1. An Analysis of Counterfactuals

1.1 Introduction

'*If kangaroos had no tails, they would topple over*' seems to me to mean something like this: in any possible state of affairs in which kangaroos have no tails, and which resembles our actual state of affairs as much as kangaroos having no tails permits it to, the kangaroos topple over. I shall give a general analysis of counterfactual conditionals along these lines.

My methods are those of much recent work in possible-world semantics for intensional logic.* I shall introduce a pair of counterfactual conditional operators intended to correspond to the various counterfactual conditional constructions of ordinary language; and I shall interpret these operators by saying how the truth value at a given possible world of a counterfactual conditional is to depend on the truth values at various possible worlds of its antecedent and consequent.

Counterfactuals are notoriously vague. That does not mean that we cannot give a clear account of their truth conditions. It does mean that such an account must either be stated in vague terms—which does *not* mean ill-understood terms—or be made relative to some parameter that is fixed only within rough limits on any given occasion of language use. It is to be hoped that this imperfectly fixed parameter is a familiar one that we would be stuck with whether or not we used it in the analysis of counterfactuals; and so it will be. It will be a relation of comparative similarity.

Let us employ a language containing these two counterfactual conditional operators:

$$\Box\!\!\rightarrow$$

* See, for instance, Saul Kripke, 'Semantical Considerations on Modal Logic', *Acta Philosophica Fennica* **16** (1963): 83–94; Richard Montague, 'Pragmatics', in R. Klibansky, *Contemporary Philosophy* (La Nuova Italie Editrice: Firenze, 1968): 102–122, reprinted in Montague, *Formal Philosophy* (Yale University Press: New Haven, 1974); Dana Scott, 'Advice on Modal Logic', in K. Lambert, *Philosophical Problems in Logic* (D. Reidel: Dordrecht, 1970); and David Lewis, 'General Semantics', *Synthese* **22** (1970): 18–67.

read as '*If it were the case that* ____, *then it would be the case that . . .*', and

$$\Diamond\!\!\rightarrow$$

read as '*If it were the case that* ____, *then it might be the case that . . .*'. For instance, the two sentences below would be symbolized as shown.

If Otto behaved himself, he would be ignored.
Otto behaves himself $\Box\!\!\rightarrow$ *Otto is ignored*

If Otto were ignored, he might behave himself.
Otto is ignored $\Diamond\!\!\rightarrow$ *Otto behaves himself*

There is to be no prohibition against embedding counterfactual conditionals within other counterfactual conditionals. A sentence of such a form as this.

$$((\psi \Box\!\!\rightarrow ((\chi \Box\!\!\rightarrow \psi) \Diamond\!\!\rightarrow \phi)) \Diamond\!\!\rightarrow \chi) \Box\!\!\rightarrow (\phi \Box\!\!\rightarrow (\psi \Diamond\!\!\rightarrow ((\chi \Box\!\!\rightarrow \phi) \Diamond\!\!\rightarrow (\phi \Box\!\!\rightarrow \psi))))$$

will be perfectly well formed and will be assigned truth conditions, although doubtless it would be such a confusing sentence that we never would have occasion to utter it.

The two counterfactual operators are to be interdefinable as follows.

$$\phi \Diamond\!\!\rightarrow \psi =^{df} \sim(\phi \Box\!\!\rightarrow \sim\psi),$$
$$\phi \Box\!\!\rightarrow \psi =^{df} \sim(\phi \Diamond\!\!\rightarrow \sim\psi).$$

Thus we can take either one as primitive. Its interpretation determines the interpretation of the other. I shall take the 'would' counterfactual $\Box\!\!\rightarrow$ as primitive.

Other operators can be introduced into our language by definition in terms of the counterfactual operators, and it will prove useful to do so. Certain modal operators will be thus introduced in Sections 1.5 and 1.7; modified versions of the counterfactual in Section 1.6; and 'comparative possibility' operators in Section 2.5.

My official English readings of my counterfactual operators must be taken with a good deal of caution. First, I do not intend that they should interfere, as the counterfactual constructions of English sometimes do, with the tenses of the antecedent and consequent. My official reading of the sentence

We were finished packing Monday night $\Box\!\!\rightarrow$ *we departed Tuesday morning*

comes out as a sentence obscure in meaning and of doubtful grammaticality:

If it were the case that we were finished packing Monday night, then it would be the case that we departed Tuesday morning.

In the correct reading, the subjunctive 'were' of the counterfactual construction and the temporal 'were' of the antecedent are transformationally combined into a past subjunctive:

If we had been finished packing Monday night, then we would have departed Tuesday morning.

Second, the 'If it were the case that ____' of my official reading of $\Box\!\!\rightarrow$ is not meant to imply that it is not the case that ____. Counterfactuals with true antecedents—counterfactuals that are not counterfactual—are not automatically false, nor do they lack truth value. This stipulation does not seem to me at all artificial. Granted, the counterfactual constructions of English do carry some sort of presupposition that the antecedent is false. It is some sort of mistake to use them unless the speaker does take the antecedent to be false, and some sort of mishap to use them when the speaker wrongly takes the antecedent to be false. But there is no reason to suppose that every sort of presupposition failure must produce automatic falsity or a truth-value gap. Some or all sorts of presupposition, and in particular the presupposition that the antecedent of a counterfactual is false, may be mere matters of conversational implicature, without any effect on truth conditions. Though it is difficult to find out the truth conditions of counterfactuals with true antecedents, since they would be asserted only by mistake, we will see later (in Section 1.7) how this may be done.

You may justly complain, therefore, that my title 'Counterfactuals' is too narrow for my subject. I agree, but I know no better. I cannot claim to be giving a theory of conditionals in general. As Ernest Adams has observed,* the first conditional below is probably true, but the second may very well be false. (Change the example if you are not a Warrenite.)

If Oswald did not kill Kennedy, then someone else did.
If Oswald had not killed Kennedy, then someone else would have.

Therefore there really are two different sorts of conditional; not a single conditional that can appear as indicative or as counterfactual depending on the speaker's opinion about the truth of the antecedent.

* 'Subjunctive and Indicative Conditionals', *Foundations of Language* **6** (1970): 89–94.

The title 'Subjunctive Conditionals' would not have delineated my subject properly. For one thing, there are shortened counterfactual conditionals like '*No Hitler, no A-bomb*' that have no subjunctives except in their—still all-too-hypothetical—deep structure. More important, there are subjunctive conditionals pertaining to the future, like '*If our ground troops entered Laos next year, there would be trouble*' that appear to have the truth conditions of indicative conditionals, rather than of the counterfactual conditionals I shall be considering.*

1.2 Strict Conditionals

We shall see that the counterfactual cannot be any strict conditional. Since it turns out to be something not too different, however, let us set the stage by reviewing the interpretation of strict conditionals in the usual possible-world semantics for modality. Generally speaking, a strict conditional is a material conditional preceded by some sort of necessity operator:

$$\Box(\phi \supset \psi).$$

With every necessity operator $\Box$ there is paired its dual possibility operator $\Diamond$. The two are interdefinable:

$$\Diamond\phi =^{df} \sim\Box \sim \phi, \quad \text{or} \quad \Box\phi =^{df} \sim\Diamond \sim \phi.$$

If we like, we can rewrite the strict conditional using the possibility operator:

$$\sim\Diamond(\phi \ \& \ \sim\psi).$$

Or we could introduce a primitive strict conditional arrow or hook, and define the necessity and possibility operators from that.‡

A necessity operator, in general, is an operator that acts like a restricted universal quantifier over possible worlds. Necessity of a certain sort is truth at all possible worlds that satisfy a certain restriction. We

* Notation: sentences of our language are mentioned by means of lower-case Greek letters ϕ, ψ, χ *et al.*; sets of sentences by means of Greek capitals. Logical symbols and the like are used autonymously, and juxtaposition of names of expressions signifies concatenation of the expressions named. Possible worlds are mentioned by means of the lower-case letters *h*, *i*, *j*, *k*; sets of worlds by means of capital letters; and sets of sets of worlds by means of script capitals.

‡ In this section only, I use the unmarked box and diamond to stand for *any* arbitrary paired necessity operator and possibility operator. When next they appear, in Section 1.5, they will be reserved thenceforth for a specific use: they will be the 'outer' necessity and possibility operators definable in a certain way from the counterfactual (or they will be analogously related to operators analogous to the counterfactual). The dotted box and diamond, $\boxdot$ and $\diamondsuit$, will be likewise reserved when they appear in Section 1.7.

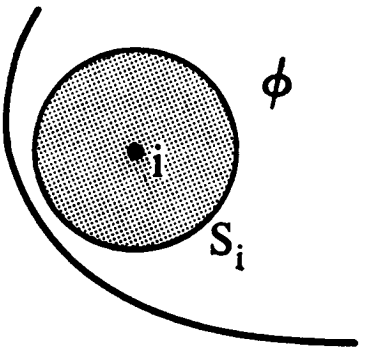
call these worlds *accessible*, meaning thereby simply that they satisfy the restriction associated with the sort of necessity under consideration. Necessity is truth at all accessible worlds, and different sorts of necessity correspond to different accessibility restrictions. A *possibility operator*, likewise, is an operator that acts like a restricted existential quantifier over worlds. Possibility is truth at some accessible world, and the accessibility restriction imposed depends on the sort of possibility under consideration. If a necessity operator and a possibility operator correspond to the same accessibility restriction on the worlds quantified over, then they will be a dual, interdefinable pair.

In the case of *physical necessity*, for instance, we have this restriction: the accessible worlds are those where the actual laws of nature hold true. Physical necessity is truth at all worlds where those laws hold true; physical possibility is truth at some worlds where those laws hold true.

In the case of physical necessity, which possible worlds are admitted as accessible depends on what the actual laws of nature happen to be. The restriction will be different from the standpoint of worlds with different laws of nature. Let i and j be worlds with different laws of nature, and let k be a world where the laws of i hold true but the different laws of j are violated. From the standpoint of i, k is an accessible world; from the standpoint of j it is not. Accessibility is in this case—and most cases—a relative matter. It is the custom, therefore, to think of accessibility as a relation between worlds: we say that k is *accessible from* i, but k is not accessible from j. We say also that i stands to k, but j does not stand to k, in the *accessibility relation* for physical necessity and possibility.

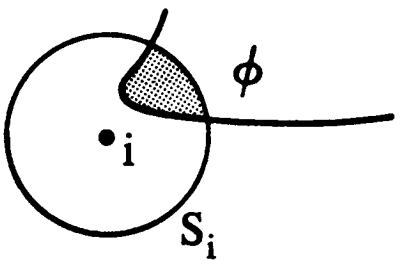
In general: to a necessity operator $\Box$ or a possibility operator $\Diamond$ there corresponds an accessibility relation. The appropriate accessibility relation serves to restrict quantification over worlds in giving the truth conditions for $\Box$ or $\Diamond$. For any possible world i and sentence ϕ, the sentence $\Box\phi$ is true at the world i if and only if, for every world j such that j is accessible from i, ϕ is true at j. Likewise $\Diamond\phi$ is true at i if and only if, for some world j such that j is accessible from i, ϕ is true at j. More concisely: $\Box\phi$ is true at i if and only if ϕ is true at every world accessible from i; $\Diamond\phi$ is true at i if and only if ϕ is true at some world accessible from i. It follows that the strict conditional $\Box(\phi \supset \psi)$ is true at i if and only if, for every world j such that j is accessible from i, the material conditional $\phi \supset \psi$ is true at j; that is, if and only if, for every world j such that j is accessible from i and ϕ is true at j, ψ is true at j. More concisely: $\Box(\phi \supset \psi)$ is true at i if and only if ψ is true at every accessible ϕ-world. ('ϕ-world', of course, abbreviates 'world at which ϕ is true', and likewise for parallel formations.)

(A) NECESSITY



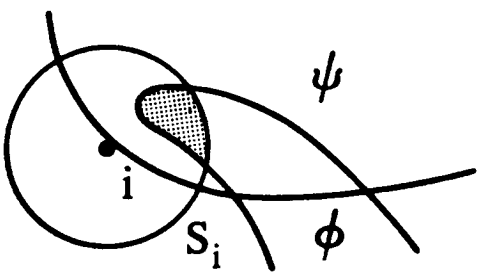
$\Box\phi$

(B) POSSIBILITY



$\Diamond\phi$

(C) STRICT CONDITIONAL



$\Box(\phi \supset \psi)$

FIGURE 1

It suits my purposes better not to use the customary accessibility relations, but instead to adopt a slightly different—but obviously equivalent—formulation. Corresponding to a necessity operator $\Box$, or a possibility operator $\Diamond$, or a kind of strict conditional, let us have an assignment to each world i of a set S_i of worlds, called the *sphere of accessibility* around i and regarded as the set of worlds accessible from i.* The assignment of spheres to worlds may be called the *accessibility assignment* corresponding to the modal operator. It is used to give the truth conditions for modal sentences as follows.

A sentence $\Box\phi$ is true at a world i if and only if ϕ is true throughout the sphere of accessibility S_i around i (as shown in Figure 1(A)).

A sentence $\Diamond\phi$ is true at a world i if and only if ϕ is true somewhere in the sphere S_i (as shown in Figure 1(B)).

A strict conditional sentence $\Box(\phi \supset \psi)$ is true at i if and only if $\phi \supset \psi$ is true throughout the sphere S_i; that is, if and only if ψ is true at every ϕ-world in S_i (as shown in Figure 1(C)).

Let us consider various examples of accessibility assignments for various sorts of necessity, with particular attention to the corresponding strict conditionals.

Corresponding to *logical necessity*, and the logical strict conditional, we assign to each world i as its sphere of accessibility S_i the set of *all* possible worlds. Thus the logical strict conditional $\Box(\phi \supset \psi)$ is true at i if and only if ψ is true at all ϕ-worlds whatever; there are no inaccessible ϕ-worlds to be left out of consideration.

Corresponding to *physical necessity*, and the physical strict conditional, we assign to each world i as its sphere of accessibility S_i the set of all worlds where the laws of nature prevailing at i hold; so the physical strict conditional $\Box(\phi \supset \psi)$ is true at i if and only if ψ is true at all those ϕ-worlds where the laws prevailing at i hold.

Corresponding to a kind of time-dependent necessity we may call *inevitability at time t*, and its strict conditional, we assign to each world i as its sphere of accessibility the set of all worlds that are exactly like i at all times up to time t, so $\Box(\phi \supset \psi)$ is true at i if and only if ψ is true at all ϕ-worlds that are exactly like i up to t.

Corresponding to what we might call *necessity in respect of facts of so-and-so kind*, and its strict conditional, we assign to each world i as its sphere of accessibility the set of all worlds that are exactly like i in respect of all facts of so-and-so kind, so $\Box(\phi \supset \psi)$ is true at i if and only if ψ is true at all ϕ-worlds that are exactly like i in respect of all facts of so-and-so kind.

* Warning: in some mathematicians' usage, a sphere is a hollow surface. Think of my spheres rather as solid regions, like spheres of influence. In mathematicians' usage, solid 'spheres' are called *balls*.

A degenerate case: corresponding to what we may call *necessity in respect of all facts*, or *fatalistic necessity*, we assign to each world i as its sphere of accessibility the set of all worlds that are exactly like i in all respects whatever. Since 'all respects whatever' includes likeness in respect of identity or nonidentity to i, i alone is like i in all respects whatever; thus each world i has as its sphere of accessibility the set $\{i\}$ having i as its sole member. Then $\Box\phi$ is true at i if and only if ϕ is true at i; and the fatalistic strict conditional $\Box(\phi \supset \psi)$ is true at i if and only if the material conditional $\phi \supset \psi$ is true at i.

Sometimes we do not insist that each world i must belong to its own sphere of accessibility S_i. Corresponding to *deontic* (or *moral*) *necessity*, we assign to each world i as its sphere of accessibility the set of all morally perfect worlds. Then $\Box\phi$ is true at i if and only if ϕ is true at every morally perfect world. A morally imperfect world like ours does not belong to its own sphere of accessibility.

We have another degenerate case: corresponding to what I may call *vacuous necessity*, we assign to each world i as its sphere of accessibility the empty set, making $\Box\phi$ true at i for any sentence ϕ and world i whatever.

We may compare the strictness of different strict conditionals. The more inclusive are the spheres of accessibility, the stricter is the conditional. Suppose we have necessity operators $\Box_1$ and $\Box_2$, corresponding to the assignment to each world i of spheres of accessibility S_i^1 and S_i^2 respectively. Then the strict conditional $\Box_2(\phi \supset \psi)$ is *stricter at* world i than $\Box_1(\phi \supset \psi)$ if and only if S_i^2 properly includes S_i^1. One strict conditional is *stricter* than another if and only if the first is stricter at every world. Note that any strict conditional is implied by any stricter conditional with the same antecedent and consequent.

Thus the logical strict conditional is stricter than any other; the material conditional is the least strict of all the conditionals that obey the constraint that every world is self-accessible; and the physical strict conditional, for instance, falls in between. The vacuous conditional is the least strict conditional of all.

It may happen, of course, that two strict conditionals are incomparable. It may be that they are incomparable at some world because neither sphere includes the other. Or they may be comparable at every world, but one may be stricter at some worlds and the other at other worlds.

Counterfactuals are related to a kind of strict conditional based on comparative similarity of possible worlds. A counterfactual $\phi \mathbin{\Box\!\!\rightarrow} \psi$ is true at a world i if and only if ψ holds at certain ϕ-worlds; but certainly not all ϕ-worlds matter. '*If kangaroos had no tails, they would topple over*' is true (or false, as the case may be) at our world, quite

without regard to those possible worlds where kangaroos walk around on crutches, and stay upright that way. Those worlds are too far away from ours. What is meant by the counterfactual is that, things being pretty much as they are—the scarcity of crutches for kangaroos being pretty much as it actually is, the kangaroos' inability to use crutches being pretty much as it actually is, and so on—if kangaroos had no tails they would topple over.

We might think it best to confine our attention to worlds where kangaroos have no tails and *everything* else is as it actually is; but there are no such worlds. Are we to suppose that kangaroos have no tails but that their tracks in the sand are as they actually are? Then we shall have to suppose that these tracks are produced in a way quite different from the actual way. Are we to suppose that kangaroos have no tails but that their genetic makeup is as it actually is? Then we shall have to suppose that genes control growth in a way quite different from the actual way (or else that there is something, unlike anything there actually is, that removes the tails). And so it goes; respects of similarity and difference trade off. If we try too hard for exact similarity to the actual world in one respect, we will get excessive differences in some other respect.

There is a simpler argument that there is no world where kangaroos have no tails and everything else is as it actually is. Consider all the material conditionals of the form

$\phi \supset$ *kangaroos have tails*

such that ϕ is true at the actual world. If kangaroos had no tails and everything else were as it actually is, then these conditionals would be true as they actually are, for these conditionals are part of the 'everything else'. Also, in most cases, the antecedents would be true as they actually are, for (at least when the antecedent is irrelevant to whether kangaroos have tails) the antecedents also are part of the 'everything else'. But then, unless the world is one where *modus ponens* goes haywire (so that logic itself is not as it actually is!), kangaroos do have tails there after all. I know of nothing wrong with this argument, but I admit that it looks like an unconvincing trick; so I prefer to rely on the considerations of the previous paragraph.

It therefore seems as if counterfactuals are strict conditionals corresponding to an accessibility assignment determined by similarity of worlds—overall similarity, with respects of difference balanced off somehow against respects of similarity. Let S_i, for each world i, be the set of all worlds that are similar to at least a certain fixed degree to the world i. Then the corresponding strict conditional is true at i if and only if the material conditional of its antecedent and consequent is true

throughout S_i; that is, if and only if the consequent holds at all antecedent-worlds similar to at least that degree to *i*.

If we take any one counterfactual, this will do nicely. But trouble may come if we consider several counterfactuals together. (1) '*If I (or you, or anyone else) walked on the lawn, no harm at all would come of it; but if everyone did that, the lawn would be ruined.*' (2) '*If the USA threw its weapons into the sea tomorrow, there would be war; but if the USA and the other nuclear powers all threw their weapons into the sea tomorrow there would be peace; but if they did so without sufficient precautions against polluting the world's fisheries there would be war; but if, after doing so, they immediately offered generous reparations for the pollution there would be peace; . . . .*'* (3) '*If Otto had come, it would have been a lively party; but if both Otto and Anna had come it would have been a dreary party; but if Waldo had come as well, it would have been lively; but. . . .*'

These sequences have the following general form. I include with each asserted counterfactual also the negated opposite, for in the cases I imagine these negated opposites also are held true.

$$\begin{array}{lll} \phi_1 \mathbin{\Box\!\!\rightarrow} \psi & \text{and} & \sim(\phi_1 \mathbin{\Box\!\!\rightarrow} \sim\psi), \\ \phi_1 \,\&\, \phi_2 \mathbin{\Box\!\!\rightarrow} \sim\psi & \text{and} & \sim(\phi_1 \,\&\, \phi_2 \mathbin{\Box\!\!\rightarrow} \psi), \\ \phi_1 \,\&\, \phi_2 \,\&\, \phi_3 \mathbin{\Box\!\!\rightarrow} \psi & \text{and} & \sim(\phi_1 \,\&\, \phi_2 \,\&\, \phi_3 \mathbin{\Box\!\!\rightarrow} \sim\psi), \\ & \vdots & \end{array}$$

With a little ingenuity, it seems possible to prolong such a sequence indefinitely. No one stage in the sequence refutes the theory that the counterfactual is a strict conditional based on similarity, but any two adjacent stages do. The counterfactual on the left at any stage contradicts the negated counterfactual on the right at the next stage. Take the first and second stages: no matter how the spheres of accessibility may be assigned, if ψ is true at every accessible ϕ_1-world, then ψ is true at every accessible $(\phi_1 \,\&\, \phi_2)$-world. So if the counterfactual is any strict conditional whatever, then $\phi_1 \mathbin{\Box\!\!\rightarrow} \psi$ implies $\phi_1 \,\&\, \phi_2 \mathbin{\Box\!\!\rightarrow} \psi$ and contradicts $\sim(\phi_1 \,\&\, \phi_2 \mathbin{\Box\!\!\rightarrow} \psi)$. Likewise $\phi_1 \,\&\, \phi_2 \mathbin{\Box\!\!\rightarrow} \sim\psi$ implies $\phi_1 \,\&\, \phi_2 \,\&\, \phi_3 \mathbin{\Box\!\!\rightarrow} \sim\psi$ and contradicts $\sim(\phi_1 \,\&\, \phi_2 \,\&\, \phi_3 \mathbin{\Box\!\!\rightarrow} \sim\psi)$, and so on down the sequence.

The left-hand counterfactuals make trouble for the theory that the counterfactual is a strict conditional, even without their negated

* J. Howard Sobel first brought such combinations of counterfactuals to my attention, pointing out that they are characteristic of the situations in which act- and rule-utilitarianism seem to prescribe different courses of action. Sobel has applied my theory of counterfactuals in examining the claim that act- and rule-utilitarianism are extensionally equivalent; see his 'Utilitarianisms: Simple and General', *Inquiry* 13 (1970): 394–449.

opposites. If those at two adjacent stages both are true, then according to the theory the second is true vacuously. So are all those beyond it. Beginning at the beginning: if ψ is true at every accessible ϕ_1-world but $\sim\psi$ is true at every accessible $(\phi_1 \,\&\, \phi_2)$-world, then there must not be any accessible $(\phi_1 \,\&\, \phi_2)$-worlds—nor any accessible $(\phi_1 \,\&\, \phi_2 \,\&\, \phi_3)$-worlds, nor.... Then if the lower counterfactuals are true, it is no thanks to their consequents: if a strict conditional is vacuously true, then so is any other with the same antecedent. From the premises that if Otto had come it would have been lively and that if Otto and Anna had come it would have been dreary, it follows that if Otto and Anna had come then the cow would have jumped over the moon. Since that does *not* follow, the counterfactual is not a strict conditional.

If we treat the counterfactual as a strict conditional based on similarity, then the best we can do for our troublesome sequences is to keep changing our minds about which such strict conditional it is. We may be able to make the two sentences at any one stage true by an appropriate choice of a sphere of accessibility based on similarity, but we must choose anew for each stage. If so, we have the situation shown

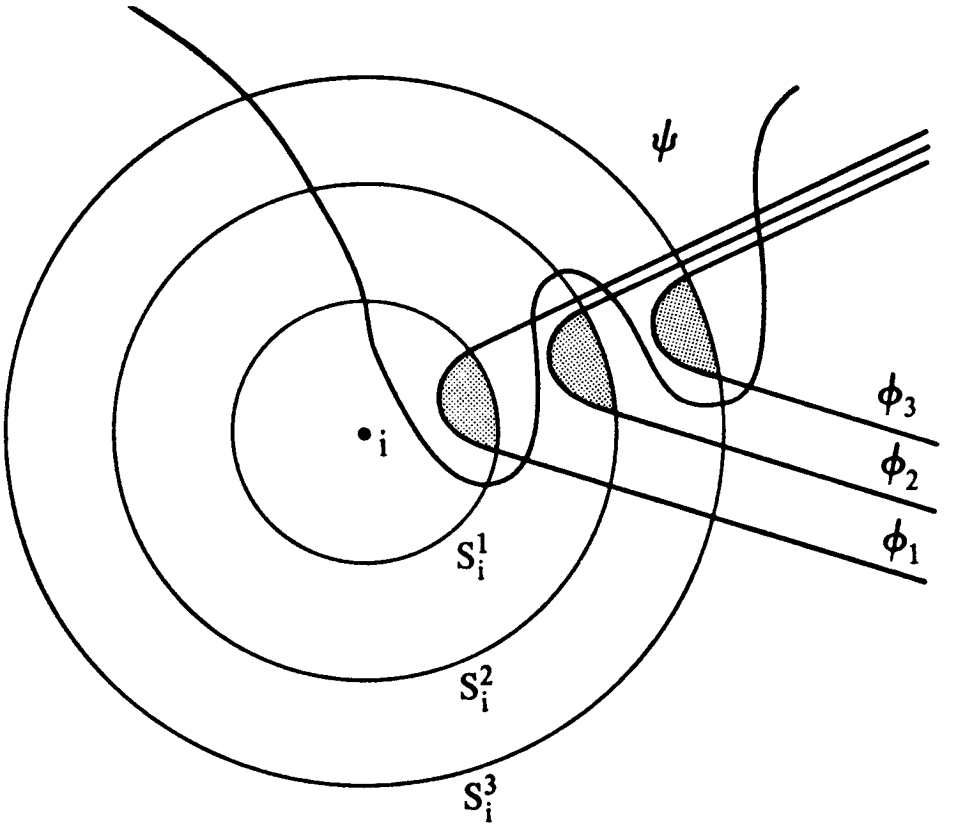


FIGURE 2

in Figure 2. Suppose we have a sphere S_i^1 around i that is right for the first stage: ψ is true at every ϕ_1-world in S_i^1, and—since there are ϕ_1-worlds in S_i^1—it is not the case that $\sim\psi$ also is true at every ϕ_1-world in S_i^1. Then S_i^1 is wrong for the second stage. So is any sphere smaller than S_i^1. But by changing our minds about the degree of similarity to i that we require, perhaps we can find a sphere S_i^2 that is right for the second stage. S_i^2 corresponds to less stringent standards of similarity than S_i^1, and to a stricter conditional. (The stringency of the standards of similarity goes inversely with the strictness of the conditional. *Less* stringent standards of similarity bring *more* worlds into accessibility, making it *more* difficult for anything to hold at all those worlds.) S_i^2 is wrong for the first stage; in order to handle the second stage we had to expand the sphere of accessibility to reach some $(\phi_1 \,\&\, \phi_2 \,\&\, \sim\psi)$-worlds, and these falsify the first-stage counterfactual. S_i^2 is wrong also for the third stage. So is any sphere smaller than S_i^2. But by changing our minds once again, perhaps we can find a still larger sphere S_i^3—a still less stringent standard of similarity, a still stricter conditional—that is right for the third stage. It is wrong for the second and first, however; and for the fourth, if the sequence continues. In short: it may be that for every stage of the sequence, there is a choice of strictness that is right for that stage. But as we go down the sequence, we need stricter and stricter conditionals. The choice that works at any one stage makes false all the counterfactuals at previous stages, and all the negated opposites at subsequent stages. If counterfactuals are strict conditionals we have no hope of deciding, once and for all, how strict they are.

It will not help to plead vagueness. If counterfactuals were strict conditionals based on similarity, indeed they would presumably be vague ones. The assignment of spheres of accessibility for them would be fixed only within rough limits. This might happen both because our ways of trading off respects of similarity and difference against each other are not well fixed and because the degree of overall similarity to a world i that is set as a condition of membership in the sphere of accessibility around i is not well fixed. Both sources of vagueness would tend to make some counterfactuals indefinite in truth value, since the truth value will come out differently under different equally acceptable resolutions of the vagueness. But the counterfactuals and their negated opposites in our troublesome sequence are *not* necessarily especially indefinite in their truth value. I think it is clear from my examples that such a sequence could consist of counterfactuals and their negated opposites all of which are as definitely true as counterfactuals ever are (except for those paragon counterfactuals in which the antecedent logically implies the consequent).

Neither will it help to plead dependence on context. If counterfactuals were vague strict conditionals, no doubt context would resolve some of the vagueness, and different contexts would sometimes resolve it differently. But our problem is not a conflict between counterfactuals in different contexts, but rather between counterfactuals in a single context. It is for this reason that I put my examples in the form of a single run-on sentence, with the counterfactuals of different stages conjoined by semicolons and 'but'. While one context may favor a delineation of baldness on which Dudley is bald, and another may favor a delineation on which he is not, no context can favor a delineation on which he both is and is not. There is no such delineation. While one context might favor a level of strictness on which the first-stage pair in our sequence are both true, and another may favor a greater strictness on which the second-stage pair are both true, and still another may favor a still greater strictness on which the third-stage pair are both true, and so on, none can favor a strictness on which the four sentences from the pairs at two adjacent stages are all true. There is no such strictness.

It is still open to say that counterfactuals are vague strict conditionals based on similarity, and that the vagueness is resolved—the strictness is fixed—by very local context: the antecedent itself. That is not altogether wrong, but it is defeatist. It consigns to the wastebasket of contextually resolved vagueness something much more amenable to systematic analysis than most of the rest of the mess in that wastebasket.

1.3 Variably Strict Conditionals

Counterfactuals are like strict conditionals based on similarity of worlds, but there is no saying how strict they are. They come in as many different strictnesses as there can be stages in my sequence of counterfactuals and their negated opposites. I suggest, therefore, that the counterfactual is not any one strict conditional, but is rather what I shall call a *variably strict conditional.* Any particular counterfactual is as strict, within limits, as it must be to escape vacuity, and no stricter.

Corresponding to any (constantly) strict conditional, as we have seen, there is an assignment to each world i of a single sphere of accessibility S_i around i. Corresponding to a variably strict conditional, on the other hand, there must be an assignment to each world i of a *set* $\$_i$ of spheres of accessibility around i, some larger and some smaller. Such an assignment is required to meet certain formal constraints, laid down in the following definition. We shall see later how, and to what extent, these constraints are justified.

Let $\$$ be an assignment to each possible world i of a set $\$_i$ of sets of

possible worlds. Then $ is called a (*centered**) *system of spheres*, and the members of each $$_i$ are called *spheres* around *i*, if and only if, for each world *i*, the following conditions hold.

(C) $$_i$ is *centered on i*; that is, the set $\{i\}$ having *i* as its only member belongs to $$_i$.

(1) $$_i$ is *nested*; that is, whenever *S* and *T* belong to $$_i$, either *S* is included in *T* or *T* is included in *S*.

(2) $$_i$ is *closed under unions*; that is, whenever $\mathcal{S}$ is a subset of $$_i$ and $\bigcup\mathcal{S}$ is the set of all worlds *j* such that *j* belongs to some member of $\mathcal{S}$, $\bigcup\mathcal{S}$ belongs to $$_i$.

(3) $$_i$ is *closed under (nonempty) intersections*; that is, whenever $\mathcal{S}$ is a nonempty subset of $$_i$ and $\bigcap\mathcal{S}$ is the set of all worlds *j* such that *j* belongs to every member of $\mathcal{S}$, $\bigcap\mathcal{S}$ belongs to $$_i$.

The system of spheres used in interpreting counterfactuals is meant to carry information about the comparative overall similarity of worlds. Any particular sphere around a world *i* is to contain just those worlds that resemble *i* to at least a certain degree. This degree is different for different spheres around *i*. The smaller the sphere, the more similar to *i* must a world be to fall within it. To say the same thing in purely comparative terms: whenever one world lies within some sphere around *i* and another world lies outside that sphere, the first world is more closely similar to *i* than the second. Conversely, if *S* is any set of worlds such that every member of *S* is more similar to *i* than any non-member of *S*, then *S* should be one of the spheres around *i*. (An exception: we may or may not count the set of *all* worlds as one of the spheres around *i*, although it vacuously meets the condition just given.)‡

Our four formal constraints in the definition of a centered system of spheres are justified because, if they were not met, the spheres could not very well be regarded as carrying information about comparative similarity of worlds.

(C) Surely each world *i* is as similar to itself as any other world is to it;

* We may omit the qualifying adjective 'centered' for the most part, restoring it only when we have need to discuss systems of spheres that are perhaps not centered: that is, assignments $ that satisfy conditions (1), (2), and (3) but perhaps not (C).

‡ Whether or not a quantitative concept of similarity 'distance' between worlds makes sense, I need only the non-quantitative, comparative concept given by means of a system of spheres. In topology also we find a non-quantitative concept of distance, given sometimes by means of a system of 'neighborhoods'. Neighborhoods are something like my spheres, but there is one important difference: because topological neighborhoods around a point are not in general nested, they yield a purely qualitative concept of distance—not quantitative, but not even comparative. We can say whether one point or point-set is at all separated from another; but if *A* and *B* both are separated from *C* we cannot say whether one separation exceeds the other.

therefore i should belong to every (nonempty) sphere around i. Almost as surely, no other world is quite as similar to a world i as i itself is; even if there were a world j qualitatively indiscernible from i (imagining for the moment that possible worlds are not the sort of things that obey a non-trivial law of identity of indiscernibles) we might still argue that i does, and j does not, resemble i in respect of being identical to i. Therefore some sphere around i should contain i and exclude all other worlds; that is, $\{i\}$ should be a sphere around i.

(1) If some $\$_i$ were not nested, we would have two spheres S and T in $\$_i$, and two worlds j and k, such that j lies within S but outside T, and k lies within T but outside S. If S and T both carried information about comparative similarity to i, then j would be more similar than k to i (because j does and k does not lie within the sphere S) but also k would be more similar than j to i (because k does and j does not lie within T). We cannot have it both ways.

(2) Suppose j does, and k does not, lie within the union $\bigcup\$$ of a set $\$$ of spheres around i. It follows that j does, and k does not, lie within some sphere S in $\$$, and hence that j is more similar than k to i. Therefore $\bigcup\$$ is a set such that any world within it is more similar to i than any world outside it, and such a set should be a sphere around i.

(3) Similarly, suppose j does, and k does not, lie within the intersection $\bigcap\$$ of a nonempty set $\$$ of spheres; then j does, and k does not, lie within some sphere S in $\$$; so j is more similar than k to i. $\bigcap\$$ is a set such that any world within it is more similar to i than any world outside it, and hence should be a sphere around i.

Note that conditions (2) and (3) of closure under union and intersection are automatically satisfied when there are only finitely many spheres around i, or in the case of a finite subset $\$$ of an infinite $\$_i$. If there is a biggest sphere in $\$$ (one that includes all the others) it is $\bigcup\$$. If there is a smallest sphere in $\$$ (one that is included in all the others) it is $\bigcap\$$. By nesting, every finite set of spheres around a world has a biggest and a smallest. But not so an infinite set: it may have bigger and bigger spheres without end, or smaller and smaller spheres without end. It would simplify things considerably if we could rule out this annoying possibility by fiat; but we shall see that such a fiat would be unjustifiable.

Condition (2) of closure under unions implies that the empty set is a sphere around each i; for in (2) I did not require $\$$ to be nonempty, and by definition the union of empty $\$$ is empty. To include the empty sphere is technically convenient, but unintuitive; however, it can easily be verified that the presence of the empty sphere has no effect at all on the truth conditions to be given with reference to the system of spheres.

More important, I have left it open whether or not the set of all

possible worlds is to be one of the spheres around each world i; or in other words, whether or not the union $\bigcup\$_i$ of all spheres around i is to exhaust the set of worlds; or, in still other words, whether or not every possible world is to lie within some or other sphere around i. If $\bigcup\$_i$ is the set of all worlds, for each i, I will call $\$$ *universal*. If not, then I regard the worlds that the spheres around i do not reach—those that lie outside $\bigcup\$_i$—as being all equally similar to i, and less similar to i than any world that the spheres do reach. We will see that any such world will be left out of consideration in determining whether a counterfactual is true at i. It is as if, from the point of view of i, these remotest worlds were not possible worlds at all.

Now that we have set up this Ptolemaic astronomy, we are ready to use it to give truth conditions for counterfactual conditionals, as follows.

$\phi \mathbin{\Box\!\!\rightarrow} \psi$ is true at a world i (according to a system of spheres $\$$) if and only if either

(1) no ϕ-world belongs to any sphere S in $\$_i$, or

(2) some sphere S in $\$_i$ does contain at least one ϕ-world, and $\phi \supset \psi$ holds at every world in S.

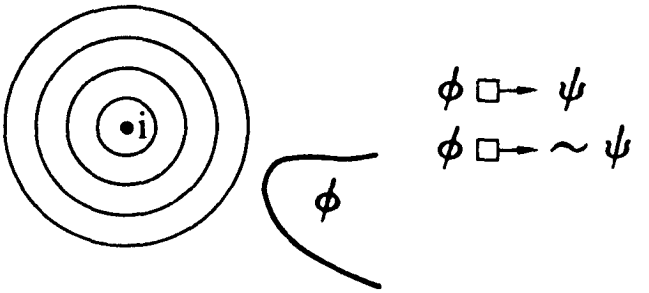
Alternative (1) gives the vacuous case: either ϕ is true at no world, or it is true only at worlds outside $\bigcup\$_i$. Then our counterfactual is vacuously true at i. We shall say in this case that ϕ is not *entertainable*, at i, as a counterfactual supposition. Alternative (2) gives the principal case: ϕ is an entertainable supposition at i, and within some sphere around i that is big enough to reach at least one ϕ-world—call such a sphere *ϕ-permitting*—ψ is true at all ϕ-worlds. In brief: a counterfactual is vacuously true if there is no antecedent-permitting sphere, non-vacuously true if there is some antecedent-permitting sphere in which the consequent holds at every antecedent-world, and false otherwise.

Figure 3 depicts the four cases that might arise for a counterfactual $\phi \mathbin{\Box\!\!\rightarrow} \psi$—two ways for it to be true at a world i, and two ways for it to be false.

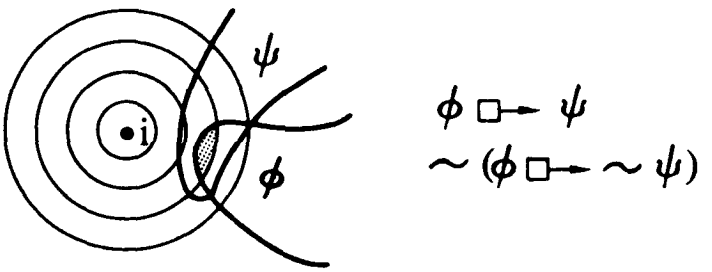
In case (A), there is no ϕ-permitting sphere. Even the outermost sphere around i does not reach the ϕ-worlds, if indeed there are any. Then every counterfactual with antecedent ϕ is vacuously true at i; for instance, both $\phi \mathbin{\Box\!\!\rightarrow} \psi$ and its opposite $\phi \mathbin{\Box\!\!\rightarrow} \sim\psi$. It does not matter (and so is not shown) where the ψ-worlds are, or even whether there are any.

In case (B), there is a ϕ-permitting sphere around i within which ψ holds at all ϕ-worlds—namely, the next-to-outermost sphere. $\phi \supset \psi$ holds throughout this sphere. Therefore $\phi \mathbin{\Box\!\!\rightarrow} \psi$ is non-vacuously true. One such sphere is enough to make it true; it does no harm that there also is a larger ϕ-permitting sphere—the outermost—that reaches

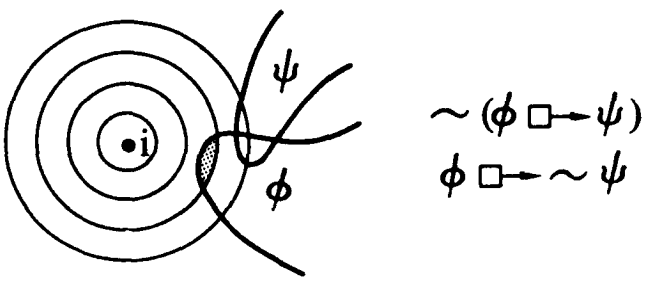
(A) VACUOUS TRUTH



(B) NON-VACUOUS TRUTH



(C) FALSITY- - OPPOSITE TRUE



(D) FALSITY - - OPPOSITE FALSE

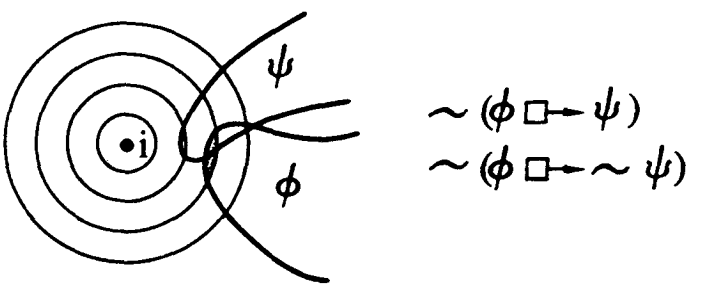


FIGURE 3

ϕ-worlds where ψ is false. The opposite counterfactual $\phi \mathbin{\Box\!\!\rightarrow} \sim\psi$ is false: there are ϕ-permitting spheres, and both of them contain ϕ-worlds where the new consequent $\sim\psi$ is false.

Case (C) is the other way around. There are ϕ-permitting spheres, but none in which ψ holds at every ϕ-world, so none throughout which $\phi \supset \psi$ holds. Therefore $\phi \mathbin{\Box\!\!\rightarrow} \psi$ is false. In the inner one of the two ϕ-permitting spheres, $\sim\psi$ holds at every ϕ-world; so the opposite counterfactual $\phi \mathbin{\Box\!\!\rightarrow} \sim\psi$ is true.

In case (D), finally, there are ϕ-permitting spheres, and both of them contain a mixture of ϕ-worlds where ψ holds and ϕ-worlds where $\sim\psi$ holds. Therefore $\phi \mathbin{\Box\!\!\rightarrow} \psi$ and its opposite $\phi \mathbin{\Box\!\!\rightarrow} \sim\psi$ both are false.

Let us reconsider the sequences of true counterfactuals and their true negated opposites that drove us to give up the theory that the counterfactual is a constantly strict conditional based on similarity:

$\phi_1 \mathbin{\Box\!\!\rightarrow} \psi$ and $\sim(\phi_1 \mathbin{\Box\!\!\rightarrow} \sim\psi)$,
$\phi_1 \,\&\, \phi_2 \mathbin{\Box\!\!\rightarrow} \sim\psi$ and $\sim(\phi_1 \,\&\, \phi_2 \mathbin{\Box\!\!\rightarrow} \psi)$,
$\phi_1 \,\&\, \phi_2 \,\&\, \phi_3 \mathbin{\Box\!\!\rightarrow} \psi$ and $\sim(\phi_1 \,\&\, \phi_2 \,\&\, \phi_3 \mathbin{\Box\!\!\rightarrow} \sim\psi)$,

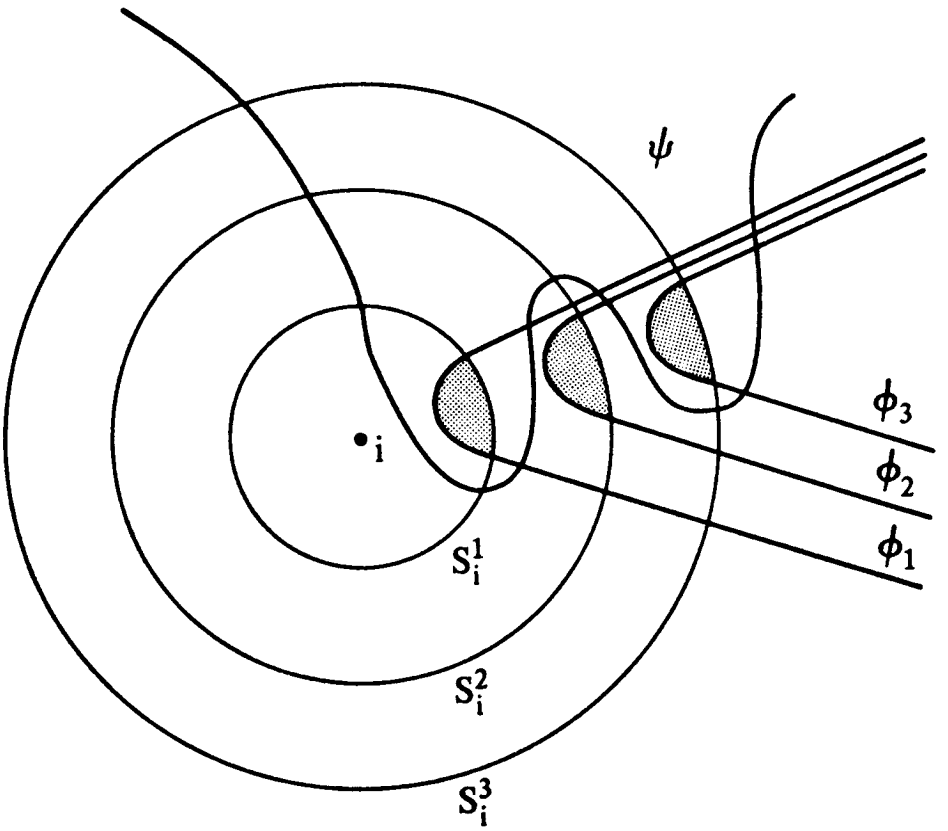


FIGURE 2

and so on. Figure 2 portrayed our difficulty: taking the counterfactual as a constantly strict conditional, we had to choose one of the spheres $S_i^1, S_i^2, S_i^3, \ldots$ to be *the* sphere of accessibility around i, but no choice was right. S_i^1 was right for the first stage of the sequence but not the second, S_i^2 was right for the second stage but not the first or third, and so on. Now Figure 2 portrays the solution: taking the counterfactual as a variably strict conditional, we do not need to choose. The several spheres are all present in $\$_i$ together. S_i^1 is there to make the first-stage counterfactual non-vacuously true, S_i^2 is there to make the second-stage counterfactual non-vacuously true, and so on. The stages can coexist in peace.

1.4 The Limit Assumption

If there are only finitely many spheres around some world i, then any nonempty set of these spheres has a smallest member: a sphere in the set that is included in every other sphere in the set. In particular, for any entertainable antecedent, the set of antecedent-permitting spheres has a smallest member. This smallest antecedent-permitting sphere is the intersection of the set of all antecedent-permitting spheres around i. It contains the antecedent-worlds closest to i: all and only those of the antecedent-worlds than which no other antecedent world is closer to i.

The same may be true even if there are infinitely many spheres around i, provided we have no infinite descending sequence of smaller and smaller spheres without end. (That is: if the ordering of the spheres by inclusion is a well-ordering.) For instance, if we could number the spheres in such a way that sphere 0 is the smallest (the empty set), sphere 1 is the next smallest (the sphere $\{i\}$), sphere 2 is the next smallest, and so on, then there would be a smallest member of every set of spheres, and in particular there would be a smallest antecedent-permitting sphere for every entertainable antecedent.

If there are sequences of smaller and smaller spheres without end, then there are sets of spheres with no smallest member: take the set of all spheres in any such sequence. Yet it might still happen that for every entertainable antecedent in our language, there is a smallest antecedent-permitting sphere. For our language may be limited in expressive power so that not just any set of worlds is the set of ϕ-worlds for some sentence ϕ; and, in that case, it may never happen that the set of ϕ-permitting spheres is one of the sets that lacks a smallest member, for any antecedent ϕ.

The assumption that, for every world i and antecedent ϕ that is entertainable at i, there is a smallest ϕ-permitting sphere, I call the

Limit Assumption. It is the assumption that as we take smaller and smaller antecedent-permitting spheres, containing antecedent-worlds closer and closer to i, we eventually reach a limit: the *smallest* antecedent-permitting sphere, and in it the *closest* antecedent-worlds.

If the consequent of a counterfactual holds at all antecedent-worlds within some antecedent-permitting sphere around i, then also the consequent holds at all antecedent-worlds in any smaller antecedent-permitting sphere. In particular, the consequent holds at all antecedent-worlds in the smallest antecedent-permitting sphere, if such there be. Conversely, if the consequent holds at all antecedent-worlds in the smallest antecedent-permitting sphere, then the consequent holds at all antecedent worlds in some antecedent-permitting sphere. Under the Limit Assumption, therefore, we could make the truth conditions for counterfactuals simpler: a counterfactual is true at i if and only if either (1) there is no antecedent-permitting sphere around i, or (2) the consequent holds at every antecedent-world in the smallest antecedent-permitting sphere around i. Simpler still: a counterfactual is true at i if and only if the consequent holds at every antecedent-world closest to i (where we do not call an antecedent-world outside $\bigcup \$_i$ 'closest', even if it is an antecedent-world than which there is none closer).

Unfortunately, we have no right to assume that there always are a smallest antecedent-permitting sphere and, within it, a set of closest antecedent-worlds. Suppose we entertain the counterfactual supposition that at this point

---

there appears a line more than an inch long. (Actually it is just under an inch.) There are worlds with a line 2″ long; worlds presumably closer to ours with a line $1\frac{1}{2}$″ long; worlds presumably still closer to ours with a line $1\frac{1}{4}$″ long; worlds presumably still closer .... But how long is the line in the *closest* worlds with a line more than an inch long? If it is $1+x$″ for any x however small, why are there not other worlds still closer to ours in which it is $1+\frac{1}{2}x$″, a length still closer to its actual length? The shorter we make the line (above 1″), the closer we come to the actual length; so the closer we come, presumably, to our actual world.* Just as there is no shortest possible length above 1″, so there is no closest world to ours among the worlds with lines more than an

* 'Presumably', here and elsewhere, because it depends on the technique of printing. Suppose the actual line was printed by a digital process of some sort, and the shortest length above 1″ that is possible using this process is $1\frac{1}{8}$″. Then perhaps some world at which this process is used to give a $1\frac{1}{8}$″ line is closest, being closer to ours than any world at which a process unlike the actual process is used to give a slightly shorter line. Thus this and other examples are not quite decisive; but they should suffice at least to deter us from rashly assuming that there *must* be a smallest antecedent-permitting sphere.

inch long, and no smallest sphere permitting the supposition that there is a line more than an inch long.

When there is no smallest antecedent-permitting sphere, our truth conditions amount to this: if there are antecedent-permitting spheres, then as we take smaller and smaller ones without end, eventually we come to ones in which the consequent holds at every antecedent-world.

1.5 'Might' Counterfactuals and Outer Modalities

My interpretation of the 'would' counterfactual as a variably strict conditional, together with my definition

$$\phi \Diamond\!\!\rightarrow \psi =^{df} \sim(\phi \Box\!\!\rightarrow \sim\psi)$$

of the 'might' counterfactual in terms of the 'would' counterfactual, yield derived truth conditions for the 'might' counterfactual as follows.

$\phi \Diamond\!\!\rightarrow \psi$ is true at a world i (according to a system of spheres \$) if and only if both

(1) some ϕ-world belongs to some sphere S in $\$_i$, and
(2) every sphere S in $\$_i$ that contains at least one ϕ-world contains at least one world where ϕ & ψ holds.

Under the Limit Assumption, we could restate the derived truth conditions for 'might' counterfactuals thus: $\phi \Diamond\!\!\rightarrow \psi$ is true at i if and only if ψ holds at some ϕ-world in the smallest ϕ-permitting sphere around i. More simply: a 'might' counterfactual is true at i if and only if the consequent holds at some antecedent-world closest to i. (Again, an antecedent-world outside $\bigcup\$_i$ must never count as a closest antecedent-world to i, not even if there are none closer.) But if the Limit Assumption does not hold, then these simplified restatements will not do; the 'might' counterfactual is then true if and only if, as we take smaller and smaller antecedent-permitting spheres around i without end, and thereby confine our attention to antecedent-worlds closer and closer to i, we never leave behind all the antecedent-worlds where the consequent holds.

If the 'would' counterfactual $\phi \Box\!\!\rightarrow \psi$ is non-vacuously true, then the 'might' counterfactual $\phi \Diamond\!\!\rightarrow \psi$ also is true. If $\phi \Box\!\!\rightarrow \psi$ and its opposite $\phi \Box\!\!\rightarrow \sim\psi$ are both false, then $\phi \Diamond\!\!\rightarrow \psi$ and its opposite $\phi \Diamond\!\!\rightarrow \sim\psi$ are both true; for this is the case in which ψ is true at some of the closest ϕ-worlds and $\sim\psi$ is true at others of them. But when $\phi \Box\!\!\rightarrow \psi$ is false and its opposite $\phi \Box\!\!\rightarrow \sim\psi$ is true, ψ holds at none of the closest ϕ-worlds and $\phi \Diamond\!\!\rightarrow \psi$ is therefore false. Finally, when ϕ is

not entertainable and $\phi \mathbin{\Box\!\!\rightarrow} \psi$ is therefore vacuously true, $\phi \mathbin{\Diamond\!\!\rightarrow} \psi$ is again false.

Let T be a sentential constant true at every world; let $\bot$ be a sentential constant false at every world. (Or, if you prefer, let T abbreviate some arbitrarily chosen truth-functional tautology and let $\bot$ abbreviate its contradictory negation.) Then the 'would' counterfactual $\phi \mathbin{\Box\!\!\rightarrow} \bot$ cannot be true otherwise than vacuously, when ϕ is not entertainable. Therefore the 'might' counterfactual $\phi \mathbin{\Diamond\!\!\rightarrow} \mathrm{T}$, definitionally equivalent to $\sim(\phi \mathbin{\Box\!\!\rightarrow} \sim \mathrm{T})$ and hence to the negation of $\phi \mathbin{\Box\!\!\rightarrow} \bot$, is a sentence true if and only if ϕ is entertainable.

We may therefore introduce into our language a pair of modal operators defined in terms of the counterfactual conditional connectives:

$$\Diamond\phi =^{\mathrm{df}} \phi \mathbin{\Diamond\!\!\rightarrow} \mathrm{T} \quad \text{(or, equivalently, } \sim(\phi \mathbin{\Box\!\!\rightarrow} \bot)\text{)},$$
$$\Box\phi =^{\mathrm{df}} \sim\Diamond\sim\phi \quad \text{(or, equivalently, } \sim\phi \mathbin{\Box\!\!\rightarrow} \bot\text{)}.$$

We may read $\Diamond$ as '*Possibly* ____' or as '*It is entertainable that* ____'; we may read $\Box$ as '*Necessarily* ____' or as '*It would be the case, no matter what, that* ____'. The two are interdefinable in the usual way: not only is $\Box\phi$ definitionally equivalent to $\sim\Diamond\sim\phi$, as stipulated above, but also it follows that $\Diamond\phi$ is equivalent to $\sim\Box\sim\phi$. Other definitions could be given of the two modal operators:

$$\Diamond\phi =^{\mathrm{df}} \phi \mathbin{\Diamond\!\!\rightarrow} \phi,$$
$$\Box\phi =^{\mathrm{df}} \sim\phi \mathbin{\Box\!\!\rightarrow} \phi,$$

for instance, are equivalent to the definitions given above.

From the truth conditions for counterfactuals and the definitions of the two modal operators, we obtain derived truth conditions for modal sentences as follows.

$\Diamond\phi$ is true at a world i (according to a system of spheres \$) if and only if ϕ is true at some world in some sphere S in $\$_i$.

$\Box\phi$ is true at a world i (according to a system of spheres \$) if and only if ϕ is true at every world in every sphere S in $\$_i$.

We can also express these truth conditions in terms of the assignment to each world i of the set of worlds $\bigcup\$_i$: the union of all the spheres around i (that is, the set of all and only those worlds that belong to some or other sphere around i). $\bigcup\$_i$ is itself a sphere around i; it is the largest, or *outermost* sphere around i.

$\Diamond\phi$ is true at i (according to \$) if and only if ϕ is true at some world in $\bigcup\$_i$.

$\Box\phi$ is true at i (according to \$) if and only if ϕ is true at every world in $\bigcup\$_i$.

Hence our defined modal operators turn out to be interpretable in the usual way by means of accessibility; they correspond to the assignment to each world i of the single sphere of accessibility $\bigcup\$_i$. Since they pertain to the outermost of our spheres around each world i, let us call them the *outer modalities*: outer necessity and outer possibility.

In case our system of spheres is universal, in that each $\bigcup\$_i$ is the set of all possible worlds, then our outer necessity and possibility will be ordinary logical necessity and possibility. If the system of spheres is not universal, so that at least for some worlds i the outermost sphere $\bigcup\$_i$ around i does not exhaust the set of all worlds, then our outer modalities may not be the same as any familiar modalities. They will be rather strict modalities: probably stricter than anything familiar except the logical modalities themselves.

Our reading of $\Diamond$ as '*It is entertainable that* ____' is justified by the fact that, as we have already noted, $\Diamond\phi$ is true at i if and only if ϕ is an entertainable counterfactual supposition at i. That is so, we recall, if and only if there is some ϕ-permitting sphere around i, so that counterfactuals with ϕ as antecedent can be false or non-vacuously true at i. In other words, that is so if and only if ϕ is true at some world in some sphere around i.

Our reading of $\Box$ as '*It would be the case, no matter what, that* ____' is justified by the fact that if $\Box\phi$ is true at a world i, then $\chi \mathbin{\Box\!\!\rightarrow} \phi$ is true at i for any antecedent χ whatever. If χ is not entertainable, then $\chi \mathbin{\Box\!\!\rightarrow} \phi$ is vacuously true; if χ is entertainable, then $\chi \mathbin{\Box\!\!\rightarrow} \phi$ is non-vacuously true because by hypothesis ϕ is true throughout every sphere around i, and hence ϕ is true throughout some χ-permitting sphere around i, and hence $\chi \supset \phi$ is true throughout some χ-permitting sphere around i.

$\Box(\phi \supset \psi)$, the *outer strict conditional*, implies the counterfactual $\phi \mathbin{\Box\!\!\rightarrow} \psi$; if $\phi \supset \psi$ is true throughout every sphere around i, then in particular, if there is any ϕ-permitting sphere around i, it is true throughout that. But not conversely: the counterfactual is not the outer strict conditional any more than it is any other constantly strict conditional, despite the fact that $\Box$ is defined from $\mathbin{\Box\!\!\rightarrow}$. $\phi \mathbin{\Box\!\!\rightarrow} \psi$ is true and $\Box(\phi \supset \psi)$ is false if $\phi \supset \psi$ is true throughout some ϕ-permitting sphere, but false somewhere in some larger ϕ-permitting sphere.

Indeed the counterfactual cannot be defined in any way whatever from the outer modalities and truth-functional connectives. Given a system of spheres, we may consider what happens to the truth values of sentences when spheres are added or deleted, but in such a way as never to change the outermost sphere around any world. The truth values at worlds of counterfactuals (with non-counterfactual antecedents and consequents) will in some cases change when non-outermost spheres

are added or deleted; but such additions or deletions never could change the truth value at any world of any sentence built up from non-counterfactual sentences by means of the outer modalities and truth-functional connectives alone. Therefore some counterfactuals cannot be definitionally equivalent to any such sentences.

1.6 Impossible Antecedents

There is at least some intuitive justification for the decision to make a 'would' counterfactual with an impossible antecedent come out vacuously true. Confronted by an antecedent that is not really an entertainable supposition, one may react by saying, with a shrug: If that were so, anything you like would be true! Further, it seems that a counterfactual in which the antecedent logically implies the consequent ought always to be true; and one sort of impossible antecedent, a self-contradictory one, logically implies any consequent.

Moreover, one sometimes asserts counterfactuals by way of *reductio* in philosophy, mathematics, and even logic. (I have done so in this very chapter.) These counterfactuals are asserted in argument, and must therefore be thought true; but their antecedents deny what are thought to be philosophical, mathematical, or even logical truths, and must therefore be thought not only false but impossible. These asserted counterphilosophicals, countermathematicals, and counterlogicals look like examples of vacuously true counterfactuals.

There are other things they might be, however. They might not really be counterfactuals, but subjunctive conditionals of some other kind. More interesting, they might be non-vacuously true counterfactuals, understood in the way I have proposed; but so understood under the pretense that along with the *possible* possible worlds that differ from our world only in matters of contingent, empirical fact, there also are some *impossible* possible worlds that differ from our world in matters of philosophical, mathematical, and even logical truth. (The pretense need not be taken very seriously to explain what happens in conversation; it just might be that this part of our conversational practice is founded upon a confused fantasy.) These alternative hypotheses have the merit that they might explain how we could discriminate in truth value between different counterfactuals with impossible antecedents, whereas my theory makes all of them alike come out vacuously true.

I do not think, however, that we need to discriminate in truth value among such counterfactuals. Of course there are some we would assert and some we would not:

If there were a largest prime p, p!+1 would be prime.
If there were a largest prime p, p!+1 would be composite.

are both sensible things to say, but

If there were a largest prime p, there would be six regular solids.
If there were a largest prime p, pigs would have wings.

are not. But what does that prove? We have to explain why things we do want to assert are true (or at least why we take them to be true, or at least why we take them to approximate to truth), but we do not have to explain why things we do not want to assert are false. We have plenty of cases in which we do not want to assert counterfactuals with impossible antecedents, but so far as I know we do not want to assert their negations either. Therefore they do not have to be made false by a correct account of truth conditions; they can be truths which (for good conversational reasons) it would always be pointless to assert.

Therefore I am fairly content to let counterfactuals with impossible antecedents be vacuously true. But my reasons are less than decisive, and some might prefer a stronger 'would' counterfactual that cannot be vacuously true. We write this as $\Box\!\!\Rightarrow$, and give it the following truth conditions:

$\phi \Box\!\!\Rightarrow \psi$ is true at a world i (according to a system of spheres \$) if and only if there is some sphere S in $\$_i$ such that S contains at least one ϕ-world, and $\phi \supset \psi$ holds at every world in S.

Preserving the interdefinability of 'would' and 'might' counterfactuals as before, we introduce also a weakened 'might' counterfactual $\Diamond\!\!\Rightarrow$, vacuously true whenever its antecedent is impossible. It is defined by

$$\phi \Diamond\!\!\Rightarrow \psi =^{df} \sim(\phi \Box\!\!\Rightarrow \sim \psi),$$

and it has the following derived truth conditions:

$\phi \Diamond\!\!\Rightarrow \psi$ is true at a world i (according to a system of spheres \$) if and only if every sphere S in $\$_i$ that contains at least one ϕ-world contains at least one ϕ-world at which $\phi \,\&\, \psi$ holds.

One might perhaps motivate this weakened 'might' in much the same way as I motivated the original, weak 'would': confronted by an antecedent that is not really entertainable, one might say, with a shrug: If that were so, anything you like *might* be true!

I find $\Box\!\!\rightarrow$ and $\Diamond\!\!\rightarrow$, taken as a pair, somewhat better intuitively than $\Box\!\!\Rightarrow$ and $\Diamond\!\!\Rightarrow$; and the simple interdefinability of 'would' and 'might' seems plausible enough to destroy the appeal of the mixed pair of $\Box\!\!\rightarrow$ and $\Diamond\!\!\Rightarrow$. There seems not to be much more to be

said; perhaps ordinary usage is insufficiently fixed to force either choice, and technical convenience may favor one or the other pair depending on how we choose to formulate our truth conditions. (On the present formulation, $\Box\!\!\rightarrow$ and $\Diamond\!\!\rightarrow$ have simpler truth conditions; on the formulation to be given in Section 2.7, $\Box\!\!\rightarrow$ and $\Diamond\!\!\rightarrow$ have simpler truth conditions.) In any case, we have both pairs in stock; and we can get either pair from the other via the following definitions:

$$\phi \Diamond\!\!\rightarrow \psi =^{df} (\phi \Diamond\!\!\rightarrow \phi) \supset (\phi \Diamond\!\!\rightarrow \psi),$$
$$\phi \Box\!\!\rightarrow \psi =^{df} (\phi \Box\!\!\rightarrow \phi) \supset (\phi \Box\!\!\rightarrow \psi).$$

1.7 True Antecedents

We noted at the outset that truth of the antecedent was a defect in a counterfactual, but not necessarily the sort of defect that produces automatic falsity or a truth-value gap. According to the truth conditions I have given, a counterfactual with true antecedent is true if and only if the consequent is true. This is so both for 'would' and 'might' counterfactuals (and for the strong 'would' and weak 'might' counterfactuals introduced in the previous section). In short: counterfactuals with true antecedents reduce to material conditionals.

Suppose the antecedent ϕ is true at a world i. Then there is a ϕ-permitting sphere around i, because $\{i\}$ is a sphere. If the consequent ψ is true at i, then there is a ϕ-permitting sphere around i throughout which $\phi \supset \psi$ holds, to wit $\{i\}$; so $\phi \Box\!\!\rightarrow \psi$ is true at i. Also every ϕ-permitting sphere around i contains a world where $\phi \,\&\, \psi$ holds, since every sphere around i, except the empty set which is not a ϕ-permitting sphere, contains the world i itself; so $\phi \Diamond\!\!\rightarrow \psi$ is true at i. If, on the other hand, the consequent ψ is false at i, then there is no ϕ-permitting sphere around i throughout which $\phi \supset \psi$ holds, since it fails at the world i which belongs to every ϕ-permitting sphere; so $\phi \Box\!\!\rightarrow \psi$ is false at i. Also there is a ϕ-permitting sphere containing no world where $\phi \,\&\, \psi$ holds, to wit $\{i\}$; so $\phi \Diamond\!\!\rightarrow \psi$ is false at i.

I can summarize the status of counterfactuals with true antecedents by noting that the following two inference-patterns are valid: that is, my truth conditions guarantee that whenever the premise is true at a world, so is the conclusion.

$$\frac{\phi \;\&\; \sim\psi}{\therefore\ \sim(\phi \Box\!\!\rightarrow \psi)} \qquad \frac{\phi \;\&\; \psi}{\therefore\ \phi \Box\!\!\rightarrow \psi}.$$

The validity of the first inference-pattern guarantees also the validity

of the inference from a counterfactual to a material conditional and the validity of *modus ponens* from a counterfactual conditional:

$$\frac{\phi \mathbin{\Box\!\!\rightarrow} \psi}{\therefore\ \phi \supset \psi} \quad \text{and} \quad \frac{\phi \mathbin{\Box\!\!\rightarrow} \psi \qquad \phi}{\therefore\ \psi}.$$

How plausible are these consequences of my truth conditions? It is hard to test them directly. It is not much help considering a counterfactual with an antecedent known to be true, and asking whether it seems to be true or false according as the consequent is thought to be true or false. Our principal response will be not that the conditional is true or that it is false, but that it is mistaken and misleading because of its true antecedent. So it is, but that is not at issue. The false information conveyed by using a counterfactual construction with a true antecedent eclipses the falsity or truth of the conditional itself.

It is not safe to put the conditional in indicative form in order to get rid of the presupposition that the antecedent is false. Sometimes when the antecedent is thought to be probably false, so that the counterfactual construction is appropriate, the counterfactual and indicative conditionals are thought to differ in truth value. (We considered Ernest Adams's example of this in Section 1.1.) Therefore we have no right to take for granted that they have the same truth values when the antecedent is thought to be true, differing only in the presuppositions they carry.

What we must do, I think, is consider a dialog in which the participants disagree on the truth of the antecedent. The first speaker does not deliberately violate the prohibition against asserting a counterfactual with a true antecedent; rather, he asserts a counterfactual with an antecedent he takes to be false. The second speaker replies, registering disagreement with the first speaker's manifest supposition that the antecedent is false, but also expressing agreement or disagreement with the first speaker's assertion.

You say: '*If Caspar had come, it would have been a good party.*' I reply: '*That's false; for he* did *come, yet it was a rotten party.*' Or else I reply: '*That's true; for he* did, *and it* was *a good party. You didn't see him because you spent the whole time in the kitchen, missing all the fun.*' Either reply seems perfectly cogent. In each reply, I correct your false belief that Caspar was absent, manifest in your use of the counterfactual form; but I do this while expressing overall disagreement or agreement with your conditional assertion. Moreover, I justify my disagreement or agreement by giving an argument. The argument is

abbreviated, but its presence is signalled by the word 'for' in my reply. The arguments I give have the forms

$$\frac{\phi \ \& \ \sim\psi}{\therefore \textit{ That's false}} \quad \text{and} \quad \frac{\phi \ \& \ \psi}{\therefore \textit{ That's true}}$$

respectively, where ϕ and ψ are the antecedent and consequent of the counterfactual you have just asserted, and 'that' in the conclusion refers to what you have asserted. Therefore my replies are cogent only if the inference-patterns that we want to test,

$$\frac{\phi \ \& \ \sim\psi}{\therefore \ \sim(\phi \mathbin{\Box\!\!\rightarrow} \psi)} \quad \text{and} \quad \frac{\phi \ \& \ \psi}{\therefore \ \phi \mathbin{\Box\!\!\rightarrow} \psi},$$

are valid. The replies do seem cogent; so the inference-patterns are valid; so my truth conditions for counterfactuals with true antecedents are confirmed.

I admit that this test is not quite decisive. It is just possible that my arguments to '*That's false*' and '*That's true*' are invalid with only the premises that appear explicitly in my reply, and depend also on further premises that are understood but not stated. Or it is just possible that what I refer to as 'that' in my reply, and judge false or true, is not the counterfactual you asserted, but rather some belief that I take to have been your reason for thinking the counterfactual true.

The test by dialog is evidence for my truth conditions. What can be said against them? So far as I know, only this: it would seem very odd to pick two completely unrelated truths ϕ and ψ and, on the strength of their truth, to deny the counterfactual $\phi \mathbin{\Box\!\!\rightarrow} \sim\psi$; and even odder to assert the counterfactual $\phi \mathbin{\Box\!\!\rightarrow} \psi$. What would we make of someone who saw fit to deny that if the sky were blue then grass would not be green, or to assert that if the sky were blue then grass would be green? It would be doubly odd. First, because he is using the counterfactual construction with an antecedent he takes to be true, though this construction is customarily reserved for antecedents taken to be false; second, because his assertions could serve no likely conversational purpose that would not be better served by separate assertions of ϕ and ψ. But oddity is not falsity; not everything true is a good thing to say. In fact, the oddity dazzles us. It blinds us to the truth value of the sentences, and we can make no confident judgements one way or the other. We ordinarily take no interest in the truth value of extreme oddities, so we cannot be expected to be good at judging them. They prove nothing at all about truth conditions.

I have claimed that the counterfactuals with true antecedent and false consequent are false, and that those with true antecedent and true

consequent are true. I am fairly sure of both claims, but surer of the first; so it may be of interest to see what changes could be made to keep the first result but not the second.

The first is a consequence of the assumption that no world is more similar to a world i than i itself is; and that seems perfectly safe. The second is a consequence of the assumption that no other world is even *as* similar to i as i itself is; and that is not quite such a safe assumption. Perhaps our discriminations of similarity are rather coarse, and some worlds different from i are enough like i so that such small differences as there are fail to register. In that case, we would need to revise the definition of a system of spheres, weakening the original centering condition (C) which stipulated that $\{i\}$ was to be a sphere around i.

Let \$ be an assignment to each world i of a set $\$_i$ of sets of worlds. Then \$ is a *weakly centered system of spheres* if and only if, for each world i, the following conditions hold.

(W) $\$_i$ is *weakly centered* on i; that is, i belongs to every nonempty sphere around i, and there is at least one nonempty sphere around i.

(1)–(3) $\$_i$ is nested, closed under unions, and closed under (nonempty) intersections; these conditions are unchanged.

In a weakly centered system of spheres, the smallest, or *innermost*, nonempty sphere around i is the intersection of all nonempty spheres around i—that is, $\bigcap(\$_i - \{\Lambda\})$. It contains the closest worlds to i. The world i itself is one of these closest worlds to i; but there may be others as well—worlds differing negligibly from i, so that they come out just as close to i as i itself.

Having weakened our conditions on the system of spheres, we can leave the truth conditions for counterfactuals unchanged and still have the intended result: a counterfactual with true antecedent and false consequent must be false, but one with true antecedent and true consequent may be either true or false. Suppose ϕ is true at a world i; then the smallest ϕ-permitting sphere around i is the innermost nonempty sphere around i. This sphere contains i itself. It may or may not contain other worlds, now that we have (temporarily!) retreated from centering to weak centering. If it does, there may or may not be ϕ-worlds other than i among them. Suppose there are; then $\phi \mathrel{\Box\!\!\rightarrow} \psi$ holds at i if and only if the consequent ψ holds not only at i itself but also at the other ϕ-worlds in the innermost nonempty sphere around i. Thus it may happen that a counterfactual with true antecedent and consequent is false if the consequent is false at a sufficiently close antecedent-world.

When we weaken centering, then a distinction appears between

truth at i itself and truth at all or some of the worlds in the innermost nonempty sphere around i. To express the latter, we may introduce a second pair of modal operators, defined ultimately in terms of the counterfactual connectives. These will pertain to the innermost nonempty sphere around each world i, so let us call them the *inner modalities:* inner necessity and inner possibility.

$$\boxdot\phi =^{df} \mathsf{T} \,\Box\!\!\Rightarrow \phi \quad \text{(or, equivalently, } \Diamond\mathsf{T} \;\&\; \mathsf{T} \,\Box\!\!\rightarrow \phi\text{)},$$
$$\diamondsuit\phi =^{df} \mathsf{T} \,\Diamond\!\!\Rightarrow \phi \quad \text{(or, equivalently, } \Diamond\mathsf{T} \supset \mathsf{T} \,\Diamond\!\!\rightarrow \phi\text{)}.$$

We obtain derived truth conditions for the inner modalities as follows.

$\boxdot\phi$ is true at i if and only if ϕ is true at every world in some nonempty sphere around i.
$\diamondsuit\phi$ is true at i if and only if ϕ is true at some world in every nonempty sphere around i.

Given that we have an innermost nonempty sphere around i, the truth conditions can be stated more simply: the inner modalities are interpreted by means of accessibility, the appropriate assignment of spheres of accessibility being the assignment to each world i of the innermost nonempty sphere around i, that is $\bigcap(\$_i - \{\Lambda\})$, as its single sphere of accessibility. $\boxdot\phi$ is true at i if and only if ϕ holds throughout the innermost nonempty sphere around i, and thus means that ϕ holds at every maximally close world. $\diamondsuit\phi$ is true at i if and only if ϕ holds somewhere in the innermost nonempty sphere around i, and thus means that ϕ holds at some maximally close world.

The outermost sphere includes the innermost nonempty sphere; therefore outer necessity is stricter than inner necessity. Therefore $\Box\phi$ implies $\boxdot\phi$ and $\diamondsuit\phi$ implies $\Diamond\phi$.

So long as we confine our attention to weakly centered systems of spheres, the inner modalities could be defined more simply as $\mathsf{T} \,\Box\!\!\rightarrow \phi$ and $\mathsf{T} \,\Diamond\!\!\rightarrow \phi$ respectively. According to any weakly centered system of spheres, these definitions are exactly equivalent to those I gave, since T is true at every world and hence never fails to be entertainable. But in Section 5.1 I shall give a deontic reinterpretation of our language, on which it will be appropriate to give up even weak centering. Then there may not be any nonempty spheres around a world; in which case nothing, not even T, is entertainable and the definitions no longer will be equivalent. Then it will prove advantageous to have defined the inner modalities as I did.

If we insist—correctly, I think—on interpreting the counterfactuals by means of a centered system of spheres, then it is pointless to con-

sider the inner modalities. Under unweakened centering, the inner modalities are trivial: both $\boxdot\phi$ and $\diamondsuit\phi$ are equivalent to ϕ itself.*

We noted that the counterfactual cannot be defined from truth-functional connectives and the outer modalities; neither can it be defined from these plus the inner modalities. That is so whether we assume centering, weak centering, or neither. The reason is that we can change truth values of counterfactuals by adding or deleting spheres that are neither outermost nor innermost, but we cannot in this way change the truth value of any sentence built up from non-counterfactual sentences by means of truth-functional connectives and outer and inner modalities.

1.8 Counterfactual Fallacies

Certain inferences are correct for the material conditional, and indeed for any constantly strict conditional, but not for variably strict conditionals. The inference fails because the strictness varies between different conditionals in the premises and conclusion. Three especially important inferences that fail for variably strict conditionals may be called the *fallacy of strengthening the antecedent*, the *fallacy of transitivity*, and the *fallacy of contraposition.*‡

The fallacy of *strengthening the antecedent* is the invalid inference-pattern:

$$\frac{\phi \mathrel{\Box\!\!\rightarrow} \psi}{\therefore\ \phi\ \&\ \chi \mathrel{\Box\!\!\rightarrow} \psi}.$$

We have already noted that the premise of such an inference may be true and the conclusion false, in connection with my sequences of counterfactuals and their negated opposites with stronger and stronger antecedents and consequents alternating between a sentence and its negation. The consistency of such sequences, and therefore the invalidity of inference by strengthening the antecedent, was indeed the principal evidence I gave that counterfactuals were variably, not constantly, strict conditionals.

Adding a conjunct to an antecedent is only one among many ways to

* The observation that two different pairs of modalities are definable from the counterfactual, both non-trival under weak centering, is due to Sobel.

‡ These three fallacies have been discussed by Robert Stalnaker from the standpoint of a theory equivalent (as explained in Section 3.4) to a special case of mine. See Stalnaker, 'A Theory of Conditionals', in N. Rescher, *Studies in Logical Theory* (Blackwell: Oxford, 1968). An extensive survey of these and other counterfactual fallacies is given in Sobel, 'Utilitarianisms: Simple and General' (appendix).

strengthen it. A more general form of the fallacy of strengthening the antecedent is as an invalid inference-pattern with two premises:

$$\frac{\begin{array}{l}\Box(\chi \supset \phi) \\ \phi \mathrel{\Box\!\!\rightarrow} \psi\end{array}}{\therefore\ \chi \mathrel{\Box\!\!\rightarrow} \psi}.$$

In the special case that χ is the conjunction of ϕ and something else, the strict conditional $\Box(\chi \supset \phi)$ will hold. The inference is fallacious even if outer necessity is logical necessity, and *a fortiori* also if it is a less strict necessity. For a counterexample to inference by strengthening the antecedent, in which the strengthening is done otherwise than by adding a conjunct, consider this invalid argument.

$\Box$ (*I started at 5 this morning* $\supset$ *I started before 6*)
If I had started before 6, I would have arrived before noon.
$\therefore$ *If I had started at 5, I would have arrived before noon.*

Certainly the first premise is true. To see how the second premise may be true and the conclusion false, suppose that in fact I started just after 6, tried out a new shortcut that turned out to cut two hours off the usual time for the journey, and arrived at noon exactly; but suppose that if I had started at 5, I would have been too sleepy to remember to try the shortcut. (I am supposing that the later I started, in the range of times permitted by the antecedent, the closer an antecedent-world is to our actual world; this may be so, but might not be if, for instance, I planned on starting at 5 and failed to do so only because my alarm did not quite wake me up.)

The *fallacy of transitivity* is the invalid inference-pattern

$$\frac{\begin{array}{l}\chi \mathrel{\Box\!\!\rightarrow} \phi \\ \phi \mathrel{\Box\!\!\rightarrow} \psi\end{array}}{\therefore\ \chi \mathrel{\Box\!\!\rightarrow} \psi}.$$

The fallacy of transitivity is a further generalization of the fallacy of strengthening the antecedent. From the strict conditional $\Box(\chi \supset \phi)$ we can correctly infer $\chi \mathrel{\Box\!\!\rightarrow} \phi$; from that and $\phi \mathrel{\Box\!\!\rightarrow} \psi$ we can fallaciously infer $\chi \mathrel{\Box\!\!\rightarrow} \psi$ by transitivity. Inference by transitivity would thus justify inference by strengthening the antecedent; since we know that the latter is fallacious, so is the former. For a direct counterexample to transitivity, consider this argument:

If Otto had gone to the party, then Anna would have gone.
If Anna had gone, then Waldo would have gone.
$\therefore$ *If Otto had gone, then Waldo would have gone.*

The fact is that Otto is Waldo's successful rival for Anna's affections. Waldo still tags around after Anna, but never runs the risk of meeting Otto. Otto was locked up at the time of the party, so that his going to it is a far-fetched supposition; but Anna almost did go. Then the premises are true and the conclusion false. Or take this counterexample, from Stalnaker:*

> *If J. Edgar Hoover had been born a Russian, then he would have been a Communist.*
> *If he had been a Communist, he would have been a traitor.*
> ∴ *If he had been born a Russian, he would have been a traitor.*

In general, transitivity fails in the situation shown in Figure 4(A). The antecedent of the first premise must be more far-fetched than the antecedent of the second, which is the consequent of the first. Then the closest worlds where the first antecedent holds are different from—and may differ in character from—the closest worlds where the second antecedent holds. That is the situation in our examples. We must go farther from actuality to find worlds where Otto went than to find worlds where Anna went. A Communist Hoover is nowhere to be found at worlds near ours, but a Russian-born Hoover is still more remote.

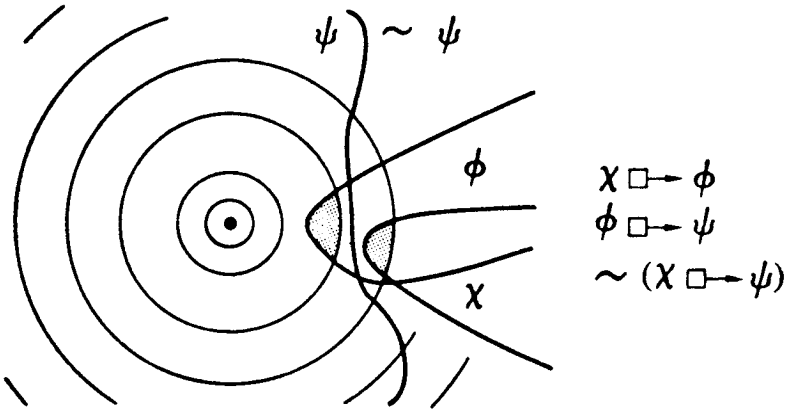
In these and all other counterexamples to transivity, the 'might' counterfactual $\phi \Diamond\!\!\rightarrow \sim\chi$ is true. In these examples, but not in all, we can say something stronger: the 'would' counterfactual $\phi \Box\!\!\rightarrow \sim\chi$ is non-vacuously true. If Anna had gone, Otto would still not have; if Hoover had been a Communist, he would still not have been born a Russian.‡ By adding a third premise to the inference by transitivity, we may rule out all cases where transitivity fails. The inference-pattern

$$\begin{array}{l} \sim(\phi \Diamond\!\!\rightarrow \sim\chi) \\ \chi \Box\!\!\rightarrow \phi \\ \phi \Box\!\!\rightarrow \psi \\ \hline \therefore\ \chi \Box\!\!\rightarrow \psi \end{array} \qquad \text{or, more simply,} \qquad \begin{array}{l} \phi \Box\!\!\rightarrow \chi \\ \chi \Box\!\!\rightarrow \phi \\ \phi \Box\!\!\rightarrow \psi \\ \hline \therefore\ \chi \Box\!\!\rightarrow \psi \end{array}$$

* 'A Theory of Conditionals'.

‡ 'Still', 'even so', etc. in the consequent, or 'even' before the antecedent, mark a presupposition that the consequent fails to contrast with something. In the cases above, it is true and so fails to contrast with the actual state of affairs; in other cases, it is false but fails to contrast with the consequent of some other counterfactual. Insofar as it is misleading to omit these contrast-marking devices when they would be appropriate, perhaps we may say that the unmarked counterfactual carries a weak presupposition that the consequent *does* contrast with something. I treat such presuppositions about the consequent, as I do the presupposed falsity of the antecedent, as matters of conversational implicature irrelevant to truth conditions.

(A) FAILURE OF TRANSITIVITY



(B) FAILURE OF CONTRAPOSITION

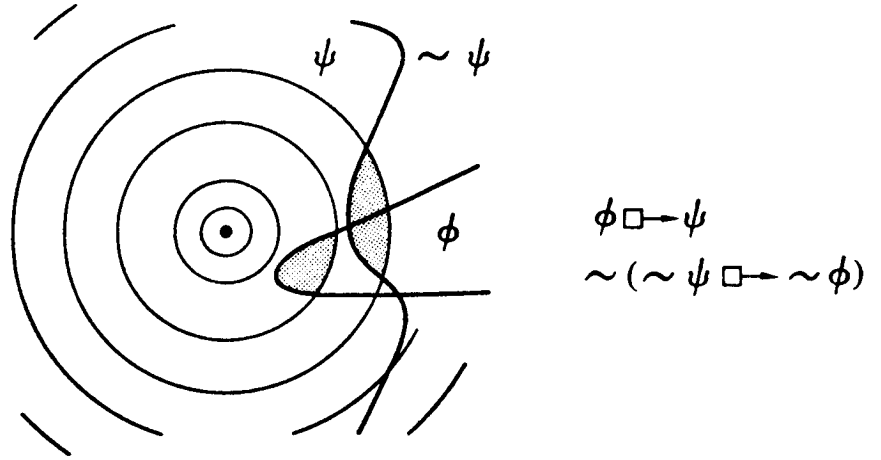


FIGURE 4

is perfectly valid. Using this valid inference-pattern, and the fact that the requisite third premise of the form $\chi \ \& \ \phi \ \Box\!\!\rightarrow \chi$ must be true, we can justify a valid special case of inference by transitivity:

$$\frac{\begin{array}{l}\chi \ \Box\!\!\rightarrow \chi \ \& \ \phi \\ \chi \ \& \ \phi \ \Box\!\!\rightarrow \psi\end{array}}{\therefore \ \chi \ \Box\!\!\rightarrow \psi}.$$

Since the inference from $\chi \ \Box\!\!\rightarrow \phi$ to $\chi \ \Box\!\!\rightarrow \chi \ \& \ \phi$ is valid, we can simplify this inference-pattern to:

$$\frac{\begin{array}{l}\chi \ \Box\!\!\rightarrow \phi \\ \chi \ \& \ \phi \ \Box\!\!\rightarrow \psi\end{array}}{\therefore \ \chi \ \Box\!\!\rightarrow \psi}.$$

Another valid inference-pattern resembling the fallacious inference by transitivity may be called inference by weakening the consequent. It is like transitivity except that the second conditional premise is a strict conditional instead of the corresponding counterfactual.

$$\frac{\begin{array}{l}\chi \ \Box\!\!\rightarrow \phi \\ \Box(\phi \supset \psi)\end{array}}{\therefore \ \chi \ \Box\!\!\rightarrow \psi}.$$

The *fallacy of contraposition* is either one of the two invalid inference-patterns

$$\frac{\phi \ \Box\!\!\rightarrow \psi}{\therefore \ \sim\psi \ \Box\!\!\rightarrow \sim\phi} \quad \text{and} \quad \frac{\sim\psi \ \Box\!\!\rightarrow \sim\phi}{\therefore \ \phi \ \Box\!\!\rightarrow \psi}.$$

Obviously, both or neither are valid; let us concentrate on the first. For instance, consider this argument.

If Boris had gone to the party, Olga would still have gone.
∴ If Olga had not gone, Boris would still not have gone.

Suppose that Boris wanted to go, but stayed away solely in order to avoid Olga, so the conclusion is false; but Olga would have gone all the more willingly if Boris had been there, so the premise is true. In general, $\phi \ \Box\!\!\rightarrow \psi$ may be true and its contrapositive false in the situation shown in Figure 4(B).

Note that we could use contraposition to justify the following inference-pattern involving 'might' counterfactuals:

$$\frac{\phi \ \Diamond\!\!\rightarrow \psi}{\therefore \ \psi \ \Diamond\!\!\rightarrow \phi}.$$

But this inference has no plausibility at all. Note also that although

contraposition of counterfactuals is invalid, nevertheless inference by *modus tollens* on a counterfactual is valid:

$$\frac{\phi \mathbin{\Box\!\!\rightarrow} \psi \quad \sim\psi}{\therefore\ \sim\phi}.$$

We cannot regard this *modus tollens* as proceeding by contraposition followed by *modus ponens*, as we can in the case of *modus tollens* on a material conditional; rather we should think of it as an inference from the counterfactual to the material conditional $\phi \supset \psi$, followed by contraposition of the material conditional, followed by *modus ponens* on the contraposed material conditional.

1.9 Potentialities

We might have occasion to complain that if the winner had not bribed the judge, then he would not have won. By this, we do not mean that the closest worlds to ours where '*The winner did not bribe the judge*' is true are worlds where '*The winner did not win*' is true. Our complaint might be true, but that construal of it certainly is false. The supposition that someone managed to win without bribing the judge—far-fetched though it might be—is entertainable; but there are no worlds at all, neither the closest worlds where that supposition holds nor any others, where anyone wins without winning. (One can win officially without 'really' winning, but that is equivocation—stick to official winning.) Our complaint therefore is not

The winner did not bribe the judge $\mathbin{\Box\!\!\rightarrow}$ *the winner did not win.*

Rather, it is *de re* with respect to 'the winner'. We are ascribing to whoever actually was the winner a counterfactual property, or *potentiality*, expressed by the formula

x did not bribe the judge $\mathbin{\Box\!\!\rightarrow}$ *x did not win.*

We are talking about what would have befallen the actual winner, not about what would have befallen whoever would have been the winner. Our supposition is that *he*—the actual winner—did not bribe the judge. It matters not that the description 'winner' we used to specify him as the one we were making suppositions about would not have fitted him (in our opinion) if the supposition had been true. The right way to symbolize what we want to say would be something like this:

The winner is an x such that
(*x did not bribe the judge* $\mathbin{\Box\!\!\rightarrow}$ *x did not win*).

This symbolization reveals that the counterfactual formula and its subformulas '*x did not bribe the judge*' and '*x did not win*' are what enter into the interpretation of the sentence, whereas the seeming antecedent and consequent '*The winner did not bribe the judge*' and '*The winner did not win*' are foisted upon us by a superficial illusion of grammar.

We could treat '*is an x such that . . .*' as a structureless abstraction operator that attaches to a formula ϕ_x to make a predicate that can combine in turn with a denoting term α to make a subject-predicate sentence. (Or to make another formula, in case there are free variables besides the free x in ϕ_x.) Alternatively, we could treat '____ *is an x such that . . .*' as a quantificational matrix with two gaps suitable to receive α and ϕ_x respectively:

$$\exists x(x = \_\_\_\_ \ \& \ \ldots).$$

Either way, a sentence 'α *is an x such that* ϕ_x' is true at a world i if and only if whatever is denoted at i by α satisfies ϕ_x at i (as a value of the variable x)—in other words, has the property expressed by the formula ϕ_x.*

Counterfactuals *de re* crop up in connection with countercomparatives. What if my yacht were longer than it is? The supposition is notoriously not that the seeming antecedent '*My yacht is longer than it is*' is true (in any straightforward way; but see Section 2.8). One way to handle counterfactuals with this seeming antecedent is as *de re* predications of a counterfactual potentiality to whatever is the actual length of my yacht.

The length of my yacht is an x such that
(*the length of my yacht exceeds x* $\boxrightarrow$ . . .)

Of course, we need an entity to have the potentiality being ascribed. It is easy enough to hypostatize lengths, but what do we make of '*If mankind were wiser . . .*'?

Potentialities expressed by counterfactual formulas are needed not only in the *de re* cases, but also for universal or existential quantifications (including the existential quantifications that arise if we take '____ *is an x such that . . .*' as a quantificational matrix). We want a way to say that any winner who would not have won if he had not bribed the judge is a knave:

$$\forall x((Wx \ \& \ (\sim Bx \boxrightarrow \sim Wx)) \supset Kx),$$

* On the use of abstraction operators to give a clear and unambiguous symbolization of modal predications *de re*, see Richmond Thomason and Robert Stalnaker, 'Modality and Reference', *Noûs* 2 (1968): 359–372.

using the obvious abbreviations. Or that there was at least one Roman emperor who, if only he had had gunpowder, would have conquered all of Europe:

$$\exists x(Rx \ \& \ (Gx \ \Box\!\!\rightarrow \ Cx)).$$

Or that a thing has disposition D if and only if, subjected to test T, it would give response R:

$$\forall x(Dx \equiv (Tx \ \Box\!\!\rightarrow \ Rx)).$$

(The point is not that I want to *believe* instances of the last—I am inclined not to—but that I want my theory of counterfactuals to explain what they mean.) There is nothing peculiar about the use of quantifiers in these sentences. To interpret them, and to interpret also our counterfactuals *de re*, all we need is an account of satisfaction by things of counterfactual formulas. Or, in the material mode: of possession by things of counterfactual potentialities.

It is enough to ask for conditions under which a single thing satisfies a formula $\phi_x \ \Box\!\!\rightarrow \ \psi_x$ with x as the only free variable. The generalization to satisfaction of formulas with arbitrarily many free variables by arbitrarily long or infinite sequences of things is routine, once we know what to do in the one-variable case. It is not required that x appear both in ϕ_x and in ψ_x. Often it does not:

The length of my yacht exceeds x $\Box\!\!\rightarrow$ *I am contented*
I am ostentatious $\Box\!\!\rightarrow$ *the length of my yacht exceeds x*

for instance, might be used in symbolizations of '*If my yacht were longer I would be contented*' and '*If I were ostentatious my yacht would be longer*'. In fact, I do not even exclude the degenerate case that x appears neither in ϕ_x nor in ψ_x.

As a first try, we could give satisfaction conditions for counterfactual formulas simply by imitating the truth conditions for counterfactual sentences, letting the thing that satisfies the formula tag along throughout. Something satisfies $\phi_x \ \Box\!\!\rightarrow \ \psi_x$ at a world i, on this proposal, if and only if either (1) no world where it satisfies ϕ_x belongs to any sphere around i (the vacuous case), or (2) some sphere S around i does contain at least one world where it satisfies ϕ_x, and at every world in S where it satisfies ϕ_x it also satisfies ψ_x. So, for example, if Ripov is the winner because he bribed the judge (here at our world), then he has the potentiality expressed by

x did not bribe the judge $\Box\!\!\rightarrow$ *x did not win*

non-vacuously if and only if there is some sphere containing worlds where he did not bribe the judge, throughout which all the worlds

where he did not bribe the judge are worlds where he did not win. Roughly: if and only if he did not win at the closest of the worlds where he did not bribe the judge.

The trouble is that this presumes that we have the very same Ripov active at several worlds: ours, where he bribes the judge and wins, and others, where he does not bribe the judge and does not win. What makes the inhabitant of another world, who does not bribe and does not win, identical with our Ripov? I suppose the answer must be *either* that his identity with our Ripov is an irreducible fact, not to be explained in terms of anything else, *or* that his identity with our Ripov is due to some sort of similarity to our Ripov—he is Ripov because he plays much the same role at the other world that our Ripov plays here. Neither answer pleases me. The first answer either posits trans-world identities between things arbitrarily different in character, thereby denying what I take to be some of the facts about *de re* modality, or else it makes a mystery of those facts by denying us any way to explain why there are some sorts of trans-world identities but not others. The second answer at least is not defeatist, but it runs into trouble because similarity relations lack the formal properties—transitivity, for instance—of identity.

The best thing to do, I think, is to escape the problems of trans-world identity by insisting that there is nothing that inhabits more than one world. There are some abstract entities, for instance numbers or properties, that inhabit no particular world but exist alike from the standpoint of all worlds, just as they have no location in time and space but exist alike from the standpoint of all times and places. Things that do inhabit worlds—people, flames, buildings, puddles, concrete particulars generally—inhabit one world each, no more. Our Ripov is a man of our world, who does not reappear elsewhere. Other worlds may have Ripovs of their own, but none of these is our Ripov. Rather, they are counterparts of our Ripov. What comes from trans-world resemblance is not trans-world identity, but a substitute for trans-world identity: the counterpart relation. What our Ripov cannot do in person at other worlds, not being present there to do it, he may do vicariously through his counterparts. He himself is not an honest man at any world —he is dishonest here, and nonexistent elsewhere—but he is vicariously honest through his honest counterparts.

In general: something has for *counterparts* at a given world those things existing there that resemble it closely enough in important respects of intrinsic quality and extrinsic relations, and that resemble it no less closely than do other things existing there. Ordinarily something will have one counterpart or none at a world, but ties in similarity may give it multiple counterparts. Two special cases: (1) anything is its own unique counterpart at its own world, and (2) the abstract

entities that exist alike from the standpoint of all worlds, but inhabit none, are their own unique counterparts at all worlds.

I have proposed elsewhere that the counterpart relation ought to be used as a substitute for trans-world identity in explaining *de re* modality.* The realm of essence and accident is the realm of the vicarious. What something *might* have done (or might have been) is what it does (or is) vicariously; and that is what its counterparts do (or are). What is essential to something is what it has in common with all its counterparts; what it nowhere vicariously lacks. Ripov's honest counterparts make him someone who might have been honest. His lack of inanimate counterparts makes him essentially animate. In terms of satisfaction of modal formulas: something satisfies $\Box\phi_x$ at a world i if and only if any counterpart of it at any world j accessible from i satisfies ϕ_x at j; something satisfies $\Diamond\phi_x$ at a world i if and only if it has some counterpart at some world j accessible from i that satisfies ϕ_x at j. Alternatively, we can say that something *vicariously satisfies* ϕ_x at a world i if and only if it has some counterpart at i that satisfies ϕ_x at i. (At one's own world, vicarious satisfaction coincides with satisfaction.) Then we can restate the conditions in terms of vicarious satisfaction. Something satisfies $\Box\phi_x$ at i if and only if there is no world accessible from i where it vicariously satisfies $\sim\phi_x$. Something satisfies $\Diamond\phi_x$ if and only if there is some world accessible from i where it vicariously satisfies ϕ_x.‡

The method of counterparts seems to me to have many virtues as a theory of *de re* modality. (1) It has the same explanatory power as a theory of *de re* modality that employs trans-world identity based on trans-world resemblance. The facts about what things might have been and might have done are explained by our standards of similarity—that is, of the comparative importances of respects of comparison—plus facts about how things actually are. Modal facts are grounded in facts about actual character, not mysteriously independent. It is because of the way Ripov actually is that certain honest men at other worlds resemble him enough to be his counterparts, and inanimate things at other worlds do not. (2) However, we are rid of the worst burden of a theory of trans-world identity based on trans-world resemblance: the counterpart relation is not identity, so we need not try to force it to

* 'Counterpart Theory and Quantified Modal Logic', *Journal of Philosophy* **65** (1968): 113–126; 'Counterparts of Persons and Their Bodies', *Journal of Philosophy* **68** (1971): 203–211.

‡ An alternative definition of vicarious satisfaction would put the double negation in the satisfaction conditions for $\Diamond\phi_x$ instead of those for $\Box\phi_x$; but we would be stuck with it one place or other. The reason is that something with more or less than one counterpart at a world may vicariously satisfy both or neither of ϕ_x and $\sim\phi_x$, so vicariously satisfying $\sim\phi_x$ is not the same as not vicariously satisfying ϕ_x.

have the logical properties of identity. (3) Therefore we have a desirable flexibility. For instance, we can say that something might have been twins because it has twin counterparts somewhere, without claiming that it is literally identical with two things not identical to one another. (4) Since the counterpart relation is based on similarity, the vagueness of similarity infects *de re* modality. And that is all to the good. It goes a long way toward explaining why questions of *de re* modality are as difficult as we have found them to be. (5) We can plead this same vagueness to explain away seeming discrepancies among our *de re* modal opinions. For instance, consider two inhabitants of a certain world that is exactly like ours in every detail until 1888, and thereafter diverges. One has exactly the ancestral origins of our Hitler; that is so in virtue of events within the region of perfect match that ended just before his birth. In that region, it is quite unequivocal what is the counterpart of what. The other has quite different ancestral origins, but as he grows up he gradually duplicates more and more of the infamous deeds of our Hitler until after 1930 his career matches our Hitler's career in every detail. Meanwhile the first lives an obscure and blameless life. Does this world prove that Hitler might have lived a blameless life? Or does it prove that he might have had different ancestral origins? I want to be able to say either—though perhaps not both in the same breath—depending on which respects of comparison are foremost in my mind; and the method of counterparts, with due allowance for vagueness, allows me to do so. (6) There are also cases where we need to mix different counterpart relations in a single sentence in order to make sense of it as a reasonable thing to think; for instance, sentences of *de re* contingent identity. We shall see other cases in connection with counterfactuals. I see no way to get the same effect by means of trans-world identity alone, though one might get it by mixing in trans-world identity along with the counterpart relations.

Now I shall use the method of counterparts to correct my previous satisfaction conditions for counterfactual formulas. The formulation I gave will not do at all. Without the trans-world identities I reject, it leads in most cases to vacuity. We need to replace trans-world identity by the counterpart relation; that is, to replace satisfaction (in the definiens) by vicarious satisfaction. Roughly speaking, I want to say that Ripov has the potentiality expressed by

$$x \textit{ reforms} \mathrel{\Box\!\!\rightarrow} x \textit{ confesses}$$

—that he satisfies that formula—because the closest worlds where he vicariously reforms are worlds where he vicariously confesses. But that is not quite right, even if we forget to doubt the Limit Assumption. What if he has multiple counterparts at one of the closest worlds

where he vicariously reforms? It is not enough if one reforms and another confesses; it is not even enough if one reforms and confesses, and another reforms without confessing. What we must require is that at every closest world where one of Ripov's counterparts reforms, all of those who reform also confess—that is, none reforms without confessing. The closest worlds where he vicariously reforms must be worlds where he does not vicariously both reform and not confess. (Distinguish between (1) vicariously both reforming and not confessing, both through the same counterpart, and (2) both vicariously reforming and vicariously not confessing, perhaps through different counterparts.)

In general: something satisfies $\phi_x \mathbin{\Box\!\!\rightarrow} \psi_x$ at a world i if and only if either (1) no world where it vicariously satisfies ϕ_x belongs to any sphere around i (the vacuous case), or (2) some sphere S around i does contain at least one world where it vicariously satisfies ϕ_x, and at no world in that sphere does it vicariously satisfy ϕ_x & $\sim\psi_x$. Putting it in terms of the counterpart relation: something satisfies $\phi_x \mathbin{\Box\!\!\rightarrow} \psi_x$ at a world i if and only if either (1) at no world j in any sphere around i does it have a counterpart that satisfies ϕ_x at j, or (2) some sphere S around i does contain at least one world j such that some counterpart of it at j satisfies ϕ_x at j, and every counterpart of it at any world k in S that satisfies ϕ_x at k also satisfies ψ_x at k.

The method of counterparts is needed not only to give an account of satisfaction of counterfactual formulas by things that inhabit our world alone, but also to interpret counterfactuals containing 'I', 'you', or demonstratives. These denote on any occasion of utterance such things as the speaker, his audience, the things he points to; and these are things that inhabit only the world of the utterance. So if I say '*If I had given you that, you would have broken it*', what are denoted by 'I', 'you', and 'that' are three things confined to our world. The closest antecedent-worlds are not worlds where those things reappear, suitably related—that way lies vacuity—but worlds where those things have suitably related counterparts. The counterfactual is true, roughly, if and only if the closest worlds where there is a triple $\langle a, b, c\rangle$ of counterparts of I, you, and that, respectively, such that a gives c to b, are worlds where there is no such triple in which b does not break c. We have two options. We could give special truth conditions for counterfactuals with 'I', 'you', or demonstratives, along the lines I have sketched; or we could use the satisfaction conditions just laid down for counterfactual formulas by insisting that all occurrences of 'I', 'you', or demonstratives should be taken as *de re*.

It would be a good idea to provide for more than one counterpart relation. Different counterpart relations might vary in the stringency

of resemblance they require; or they might stress different respects of comparison.

We can explain the simultaneous truth of Goodman's sentences

(1) *If New York City were in Georgia, New York City would be in the South.*

and

(2) *If Georgia included New York City, Georgia would not be entirely in the South.*

by the hypothesis that both are *de re* both with respect to 'New York City' and with respect to 'Georgia', and that a less stringent counterpart relation is summoned up by the subject terms 'New York City' in (1) and 'Georgia' in (2) than by the object terms 'Georgia' in (1) and 'New York City' in (2). Then in (1) we are concerned with the closest worlds to ours where a not-too-close counterpart of our New York is in a close counterpart of our Georgia, and hence is in (a counterpart of?) the South; whereas in (2) we are concerned with the closest worlds to ours where a not-too-close counterpart of our Georgia includes a close counterpart of our New York City, and hence is not entirely included in the South.*

For a familiar illustration of the need for counterpart relations stressing different respects of comparison, take '*If I were you . . .*'. The antecedent-worlds are worlds where you and I are vicariously identical; that is, we share a common counterpart. But we want him to be in *your* predicament with *my* ideas, not the other way around. He should be your counterpart under a counterpart relation that stresses similarity of predicament; mine under a different counterpart relation that stresses similarity of ideas.

* Alternatively, perhaps each is *de re* with respect to one of the two names—perhaps the subject in both, perhaps the object in both—and we are seeing a difference in stringency between a counterpart relation involved in satisfaction of counterfactual formulas and a counterpart relation involved in determining the denotation at other worlds of a proper name originally bestowed by an episode of naming that involved some inhabitant of our world.