \not\models \neg\psi$, $\neg B(T)$, and T is maximal with respect to $<$, given these restrictions. The set of such worlds, $W(\psi, S)$, is given by

$$\{T \subseteq S \mid T \not\models \neg\psi \wedge \neg B(T) \wedge \{U \subseteq S \mid T < U \wedge U \not\models \neg\psi \wedge \neg B(U)\} = \emptyset\}.$$

A counterfactual $\psi > \phi$ is now true with respect to S iff $T \cup \{\psi\} \models \phi$ for all $T \in W(\psi, S)$.

Treating 2^S as a set of possible worlds confuses syntactic objects, such as sets of sentences, with semantic objects, such as models. In fact, each $T \in 2^S$ represents a set of worlds in its own right, namely those worlds that satisfy every sentence in T. It can be dangerous to treat arbitrary theories as 'partial possible worlds', because theories are more complex than models. This construction is known as the Possible Worlds Approach (PWA), although the 'worlds' in $W(\psi, S)$ are partial. The basic idea is that if $T \in W(\psi, S)$ then T represents a minimum revision to S such that $T \cup \{\psi\}$ is consistent, T is not implausible, and no other revision of S is more plausible than T.

In the rest of this paper, a possible world ω will be represented by a set of atomic propositions, $\{\psi_1, ..., \psi_n\}$, such that if ψ is an atom then ω satisfies ψ, written $\omega \models \psi$, if and only if $\psi \in \omega$, and if $\psi \notin \omega$ then $\omega \models \neg\psi$. Such worlds are essentially propositional calculus models as specified in Chang & Keisler [1973, Ch.1]. Satisfaction conditions for a compound statement ψ follow the normal truth-functional recursion on the complexity of ψ.

2.2 Semantics of counterfactuals

In this subsection, we review the model theory of Ginsberg's construction, as well as Winslett's reconstruction of the semantics of counterfactual consequence.

2.2.1 PWA and equivalent theories

Ginsberg notes that logically equivalent theories do not have identical counterfactual consequences in PWA.

Example 1 Let $S = \{p, p \supset q\}$ and $T = \{p, q\}$. $\neg q > p$ is a counterfactual consequence of T but not of S. Neither is it the case in general that a theory and its deductive closure have the same counterfactual consequences.

Ginsberg writes: 'It does not seem to me that this dependence upon representation is inappropriate when investigating counterfactuals, so that we should not be overly concerned over the fact that our construction depends upon more than merely the model-theoretic information contained in the theory S'.

On the contrary, we believe that the counterfactual consequences of a theory should not depend upon the vagaries of its syntactic representation, and construe this dependency as further evidence of a confusion between syntactic and semantic entities. Section 3 presents a semantics in which logically equivalent theories have identical counterfactual consequences.

Example 2 Let t denote 'thunder' and l denote 'lightning'. An agent unaware of the connection between them would describe the world as $S = \{t, l\}$, giving $\neg t > l$.

Ginsberg uses this example to suggest that without the vices of his model theory, it would be impossible for any fact to be irrelevant to any other. This turns out not to be the case. Section 3 presents a model theory without these vices in which irrelevance is possible.

2.2.2 The Possible Models Approach

Winslett's [1988] Possible Models Approach (PMA) attempts to regularize the model theory of counterfactual consequence. Given a set of formulas S and a theory T, PMA computes a set of models Incorporate(S, M) produced by incorporating S into T, where M is a model of T. The formal definition can be expressed as follows.

Definition 2 Let T be a theory with protected formulas $T^* \subset T$, let M be a model of T, and let S be a set of formulas. Incorporate(S, M) is the set of all models M' such that

(i) $M' \models S$ and $M' \models T^*$; and
(ii) no other model satisfying (i) differs from M on fewer atoms

where 'fewer' is defined by set inclusion. If Models(T) is the set of all models of T, then the set of models of the revised theory is given by: $\cup_{M \in \text{Models}(T)}$ Incorporate(S, M).

It is easy to verify that logically equivalent theories have identical counterfactual consequences in PMA. However, we shall argue that its results are sometimes counterintuitive, especially when we iterate the conditional operator. We return to Winslett's construction in Section 3.2 (Example 4), and show that its revisions are not always minimal.

3 Situation semantics for counterfactuals

The following account assumes a propositional language L, defined over a finite alphabet A. We can construct $2^{|A|}$ interpretations over this alphabet, and consider each as a possible world. Let this set of interpretations be W. The empty theory, Ø, is satisfied by every interpretation, so we label this set of worlds $W_\emptyset$, i.e. the class containing all models of Ø. Clearly $W_\emptyset = W$. A non-empty theory $S \subset L$ describes a *situation*, $W_S \in 2^W$. Thus $W_S \subset W$ is the set containing just those possible worlds which satisfy S.

The easiest way to formalize situations is in terms of model-theoretic forcing [Keisler, 1977]. If S is a theory, then a *condition* for S is a finite set of literals, C, consistent with S, and C ||- ψ denotes that C *forces* ψ, i.e. that S, C |- ψ. G is a *generic set* for S iff each $H \subseteq G$ is a condition for S and G ||- ψ or G ||- $\neg\psi$ for all propositions ψ.

Definition 3 If S is a theory, then a *situation for S*, W_S, is a set of possible worlds constructed as follows. For each generic set G for S, we can construct a world

$$\omega = \{\psi \mid S, G \vdash \psi\}$$

where ψ is an atomic proposition. W_s is the set of such worlds.

The semantics that we shall give for counterfactuals of the form $\psi > \phi$ with respect to a theory S depends upon a very simple idea. We consider ψ as a *revision function* that we can apply to S to return those plausible worlds where ψ holds which are most similar to some world in W_s. $\psi > \phi$ is then a consequence of S just in case ϕ holds in each of these worlds.

3.1 Propositions as functions

A proposition ψ can be considered as a function, $\psi: 2^W \rightarrow 2^W$, from situations to situations. For example, if $S \subset L$ is $\{p \supset q\}$, then p(S) denotes $p(W_s)$, where *p* is the function associated with P. We want the value of $p(W_S)$ to be the set of worlds most similar to S in which *p* holds. If *A* = {p, *q]* is the alphabet of L, then W_s = {Ø, *[q]*, {p, *q)*}, and the application of *p* to W_s should return {{p, *q]],* the only model of $\{p, p \supset q\}$.

The above example was rather straightforward, *p* was consistent with S, and we did not place any restrictions upon the range of/?, i.e. all worlds in which *p* held were deemed to be plausible. As a result, we can consider the computation performed by *p* as an instance of forcing: *{p}* was a condition for S that forced *q.* But we are most interested in the case where *p* is inconsistent with S, and cannot therefore be a condition for S. What general properties should propositions considered as *revision* functions possess?

P1. If S *logically implies* ψ, then there is no need to change W_s. Otherwise, some revision must be effected, else there will be a world in Ws which is not a model of Ψ-

P2. If ψ is *consistent* with S but does not follow from it, then we compute a new situation $\psi(S) = W_{\{\psi\}} \cap W_S$, containing all the possible worlds which satisfy $\{\psi\} \cup S$.

P3. If ψ is *inconsistent* with S, then we compute some minimum revision, $\psi(S)$, of W_s such that $\psi(S) \subseteq W_{\{\psi\}}$. This inclusion must hold if ψ is to be true at every world in $\psi(S)$. If ψ is inconsistent with *protected* propositions in S, then $\psi(S) = \emptyset$.

Let us concentrate on P3, since the other three cases are straightforward. $\psi(S)$ cannot be just any subset of $W_{\{\psi\}}$, because there may be propositions in S that we wish to protect. If $S^* \subset$ S is the (proper) subset containing the protected propositions, then we require that $\psi(S) \subseteq (W_{\{\psi\}} \cap W_{S*})$, i.e. $\psi(S)$ must contain only plausible worlds.

We can generalize the notion of a revision function as follows. Rather than generating an individual function for each proposition, and composing these functions for compound propositions, we introduce a two-place operation, $\Rightarrow$, upon sets of worlds, such that $\psi(S) = (W_{\{\psi\}} \Rightarrow W_S)$. In so doing, we use the notion of a *world lattice:* If $S \subset L$ and $B \subseteq A$ contains those members of the alphabet of *L* occuring in S, then $\Lambda_S = (2^B, \subset)$ is a world lattice for S.

Definition 4 Let $S^* \subset S \subset L$, $\psi \in L$ and $W_{\{\psi\}} \cap W_S = \emptyset$. Then $\psi(S) = W_{\{\psi\}} \Rightarrow W_S$ contains just those worlds $\omega \in (W_{\{\psi\}} \cap W_{S*})$ that satisfy the following condition:

There is a world $\upsilon \in W_S$ such that
(i) a) is a glb or lub of υ in Λ_S; *or*
(ii) there is a world $\varpi \in W_{\{\psi\}}$ such that
(a ϖ s a glb or lub of v i Λ_S n d
(b) $\varpi \notin W_s^*$ and
(c ω s the smallest superset or largest subset of ϖ in $(W_{\{\psi\}} \cap W_{S*})$

where glb denotes the greatest lower bound and lub denotes the least upper bound.

Thus each world in $W_{\{\psi\}} \Rightarrow W_S$ is a world from $W_{\{\psi\}}$. For each world ω i $W_{\{\psi\}}$, ω in the revision if and only if: either (i) ω satisfies the protected propositions and is a *maximal subset* or *minimal superset* of a world in Ws; or (ii) there is an implausible world ϖ in $W_{\{\psi\}}$ which is a maximal subset or minimal superset of a world in Ws, and ω is a maximal subset or minimal superset of ϖ with respect to the plausible worlds. Hence the worlds most similar to worlds in W_s are those plausible worlds which are closest to such worlds in the lattice.

Each proposition is therefore a function of the following kind.

Definition 5 If ψ is a proposition and S a theory, then

$$\psi(S) = W_{\{\psi\}} \cap W_S \text{ if } S \not\models \neg\psi, \text{ else } W_{\{\psi\}} \Rightarrow W_S.$$

The reader can verify that the definition satisfies each of the properties P1-P3.

Definition 6 $\psi > \phi$ is a (r f a c t u a l consequence of S iff $\omega \models \phi$ for all $\omega \in \psi(S)$

This completes our semantic account of counterfactual consequence; we now return to the examples introduced in Section 2.2.1. The recomputation of the set of plausible, similar worlds for each case will illustrate the model theory, as well as demonstrating its advantages. Let us agree to call the logic of Definitions 3-6 'Belief Revision Logic', or BERYL for short.

3.2 The examples revisited

In this subsection, we show that the problems noted in Section 2 do not arise in BERYL.

Example 1 revisited If S = $\{p, p \supset q\}$, T = {p, *q}* and there are no protected propositions in either theory, then

$$\neg q(S) = \{\emptyset, \{p\}\} \Rightarrow \{\{p, q\}\} = \{\{p\}\} = \neg q(T).$$

Thus $\neg q > p$ is a counterfactual consequence of both theories. It is easy to see that if S and T are *any* pair of equivalent theories, then they will have identical counterfactual consequences, so long as S^* and T^* arc equivalent. (This result also holds for Winslett's PMA.)

Theorem 1 If S and T are equivalent propositional theories and S* and T* are equivalent, then for all propositions ψ and ϕ, $\psi > \phi$ is a consequence of S iff $\psi > \phi$ is a consequence of T.

Proof Follows straightforwardly from Definitions 3-6. If the antecedent holds, then $\psi(S) = \psi(T)$. Thus $\omega \models \phi$ for all $\omega \in \psi(S)$ iff $\omega \models \phi$ for all $\omega \in \psi(T)$.

Example 2 revisited If $S = \{t, l\}$ and $S^* = \emptyset$, then

$$\neg t\,(S) = \{\emptyset, \{l\}\} \Rightarrow \{\{t, l\}\} = \{\{l\}\}$$

so $\neg t > l$, and t is irrelevant to l. But if $T = \{l, \neg t\}$ and $T^* = \emptyset$, then

$$(t \equiv l)(T) = \{\emptyset, \{t, l\}\} \Rightarrow \{\{l\}\} = \{\emptyset, \{t, l\}\}$$

and t and l are now dependent, and will remain so as long as $t \equiv l$ is protected. But if $t \equiv l$ is not protected, then it will be retracted in the face of $\neg t$ or $\neg l$. Note also that, if $t \equiv l$ is not protected, we have $(t \equiv l)(\neg t\,(S)) \neq \neg t\,((t \equiv l)(S))$, even though $\neg t$ and $t \equiv l$ are consistent:

$$(t \equiv l)(\neg t\,(S)) = \{\emptyset, \{l, t\}\}$$
$$\neg t\,((t \equiv l)(S)) = \{\{l\}\}.$$

It follows that $\neg t > ((t \equiv l) > l)$ is a counterfactual consequence of S, but not $(t \equiv l) > (\neg t > l)$. The order in which we entertain the two hypotheses makes a difference.

The point is that the outcomes in Example 2 are identical to those that would be derived by PWA. Thus the fact that equivalent theories have identical counterfactual consequences in BERYL has no bearing on the issue of irrelevance. Nor does it prevent counterfactual consequence from being sensitive to either the order in which we entertain hypotheses or the protection of certain propositions in the face of hypotheses. (The same results hold for PMA.)

The next example highlights an essential difference between PWA and BERYL.

Example 3 Let $S = \{p \supset q, \neg p\}$, $T = \{p \supset q, \neg p, \neg q\}$, and $S^* = T^* = \emptyset$.

In PWA, there is a single revision of S in the light of p: we retract $\neg p$. But there are two alternative revisions of T: we can retract $\neg q$ or we can retract $p \supset q$. The possible worlds associated with the revised theories S' and T' are $\{\{p, q\}\}$ and $\{\{p\}, \{p, q\}\}$ respectively. By contrast, in BERYL:

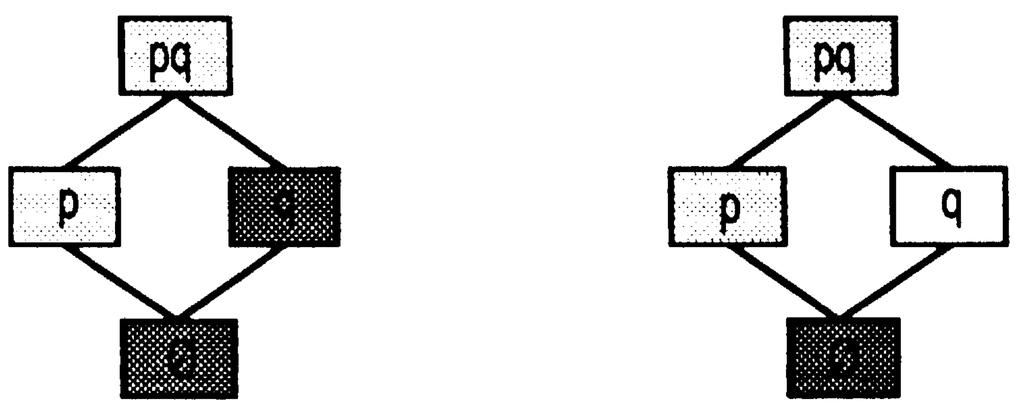
$$p(S) = \{\{p\}, \{p, q\}\} \Rightarrow \{\emptyset, \{q\}\} = \{\{p\}, \{p, q\}\}$$
$$p(T) = \{\{p\}, \{p, q\}\} \Rightarrow \{\emptyset\} = \{\{p\}\}.$$

Thus $p > q$ is a counterfactual consequence of S in Ginsberg's method but not in BERYL, while $p > \neg q$ is a consequence of T in BERYL but not in Ginsberg's method. This conflict is worth examining in more detail. The reader is invited to consult the world lattices of Figure 1, where dark shading signifies the models of the theory in question, light shading signifies models of the counterfactual premise.

With respect to S, the minimal change is to retract $\neg p$ *and* $p \supset q$. From a semantic point of view, $\{p\}$ is the most similar world to $\emptyset$ and $\{p, q\}$ is the most similar world to $\{q\}$. $\{p, q\}$ is not the most similar world to $\emptyset$, and so moving from $\emptyset$ to $\{p, q\}$ is not a minimal change. This much is clear from Figure 1a.

BERYL's judgement is also justifiable from a syntactic point of view, so long as that view is *purely* syntactic. In the propositional calculus, $p \supset q$ is just another way of writing the disjunction $\neg p \vee q$. Relinquishing $\neg p \vee q$ must involve a smaller change to S than retaining it and admitting q, since $q \vdash \neg p \vee q$.

A similar argument holds with respect to the revision of T. Figure 1b shows that $\{p\}$ is the world closest to $\emptyset$. The minimal change is to reliquish $p \supset q$ before admitting q.



(a) (b)

Figure 1 World lattices for Example 3, (a) representing $p(S)$ and (b) representing $p(T)$.

If we protect $p \supset q$ in both theories, then BERYL and PWA get the same results. But in so doing we have implicitly gone beyond the syntax (and semantics) of the propositional calculus. $p \supset q$ then signifies something closer to $L(p \supset q)$ in a system of modal logic (where L is the necessity operator), since we insist that $p \supset q$ hold in every possible world. (PMA agrees with BERYL, not PWA.)

The final example, also taken from Ginsberg's paper, serves to differentiate BERYL from both PWA and PMA.

Example 4 Let $S = \{\neg v, b, s\}$ with $S^* = \emptyset$, where v denotes that Verdi is French, b denotes that Bizet is French, and s denotes that Satie is French.

What are the counterfactual consequences of $v \equiv b$, i.e. the consequences of Verdi and Bizet being compatriots? PWA, PMA and BERYL agree about this. PWA would derive the two alternative theories $S_1 = \{v, b, s\}$ and $S_2 = \{\neg v, \neg b, s\}$, while, in BERYL's notation:

$$(v \equiv b)(S) = \{\{v, b, s\}, \{s\}\}.$$

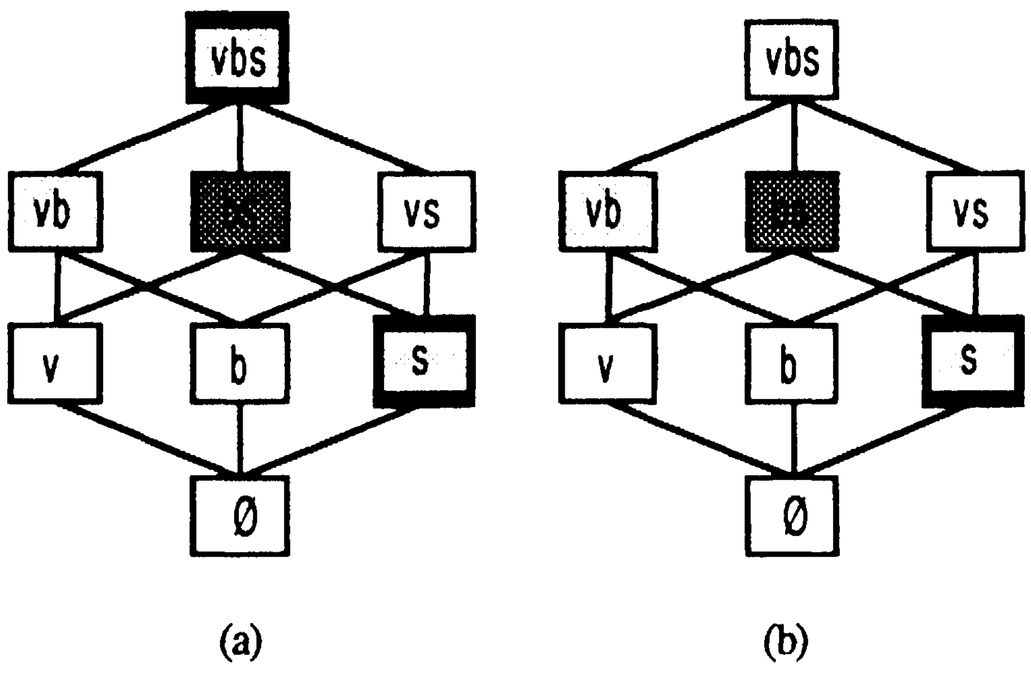
Thus $(v \equiv b) > s$, which seems intuitively right: Satie

remains French in both worlds. This result can be read off from Figure 2a, where dark shading distinguishes models of S, light shading distinguishes models of the counterfactual premise, and models of the revised theory are heavily outlined.

But what of the additional premise $\neg v \equiv s$? What are the counterfactual consequences of Verdi and Satie not being compatriots while Verdi and Bizet are compatriots?

PWA and PMA agree on the set of worlds $\{\{s\}, \{v, b\}\}$. In PWA, we can retract the Frenchness of Bizet to obtain the world $\{s\}$, or we can retract the Frenchness of Satie to obtain the world $\{v, b\}$. In PMA, the models $\{s\}$ and $\{v, b\}$ each satisfy $(\neg v \equiv s) \wedge (v \equiv b)$ and differ minimally from $\{b, s\}$, the only model of S. Neither of these worlds is preferred to the other, according to Definition 2. BERYL's answer is $((\neg v \equiv s) \wedge (v \equiv b))(S)$, which evaluates to

$$\{\{v, b\}, \{s\}\} \Rightarrow \{\{b, s\}\} = \{\{s\}\}.$$



(a) (b)

Figure 2 World lattices for Example 4, (a) representing $(v \equiv b)(S)$ and (b) representing $((\neg v \equiv s) \wedge (v \equiv b))(S)$.

Thus $((\neg v \equiv s) \wedge (v \equiv b)) > s$, and we conserve Satie's Frenchness (see Figure 2b). The minimal change to S is to let Satie stay French in the world where Bizet is Italian, since one of Bizet and Verdi must change nationality anyway. The inclusion of $\{v, b\}$ by PWA and PMA seems counterintuitive. $\{v, b\}$ differs from $\{b, s\}$ on atoms v and s, while $\{s\}$ only differs on b, so $\{s\}$ conserves more of S, and therefore seems more similar to $\{b, s\}$ than does $\{v, b\}$.

In any event, we have a clash of intuitions, and BERYL's result is perfectly arguable. The difference betwen BERYL and PMA is a direct consequence of Definition 4, which ensures that $\psi(S)$ contains worlds in $W_{\{\psi\}}$ which are maximal subsets or minimal supersets of worlds in W_S, rather than worlds which differ minimally on atoms from worlds in W_S, as in PMA.

We can use this example to further distinguish between BERYL and PMA. The most illuminating exercise is to compute $(\neg v \equiv s)((v \equiv b)(S))$, i.e. to compute the worlds in which $\neg v \equiv s$ holds which are most similar to the worlds most similar to S where $v \equiv b$ holds. BERYL's answer is identical to $((\neg v \equiv s) \wedge (v \equiv b))(S)$:

$$\{\{s\}, \{v\}, \{v, b\}, \{b, s\}\} \cap \{\{v, b, s\}, \{s\}\} = \{\{s\}\}$$

whether or not $v \equiv b$ is protected, because $\neg v \equiv s$ is consistent with $(v \equiv b)(S)$, and so the revision is given by $W_{\{\neg v \equiv s\}} \cap (v \equiv b)(S)$, courtesy of Definition 5. Hence $(v \equiv b) > ((\neg v \equiv s) > s)$, and we conserve Satie's Frenchness again.

If $v \equiv b$ is unprotected, PMA's answer is

$$\text{Incorporate}(\{\neg v \equiv s\}, \text{Incorporate}(\{v \equiv b\}, \{b, s\})) = \{\{v, b\}, \{b, s\}, \{s\}\}.$$

If $v \equiv b$ is protected, then PMA's answer is $\{\{v, b\}, \{s\}\}$. Yet the monotonic revision[1] $\{s\}$ to $(v \equiv b)(S)$ seems preferable to either of the *non*monotonic revisions $\{v, b\}$ and $\{b, s\}$, whether $v \equiv b$ is protected or not. It is hard to see how a nonmonotonic revision can be as small as a monotonic revision, whatever one's intuitions. Remember that all three systems agree on $(v \equiv b)(S)$. In fact, PWA agrees with BERYL here.

4 Summary and related work

We have presented a semantics for counterfactual implication which uses nothing more than set inclusion to compute the set of plausible, similar worlds in which a counterfactual premise holds. The function that computes this set, for a given theory S with protected propositions S*, corresponds to the premise itself. Thus if ψ is a counterfactual premise inconsistent with S, then ψ can be understood as the function $(\lambda X)(W_{\{\psi\}} \Rightarrow W_X)$, whose argument is S. The idea of using propositions as functions from one epistemic state to another is entirely due to Gärdenfors (1988, Ch.6). Our use differs only in that such functions can be nonmonotonic.

The most interesting comparison is that between BERYL and Gärdenfors' (op cit, Ch.7) axiomatization of Lewis' (1973) counterfactual logic, VC.

Theorem 2 The construction of Definitions 3-6 satisfies the ten axioms and two derivation rules of Gärdenfors' axiomatization of VC.

Proof Let ψ, ϕ and χ be any propositions, and let $S \subset L$ be any consistent theory.

A1. All truth functional tautologies.
$\vdash \psi \supset \phi$ iff $W_{\{\psi\}} \subseteq W_{\{\phi\}}$ as usual.
A2. $(\psi > \phi) \wedge (\psi > \chi) \supset (\psi > (\phi \wedge \chi))$.
If $\omega \models \phi$ and $\omega \models \chi$ for all $\omega \in \psi(S)$, then $\omega \models (\phi \wedge \chi)$.
A3. $\psi > T$. If $\vdash \phi$, then $\omega \models \phi$ for all $\omega \in \psi(S)$.
A4. $\psi > \psi$. $\omega \models \psi$ for all $\omega \in \psi(S)$.
A5. $(\psi > \phi) \supset (\psi \supset \phi)$.
If $\omega \models \phi$ for all $\omega \in \psi(S)$, then ϕ follows from S and ψ.

[1] The model $\{s\}$ represents a monotonic revision because it only requires that we *discard* models from Ws; thus the process is analogous to theory extension. By contrast, the revisions $\{v, b\}$ and $\{b, s\}$ each require that we introduce *new* models, and are therefore nonmonotonic.

A6. $(\psi \wedge \phi) \supset (\psi > \phi)$.
This follows immediately from Definition 5.
A7. $(\psi > \neg\psi) \supset (\phi > \neg\psi)$.
$\omega \models \neg\psi$ for all $\omega \in \psi(S)$ iff $\psi(S) = \emptyset$. Thus $\omega \models \neg\psi$ for all $\omega \in \phi(S)$ and any ϕ, since $S^* \vdash \neg\psi$. (If $S^* = \emptyset$, then $\vdash \neg\psi$.)
A8. $((\psi > \phi) \wedge (\phi > \psi)) \supset ((\psi > \chi) \supset (\phi > \chi))$.
If $\psi \equiv \phi$, then $W_{\{\psi\}} = W_{\{\phi\}}$, so $\psi(S) = \phi(S)$.
A9. $((\psi > \chi) \wedge (\phi > \chi)) \supset ((\psi \vee \phi) > \chi)$.
$(\psi \vee \phi)(S) \subseteq W_{\{\chi\}}$ if $\psi(S) \subseteq W_{\{\chi\}}$ and $\phi(S) \subseteq W_{\{\chi\}}$.
A10. $((\psi > \phi) \wedge \neg(\psi > \neg\chi)) \supset ((\psi \wedge \chi) > \phi)$.
$(\psi \wedge \chi)(S) \subseteq \psi(S)$ if $\omega \not\models \neg\chi$ for any $\omega \in \psi(S)$.
DR1. $\vdash (\psi \wedge (\psi \supset \phi)) \supset \phi$.
$(W_{\{\psi\}} \cap W_{\{\psi \supset \phi\}}) \subseteq W_{\{\phi\}}$ as usual.
DR2. If $\vdash \phi \supset \chi$, then $\vdash (\psi > \phi) \supset (\psi > \chi)$.
$W_{\{\phi\}} \subseteq W_{\{\chi\}}$ so $\psi(S) \subseteq W_{\{\chi\}}$ if $\psi(S) \subseteq W_{\{\phi\}}$.

Ginsberg's construction satisfies A10 only if the partial order is *modular,* i.e. for any worlds ω and ϖ such that neither $\omega < \varpi$ nor $\varpi < \omega$, if $\upsilon < \omega$ then $\upsilon < \varpi$. Partial orders based solely on set inclusion do not have this property, although orders based on cardinality, for example, do. Nevertheless, BERYL satisfies A10 without modularity, so this suggests that the requirement is a property of PWA, not counterfactual logic.

Gardenfors identifies a number of criteria for the classification of belief revision functions, two of which are the *preservation criterion* (K*P) and the monotonicity *criterion* (K*M). The former states that if ϕ follows from S and Ψ is consistent with S, then ϕ will still follow from the revision of S by ψ. The latter states that if S1 and S_2 are theories and S_2 contains S1, then the revision of S_2 will contain the revision of S1 We can show that the revisions sanctioned by BERYL arc always preservative but not always monotonic. (K*P and K*M are translated into the present notation in the following theorems.)

Theorem 3 BERYL satisfies the preservation criterion, K*P: **If $S \not\models \neg\psi$ and $S \models \phi$, then $\psi > \phi$.**

Proof $S \not\models \neg\psi$ and $S \models \phi$ by hypothesis, so $W_{\{\psi\}} \cap W_S \neq \emptyset$ and $W_S \subseteq W_{\{\phi\}}$. But then $W_{\{\psi\}} \cap W_S \subseteq W_{\{\phi\}}$. So $\omega \models \phi$ for all $\omega \in \psi(S)$, and $\psi > \phi$.

It is easy to show that PWA satisfies K*P.

Theorem 4 PMA docs not satisfy the preservation criterion.

Proof By counterexample. In Example 4, (v = *b)* > *s* with respect to the theory $S = \{\neg v, b, s\}$, but it is not the case that $(\neg v \equiv s) > s$ with respect to the theory $S' = \{v \equiv b, s\}$, which is just S revised by v = *b.* Yet $S' \not\models \neg(\neg v \equiv s)$.

These theorems are important if one wishes to extend either system in the direction of a probabilistic model, since Bayes' Theorem endorses the preservation criterion. Thus the revision functions of BERYL and PWA are amenable to a Bayesian extension, while that of PMA is not as it stands.

Theorem 5 BERYL does not satisfy the monotonicity criterion, K*M: **If $S_1 \subseteq S_2$, then $\psi(S_2) \subseteq \psi(S_1)$.**

Proof By counterexample. Let $S_1 = \{p \equiv q\}$ and $S_2 = \{p \equiv q, p\}$. $\neg q(S_1) = \{\emptyset\}$ while $\neg q(S_2) = \{\{p\}\}$. Hence $S_1 \subseteq S_2$, but not $\psi(S_2) \subseteq \psi(S_1)$

Not surprisingly, BERYL is a nonmonotonic logic. Note that we render the consequent of K*M by $\psi(S_2) \subseteq \psi(S_1)$ **and not** $\psi(S_1) \subseteq \psi(S_2)$, since there is an inverse relation between the specificity of a theory and the number of models that satisfy it.

In conclusion, we feel that nothing in this paper detracts from Ginsberg's argument that counterfactual reasoning is important for artificial intelligence. We criticize his construction because (i) logically equivalent theories can differ in their counterfactual consequences, and (ii) it does not always compute the smallest revision necessary to admit a proposition to a theory. Winslett's PMA avoids the confusion of proof theory and model theory found in PWA. Like BERYL, PMA uses no order on models other than set inclusion to compute similarity, and logically equivalent theories have identical counterfactual consequences. Yet Example 4 shows that the definition of 'Incorporate' (given here as Definition 2) is not equivalent to Definitions 4 and 5, as it docs not always prefer monotonic revisions to nonmonotonic ones.

References

[Chang and Keisler, 1973] C. C. Chang and H. J. Keisler. *Model Theory.* New York: Elsevier North Holland, 1973.

[Gardcnfors, 1988] P. Gardenfors. *Knowledge in Flux.* Boston, MA: MIT Press, 1988.

[Ginsberg, 1986] M. L. Ginsberg. Counterfactuals. *Artificial Intelligence,* 30, 35-79, 1986.

[Keisler, 1977) H. J. Keisler. Fundamentals of model theory. In Barwise, J. (ed.) *Handbook of Mathematical Logic,* New York: Elsevier North-Holland.

[Lewis, 1973] D. K. Lewis. *Countcrfactuals.* Oxford: Blackwell, 1973.

[Winslett, 1988] M. Winslett. Reasoning about action using a possible models approach. In *Proceedings of the 7th National Conference on Artificial Intelligence,* pages 89-93, St Paul, Minnesota, August 1988. American Association for Artificial Intelligence.