Counterfactuals

Matthew L. Ginsberg

LOGIC GROUP
KNOWLEDGE SYSTEMS LABORATORY
Stanford University
Stanford, California 94305

Abstract

Counterfactuals arc a form of commonsense non-monotonic inference that has been of long-term interest to philosophers. In this paper, we begin by describing some of the impact counterfactuals can be exported to have in artificial intelligence, and by reviewing briefly some of the philosophical conclusions which have been drawn about them. Philosophers have noted that the content of any particular counterfactual is in part context-dependent; we present a formal description of counterfactuals that is formally identical to the "possible worlds" interpretation due to David Lewis and which allows us to encode this context-dependent information clearly in the choice of a sublanguage of the logical language in which we are working. Finally, we examine the application of our ideas in the domain of automated diagnosis of hardware faults.

§1. Introduction

A counter factual is a statement such as, "if p, then *q"* where *p* is expected to be false. Typical examples are, "If the electricity hadn't failed, dinner would have been ready on time," or "If the bedroom door were open, I could get the widget I left in there."

From the point of view of logical semantics, counterfactuals arc always true. Thin is in contrast with our intuitive understanding of their content, which might well accept the statements in the last paragraph while rejecting, for example, "If the power hadn't failed, pigs would fly."

Indeed, the distinction between true and false counterfactuals seems to underly much of our use of knowledge. When planning the solution to a complicated problem, we reduce it to subproblems by realizing that we can prove a counterfactual of the form, "If only thus-and-so were true, T would be able to solve the original problem." The original problem reduces to proving the couuterfactual (in some suitable sense) and to arranging for thus-and-so to be true.

Consider the problem of crossing a river if the only boat available has no oars. The countcrfactual, "If I had some oars, I'd be able to cross the river," suggests replacing the original problem with that of finding something with which to row. This is a fairly general phenomenon: counterfactuals suggest goal regressions.

Counterfactuals also enable us to describe why plans fail. The example we have already mentioned, "If the electricity hadn't failed, dinner would have been ready on time," is typical. The electricity *did* fail. But in spite of its lack of logical content, the statement does explain why the soup isn't ready.

Dave Smith has pointed out that additional applications can be found in the area of design. Suppose that a machine contains an on-line representation of the design of a complex device. Questions of the form, "If I were to remove the pullup resistor connected to the output of the OR gate, would the output of the circuit change?" are countcrfactual in nature.

Diagnosis is similar. The couuterfactual, "If the device fails in this fashion, the AND gate is not functioning," allows us to reason directly from the *intended* description of the design of the device in question, as opposed to reasoning from a description that explicitly allows for the failure of some component, as in [2]. We will also see that the nature of countcrfactual implication is also such as to subsume the minimal fault assumption.

Finally, counterfactuals will necessarily play a part in natural language understanding. The extent to which they pervade our communications makes it inevitable that we will eventually need a formal description of them.

There is also a very loose connection between counterfactuals and causality. In the planning examples we have given, the countcrfactual "if p, then *q"* corresponds to "-ip is a cause for *-q"* The electricity failure is the cause of the lateness of the dinner. The lack of oars prevents us from crossing the river.

This connection cannot be pushed too far, however. Suppose that after a MYCIN run [7], we want to know why the machine asked that a certain clinical test be taken. The response is that, "If the result had been positive, the organism would have been rodlike." This is a useful countcrfactual, but the causal connection is from the conclusion to the premise, as opposed to the reverse.

An example where there is no causal connection at all can be found in [5]. Suppose that Olga attends a certain party, but that Boris, who is trying to avoid Olga, does not. If Olga has no similar aversion to Boris, we would have that, "(Even) if Boris had come, Olga would (still)

Research supported by the Office of Naval Rescrach under grant number N00014-81-K-0004.

♦

have come." Here, the counterfactual describes the *lack* of a causal connection.

My aim in this paper is twofold, Firstly, 1 would like to describe briefly some of the existing philosophical work that has been done on counterfactuals, although with an eye toward eventual applications in artificial intelligence. Secondly, 1 will present a formal description of counterfactuals that is precise enough to admit a machine implementation.

§2. Properties of counterfactuals

In his excellent book on counterfactuals, Lewis [5] clarifies the distinction between counterfactuals and standard logical implications by listing some of the properties that distinguish them. The results of this section are not new, but may be unfamiliar to an AI audience. A fairly complete treatment of this topic can also be found in [8].

Contraposition is not necessarily valid for counter/actuals. If we denote the counterfactual, "if *p,* then *q"* by $p > q$, we cannot conclude $\neg q > \neg p$ from $p > q$. Returning to our power failure example, it may well be the case that if the power hadn't failed, dinner would have been on time:

The power didn't fail > dinner was on time.

It does not follow from this that the electricity *would* have failed if dinner had been late—there may well be an alternative possible cause, such as tardiness on the part of the chef.

Counterfactuals arc not necessarily transitive. From $p > q$ and $q > r$ we cannot necessarily conclude $p > r$. The standard example is due to Stalnaker [9j:

If J. Edgar Hoover had been bom a Russian, then he would have *been a Communist,* and

If he had been a *Communist, he would have been* a *traitor*

do not together imply

//' Hoover had been horn a Russian, he would have been a traitor,

Finally, counterfactuals arc non-mono tonic. Given $p > r$, we cannot necessarily conclude $p \wedge q > r$. In fact, it is possible to have $p > r$ and $p \wedge q > \neg r$: The two statements, "If the electricity hadn't failed, dinner would have been ready on time," and "If the electricity hadn't failed, but 1 had been elected president, dinner would have been late," are completely consistent.

Glymour and Thomason [4] seem to infer from this last observation that the study of non-monotonic inference generally can be subsumed to some extent under an investigation of counterfactuals, but in light of the breadth of the non-monotonic nature of commonsense reasoning (the frame problem, default rides, etc.), this seems to me to miss the mark. Counterfactuals in fact seem to be a distinct type of non-monotonic reasoning.

§3. Possible worlds

Following an idea of Stainaker's [9], most modern investigations of counterfactuals arc based on the notion of possible worlds. Loosely speaking, we analyze a counterfactual $p > q$ by considering the "possible world" that is as similar to our (real) world as possible, given that *p* is true in it. The counterfactual is true or not depending upon whether or not *q* holds in this world.

Lewis [5] has observed that this "most similar possible world" may not be unique. He defines a counterfactual to be true if and only if it holds uniformly in the most similar possible worlds in which the premise holds.

This distinction is apparent if we consider the law of the counterfactual excluded middle:

$$(p > q) \vee (p > \neg q).$$

In Stalnaker's view, either *q* or - *q* will hold in the nearest possible world where *p* holds, so that the disjunction above will be valid. Lewis points out that this may not be the case by examining the counterfactuals, "If Bizet and Verdi had been compatriots, Bizet would have been Italian," and, "If Bizet and Verdi had been compatriots, Bizet would not have been Italian." Both of these appear to be invalid.

It is possible to understand the discussion of section 2 in terms of possible worlds; we will discuss only the non-monotonic nature of counterfactuals here. The other two properties described earlier arc similar.

The basic reason that $p > r$ and $p \wedge q > \neg r$ are consistent is that worlds where $p \wedge q$ hold may be much less similar to our own than worlds where *p* alone holds. It is entirely possible that r holds uniformly in the nearest of the *p*-worlds while $\neg r$ holds uniformly in the (more distant) nearest of the *p* A (/-worlds. It is also possible that r holds in some of the nearest of the $p \wedge q$-worlds while $\neg r$ holds in others. In this case we would have only $p > r$ and $\neg(p \wedge q > r)$.

Returning to our power failure example, the nearest possible worlds in which the power remained on are worlds in which dinner was ready on time. In the nearest of the (much) more distant worlds where 1 was elected president, dinner was late. In still more distant worlds, such as those where the power remained on and I was elected president, but no one bothered to inform me, dinner will once again be prompt.

§4. Framework

From an AI perspective, the difficulty with the possible worlds interpretation of counterfactuals is that the notion of "similarity" is too vaguely defined. Our main

intention in this paper is to present a sharper definition of it.

Intuitively, there are at least two measures of the similarity, or lack thereof, between different possible worlds. These correspond loosely to the number of propositions whose truth values change and to the relative importance of the propositions involved.

The latter is the most difficult to understand in any formal sense, since (as we will see in the next section) it is fundamentally dependent upon context. In fact, we will be able to do little more than to provide a way to encode information concerning the relative importance of the propositions being considered. It is of some interest to note that the scheme we will present can be used to define a notion of similarity that is unchanged from one possible world to another.

The other source of similarity is more syntactic. If the truth value of a proposition changes unnecessarily, in the sense that the possible world without the change is consistent, the possible world which incorporates the change should be deemed "more distant" from our own than the one which doesn't.

In order to understand this, we will work in a logical system that allows us to assign to a proposition the truth value "unknown" (u) in addition to the more usual ones of "true" (t) or "false" (/). We therefore let $T = \{t, f, u\}$ be the set of allowable truth values, take L to be the set of sentences in our language, and define a *truth function* to be a mapping

$$\phi : L \rightarrow T.$$

Intuitively, $\phi(p) = u$ if we are uncertain as to the truth or falsity of p.

We will say that a truth function > is an *extension* of a truth function ψ, writing $\phi \leq \psi$, if, for all $p \in L$, $\phi(p) = \psi(p)$ or $\psi(p) = u$. We will call <f> a *simple extension* if $\phi(p) \neq \psi(p)$ for only a single $p \in L$.

We will also assume that we have some way of determining whether or not a truth function is consistent. In the predicate calculus case upon which we are modelling our analysis, however, the truth value of a compound sentence is defined recursively in terms of the truth values of its components; furthermore, the definition only applies if the truth values of these components are *t* or / (as opposed to it). The (consistent) assignment of the truth values *t* or / to all of the sentences in *L* constitutes an *interpretation* forL.

This leads us to define a truth function ϕ to be an *interpretation* if $\phi(p) \neq u$ for all $p \in L$. If a truth function ψ is an interpretation that is an extension of the truth function ϕ, we will say that ψ is a *complete* extension of ϕ. Assuming that consistency is defined as a primitive for interpretations, we will say that a truth function ϕ is *consistent* iff ϕ has a consistent complete extension.-

Here are some examples. In the first two cases, is a consistent complete extension of ϕ.. Since ϕ in the third case has no consistent complete extension, it is inconsistent.

x	$\phi(x)$	$\psi(x)$	$\phi(x)$	$\psi(x)$	$\phi(x)$	$\psi(x)$
A	t	t	f	f	f	f
B	f	f	u	t (or f)	u	?
$A \wedge B$	f	f	u	f	t	t

Lemma 1. *No extension of an inconsistent truth function is consistent.*

Proof. The consistent complete extension of such an extension would be a consistent complete extension of the original (inconsistent) truth function. a

Equivalently, any truth function with a consistent extension is consistent.

For a consistent truth function ϕ, the *closure* of ϕ will be the maximally extended truth function ψ such that every consistent complete extension of ϕ is an extension of ψ. The closure of ϕ will be denoted $\mathrm{cl}(\phi)$. If $\phi = \mathrm{cl}(\phi)$, ϕ will be called *closed*; it is not hard to see that this corresponds precisely to logical closure.

In the example below, ϕ_1 and ϕ_2 are the only consistent complete extensions of ϕ:

	ϕ	ϕ_1	ϕ_2	$\mathrm{cl}(\phi)$
A	t	t	t	t
B	u	t	f	u
$A \vee B$	u	t	t	t

Lemma 2. $\mathrm{cl}(\phi) \leq \phi$.

Proof. Let $\{\phi_i\}$ be the consistent complete extensions of ϕ. T h e$\phi_i \leq \phi$ r all i, so that $\mathrm{cl}(\phi) \leq \phi$. □

Related to the notion of closure is that of reduction: a truth function ϕ will be called *reduced* if all of its simple extensions arc consistent. The idea is that a simple extension of ϕ corresponds to the acquisition of more knowledge about some specific proposition; if every such extension is consistent, the original truth function must have been minimal in the sense that it had no extensions which were "necessary" consequences of it.

Lemma 3. *A consistent truth function ϕ is reduced if and only if it closed.*

Proof. For any $p \in L$, if $\phi(p) = u$ and ϕ is reduced, the truth function obtained by replacing $\phi(p)$ with *t* (respectively /) is consistent and therefore has a consistent

complete extension which we will denote $\phi_{p:t}$ (respectively $\phi_{p:f}$). Since $\phi_{p:t}(p) = t$ and $\phi_{p:f}(p) = f$, $\mathrm{cl}(\phi)(p) = u$; it follows that ϕ is closed if it is reduced.

The reverse implication is similar. □

§5. Similarity

The terminology introduced in the last section allows us to make precise some of the ideas in section 3: possible worlds correspond to consistent interpretations, and sets of possible worlds to consistent truth functions.

The difficulty with this is that we still lack a formal notion of similarity. Part of the problem is context-dependent, as we can see from the pair of counterfactuals

If Caesar had been in command [in Korea], he would have used the atom bomb

and

If Caesar had been in command, he would have used catapults.

This example is Quinc's [6]. Either counterfactual may well be true (although not both); if the first, Caesar's character is important to our notion of similarity; if the second, it is the tools he had available which arc relevant.

It is clearly impossible to select between these two counterfactuals in advance; the best we can do is to present a method for encoding in our semantic machinery the information leading to the choice. In order to do this, let V be a subset of L, and suppose that ϕ is a fixed truth function defined on L — //. We can now define a truth function ψ on V to be consistent if and only if the truth function

$$\psi'(p) = \begin{cases} \psi(p), & \text{for } p \in L', \\ \phi(p), & \text{for } p \notin L' \end{cases}$$

is consistent. The effect of this is to fix the truth values outside of L', so that any consistent truth function on V must be consistent with them.

In the above example, if V includes "Caesar was a ruthless military leader," but not, "Caesar's military tools were those of the Roman Empire at its height," we will have to accept Caesar's use of catapults without question, regardless of the weapons available at the time of the engagement, resulting in the validity of the counterfactual which concludes that he would have used catapults. If the situation were reversed, the other counterfactual would be valid. If V includes both of the statements describing Caesar, the construction we will present will ambiguously select either of the counterfactuals, while if both of the descriptions arc relegated to L, both counterfactuals will be vacuously true because no truth function > with ϕ(Caesar in command) = t will have a consistent extension to all of L.

Given the choice of a (possibly restricted) language //, let p be a sentence in V and ϕ a closed truth function with $\phi(p) = f$. We will investigate the consequences of the counterfactual premise p be defining a new truth function, to be denoted $\phi|_p$, and corresponding to ϕ *with the truth value at p replaced by t.*

Throughout this section, we will follow the Caesarian example carefully. We have the following propositions in L':

K = *Caesar in command in Korea*
r = *Caesar was ruthless*
R = *Caesar's tools were those of the Romans*
a = *the atom bomb was used in Korea*
c = *catapults were used in Korea.*

In L are the axioms:

$$K \wedge r \rightarrow a$$
$$K \wedge R \rightarrow c$$
$$\neg(a \wedge c).$$

The last of these is equivalent to $\neg(K \wedge r \wedge R)$. Our initial truth assignment is given by:

	ϕ
K	f
r	t
R	t
a	f
c	f

Of interest to us is the truth function $\phi|_K$; what if Caesar had been in command in Korea?

The general difficulty is that simply changing $\phi(p)$ to t may well produce an inconsistent truth function. We begin therefore by replacing $\phi(p)$ not with t, but with u. This truth function ϕ' must be consistent (since ϕ is an extension of it), but need not be closed.

	ϕ	ϕ'	$\mathrm{cl}(\phi')$
K	f	u	f
r	t	t	t
R	t	t	t
a	f	f	f
c	f	f	f

Assume now that there is at least one reduced truth function ψ of which ϕ' is an extension. (We can genearlly take for ψ the closure of the truth function which assigns u to every $p \in L$.) The set of all such ψ's is partially ordered under extension; let ϕ'' be a minimal element of it.

	ϕ	ϕ'	ψ (typical)	ϕ''_1	ϕ''_2
K	f	u	u	u	u
r	t	t	t	t	u
R	t	t	u	u	t
a	f	f	u	u	f
c	f	f	u	f	u

The result of replacing $\phi''(p)$ with t is necessarily consistent, since ϕ'' is reduced and $\phi''(p) = u$ initially. Let θ be the truth function obtained by making this replacement, and take $\phi|_p$ to be $\mathrm{cl}(\theta)$. We also define $\phi(p > q) = \phi|_p(q)$.

	ϕ''_1	θ_1	$\phi\|_K$	ϕ''_2	θ_2	$\phi\|_K$
K	u	t	t	u	t	t
r	t	t	t	u	u	f
R	u	u	f	t	t	t
a	u	u	t	f	f	f
c	f	f	f	u	u	f
$K > a$			t			f
$K > c$			f			t

The construction we have given works through the construction of the "intermediate" truth function ϕ''. This is in some sense a set of the most similar possible worlds in which we might have $\phi(p)$ being either true or false. The reason we have done this is because it seems to be technically easier to define similarity for a world of which ours is an extension (ϕ' in the above construction) than for one which is not directly comparable with it (as ϕ with $\phi(p)$ replaced by t would be).

Having constructed similar worlds where p *might* hold, it is straightforward to investigate the consequences if we assume that it *does* hold. This is the content of our taking $\phi|_p$ to be the closure of ϕ'' with $\phi''(p)$ replaced by t.

The choice of truth function ϕ'' corresponds to choosing a *subset* of the possible worlds in which $\phi(p) = t$. In our military example, assuming Caesar to be in command in Korea required our abandoning the truth of either r or R, and the choice of which to discard corresponded to our choice of ϕ''.

An alternative would be to assume that the worlds where we abandon one description are just as similar to our own as those in which we abandon another. Lewis might well make this choice; it corresponds to defining $\phi(p > q)$ to be t or f if it takes this value independent of the choice of ϕ'', and to be u otherwise.

Theorem 4. *With the above definition, our construction is formally identical to Lewis' possible world interpretation.*

Proof. See [3]. a

§6. Example: diagnosis

6.1 Setting

Genesercth has proposed [2] that it is possible for machines to be used in automated diagnosis, provided that the machines are given both a design for the device in question, and the ability to manipulate the device by varying its inputs and observing the results. He investigates the diagnosis of a full adder (top of next page) in considerable detail.

Genesereth describes the design of the full adder in a variant of prefix predicate calculus. Quoting him:

```
SD1:  (XORG X1)
SD2:  (XORG X2)
SD3:  (AVNDG A1)
SD4:  (ANDG A2)
SD5:  (ORG 01)
SD6:  (CONN (IN 1 F1) (IN 1 X1))
SD7:  (CONN (IN 1 FI) (IN 1 AI))
SD8:  (CONN (IN 2 FI) (IN 2 X1))
SD9:  (CONN (IN 2 FI) (IN 2 A1))
SD10: (CONN (IN 3 FI) (IN 2 X2))
SD11: (CONN (IN 3 FI) (IN 1 A2))
SD12: (CONN (OUT 1 X1) (IN 1 X2))
SD13: (CONN (OUT 1 X1) (IN 2 A2))
SD14: (CONN (OUT 1 AI) (IN 2 01))
SD15: (CONN (OUT 1 A2) (IN 1 01))
SD16: (CONN (OUT 1 X2) (OUT 1 FI))
SD17: (CONN (OUT 1 01) (OUT 2 FI))
```

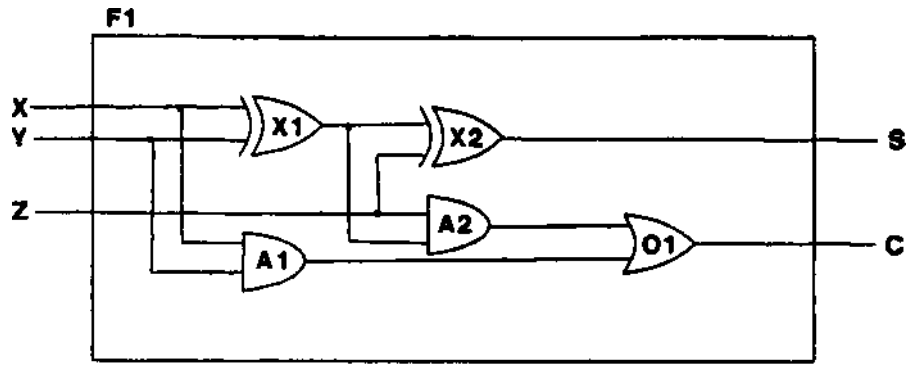
These axioms describe the *structural description* of the full adder. SD1, for example, states that X1 is an exclusive-or gate. SD 13 states that the first output of X1 is connected to the second input of A2.

Cenesercth also states in a similar form results describing the behavior of the various sorts of gates, and describing what it means for two points in the circuit to be connected. Using these additional axioms, it is possible to prove that if, for example, the first input to the full adder is on while the other two arc off, the first output should be on while the second should be off.

6.2 Diagnosis using predicate calculus

The situation of interest, of course, is that in which the outputs of the adder arc not as predicted by the design. Gcnesercth assumes that we have:

```
AC1:  (VAL (IN 1 FI) ON)
AC2:  (VAL (IN 2 FI) OFF)
AC3:  (VAL (IN 3 FI) OFF)
OBI:  (VAL (OUT 1 FI) OFF)
0B2:  (VAL (OUT 2 FI) OFF)
```



A full adder is essentially a one bit adder with carry in and carry out, and it is usually used as one of n elements in an n bit adder. A graphical representation of its design is given [above]. It has three inputs and two outputs and consists *of* two "xor" gates (X1 and X2), two "and" gates (AI and A2), and an "or" gate (OI) ... In normal operation, the first output (the "sum" line) is "on" if and only if an odd number of inputs is "on"; the second output (the "carry" line) is "on" if and only if at least two inputs are "on". [4]

In other words, the values of the three inputs to the adder are as described at the end of the last subsection, but both of the outputs are off.

The observed behavior characterized by OB1 OB2 is inconsistent with the design given by SD1 SD17 and the inputs AC1-AC3. Diagnosis is a matter of resolving this inconsistency.

To do so using predicate calculus, Gcnesereth assumes (correctly!) that the device in question does not satisfy the design description given earlier, but instead satisfies some weaker "device assumptions". In the example we are considering, he assumes that:

(1) The connections are all as described in the design, and
(2) At most one of the gates is broken (the *single fault assumption).*

These device assumptions can be encoded by replacing the structural description axioms SD1-SD5 with axioms such as:

```
(IF (NOT (XORG X1)) (AND (XORG X2) (ANDG AI)
                         (ANDG A2) (ORG 01)))
```

These new axioms are consistent with the observed behavior of the adder, and lead to the conclusions that

```
(OR (NOT (XORG X1)) (NOT (XORG X2)))        (10)
```

and

```
(AND (ANDG AI) (ANDG A2) (ORG OI).          (11I)
```

In other words, one of the exclusive-or gates is broken, and the remaining components arc functioning satisfactorily. The information about the functionality of the AND and OR gates is useful because it enables us to generate a test to determine which of the two exclusivc-or gates is in fact faulty.

The difficulty with this approach is that it requires us to generate device assumptions to replace the structural description SD1-SD17. It is possible that the fatilt(s) in the device are such that these assumptions are invalid, in which case the system will be unable to diagnose the device without a replacement set of device assumptions.

6.3 Diagnosis using counterfactuals

The device assumptions in the last subsection were introduced to encode our belief that the wiring in the adder was correct, and to enable us to take advantage of the simplifying assumption that only one of the remaining components was damaged. Both of these conditions can be recast naturally in the framework of counterfactuals.

To perform the diagnosis using the methods we have presented, we will assume the structural description DS1-DS17 and the inputs achieved by ACI AC3, and examine the count erf actual consequences of the observed outputs OB1-OB2. There are three possible explanations for the fault:

(J) One *or* more of the components is faulty.
(2) The wiring is faulty.
(3) The inputs were not as expected.

We eliminate all but the first from our countcrfactual reasoning by including only the component assumptions SD1-SD5 in our restricted language *V.* The wiring and input

assumptions SD6 -SD17 and AC1-AC3 are therefore assumed to be independent of the counterfactual assumptions corresponding to the observed misbehavior of the device.

The conclusion (10) that one of the two exclusive-or gates must be faulty is in fact a *logical* consequence of the behavior of the device:

$$OB1 \wedge OB2 \rightarrow (OR\ (NOT\ (XORG\ X1))\ (NOT\ (XORG\ X2))).$$

Meanwhile, because the remaining components need not be contributing to the observed fault, their continued performance is *counterfactually* implied by the observed behavior:

$$OB1 \wedge OB2 > (AND\ (ANDG\ A1)\ (ANDG\ A2)\ (ORG\ O1)).$$

This reappearance of (11) is especially useful. Because of the non-monotonic nature of countcrfactual reasoning, it is of course possible that additional observations appended to the lefthand side of the above equation will invalidate its conclusion; this will happen whenever the single fault assumption is violated. In this case, however, rather than generating a contradiction, the counterfactual analysis will automatically produce a new diagnosis which once again involves failure for a minimal set of components.

It is possible, however, for a counterfactual analysis to suggest a violation of the single fault assumption when one is not required. If the observed behavior can be explained either by the failure of a single component, or by the failure of a pair of different components, both will be proposed. There is nothing counterintuitive about this, however—it is quite normal to assume that a group of normally undependable components has failed before questioning a single part of proven reliability. In any event, we can if necessary retain the single fault assumption by using it to select among the possible ϕ''*s in the counterfactual construction itself.

§7. Conclusion

Our aim in this paper has been to present a formal description of counterfactuals, describing them in terms of existing logical operators instead of following the usual practice of developing a "countcrfactual calculus" to describe their behavior.

The construction we have presented seems to meet this objective. It has indeed described coun]terfactuals in terms of existing logical primitives, and reduces to the "possible worlds" interpretation of counterfactuals that is accepted by philosophers.

Our construction also distinguishes clearly between the context dependent and context independent features of countcrfactual implication. It provides us with a precise method for selecting those aspects of our world which are to be considered inviolable even under a counterfactual assumption; having made such a choice, we proceed to generate possible worlds which respect it.

The biggest difficulty with the approach we have described is the rather heavy-handed nature of the choice described in the last paragraph. Although it is possible to clearly recognize ambiguities remaining in the analysis of any particular counterfactual (they correspond to the choice of " in our construction), we have no method for choosing consistently between them. In any specific implementation, it will of course be possible to select a ø" when one is needed, but we have not considered the nature of the formalism that should govern this choice.

Acknowledgement

A preliminary version of this research was presented to MUGS at Stanford, and the author would like to thank the participants in that seminar for their useful comments on this work. Mike Genesereth and Dave Smith were especially helpful.

References

[1] Adams, E., The logic of conditionals, *Inquiry* 8 (1965), 166-197.

[2] Genescreth, M.R., The use of design descriptions in automated diagnosis, *Artificial Intelligence* 24 (1984), 411-436.

[3] Ginsberg, M.L., Counterfactuals, Stanford Computer Science Report STAN-CS-84-1029, Stanford University (1984)

[4] Glymour, C., and Thomason, R.H., Default reasoning and the logic of theory perturbation, *Non-monotonic Reasoning Workshop,* American Association for Artificial Intelligence (1984) 93-102.

[5] Lewis, D., *Counterfactuals,* Harvard University Press, Cambridge (1973).

[6] Quine, W.V., *Methods of Logic,* Holt, Rinchart, and Winston, New York (1950).

[7] Shortliffe, E.IL, *Computer-based Medical Consultations: MYCIN,* Elsevier, New York (1976).

[8] Sobel, J.II., Utilitarianisms: simple and general, *Inquiry* 13 (1970) 394-449.

[9] Stalnaker, It., A theory of conditionals, in *Studies in Logical Theory, R*eschcr, N. (cd.), Oxford University Press, Oxford (1968).