Philosophical Papers

Volume II

DAVID LEWIS

Oxford University Press

Oxford New York Toronto
Delhi Bombay Calcutta Madras Karachi
Petaling Jaya Singapore Hong Kong Tokyo
Nairobi Dar es Salaam Cape Town
Melbourne Auckland

and associated companies in
Beirut Berlin Ibadan Nicosia

Published by Oxford University Press Inc
200 Madison Avenue New York New York 10016

Oxford is a registered trademark of Oxford University Press

Library of Congress Cataloging in Publication Data
(Revised for vol 2)

Lewis David K
Philosophical papers

Includes bibliographies and indexes

1 Philosophy I Title
B29 L49 1983 191 82–22551
ISBN 0–19–503203–9 (v 1)
ISBN 0–19–503204 7 (pbk v 1)
ISBN 0–19–503645–X (v 2)
ISBN 0–19–503646–8 (pbk v 2)

2 4 6 8 9 7 5 3 1

Printed in the United States of America

Acknowledgments

I thank the editors and publishers who have kindly granted permission to reprint eleven of the essays that appears in this volume The places of first publication are as follows

"Counterfactuals and Comparative Possibility," *Journal of Philosophical Logic* 2 (1973) 418–46

"Counterfactual Dependence and Time's Arrow," *Noûs* 13 (1979) 455–76

"The Paradoxes of Time Travel," *American Philosophical Quarterly* 13 (1976) 145–52

"A Subjectivist's Guide to Objective Chance," in Richard C Jeffrey, ed , *Studies in Inductive Logic and Probability*, Volume II (University of California Press, 1980)

"Probabilities of Conditionals and Conditional Probabilities," *Philosophical Review* 85 (1976) 297–315

"Causation," *Journal of Philosophy* 70 (1973) 556–67, the postscripts incorporate parts of those that are translated as "Kausalitat Epilog 1978" in Gunter Posch, ed , *Kausalitat—Neue Texte* (Philip Reclam, 1981)

"Veridical Hallucination and Prosthetic Vision," *Australasian Journal of Philosophy* 58 (1980) 239–49

"Are We Free to Break the Laws?" *Theoria* 47 (1981) 113–21

"Prisoners' Dilemma is a Newcomb Problem," *Philosophy and Public Affairs* 8 (1979) 235–40

"Causal Decision Theory," *Australasian Journal of Philosophy* 59 (1981) 5–30

"Utilitarianism and Truthfulness," *Australasian Journal of Philosophy* 50 (1972) 17–19

Most of the numerous abstracts that appear in the bibliography to this volume were previously published either in *The Philosopher's Index* or in *The Review of Metaphysics*, I thank the editors thereof for their kind permission to reprint those abstracts here

Contents

Introduction

Eleven of the papers in this volume were originally published from 1972 to 1981, misprints apart, they are reprinted in their original form In some cases, where retractions or additions seemed urgently needed, I have appended postscripts Two other papers appear here for the first time The papers in this volume deal with topics concerning counterfactuals, causation, and related matters Papers in ontology, philosophy of mind, and philosophy of language have appeared in Volume I I have left out papers which are rejoinders, or which are of primarily technical interest, or which overlap too much with the papers I have included Abstracts of the omitted papers may be found here, in the bibliography of my writings

Many of the papers, here and in Volume I, seem to me in hindsight to fall into place within a prolonged campaign on behalf of the thesis I call "Humean supervenience" Explicit discussion of that thesis appears only in "A Subjectivist's Guide to Objective Chance", but it motivates much of the book

Humean supervenience is named in honor of the greater denier of necessary connections It is the doctrine that all there is to the world is a vast mosaic of local matters of particular fact, just one little thing and then another (But it is no part of the thesis that these local matters are mental) We have geometry a system of external relations of spatio-

temporal distance between points Maybe points of spacetime itself, maybe point-sized bits of matter or aether or fields, maybe both And at those points we have local qualities perfectly natural intrinsic properties which need nothing bigger than a point at which to be instantiated [1] For short we have an arrangement of qualities And that is all There is no difference without difference in the arrangement of qualities All else supervenes on that

First say it, then qualify it I don't really mean to say that no two possible worlds whatsoever differ in any way without differing in their arrangements of qualities For I condede that Humean supervenience is at best a contingent truth Two worlds might indeed differ only in unHumean ways, if one or both of them is a world where Humean supervenience fails Perhaps there might be extra, irreducible external relations, besides the spatiotemporal ones, there might be emergent natural properties of more-than-point-sized things, there might be things that endure identically through time or space, and trace out loci that cut across all lines of qualitative continuity It is not, alas, unintelligible that there might be suchlike rubbish Some worlds have it And when they do, it can make differences between worlds even if they match perfectly in their arrangements of qualities

But if there is suchlike rubbish, say I, then there would have to be extra natural properties or relations that are altogether alien to this world Within the inner sphere of possibility, from which these alien intrusions are absent, there is indeed no difference of worlds without a difference in their arrangements of qualities [2]

Is this materialism?—no and yes I take it that materialism is metaphysics built to endorse the truth and descriptive completeness of physics more or less as we know it, and it just might be that Humean supervenience is true, but our best physics is dead wrong in its inventory of the qualities Maybe, but I doubt it Most likely, if Humean

[1] For ways to explain what makes a property natural and intrinsic, see my New Work for a Theory of Universals, *Australasian Journal of Philosophy* 61 (1983) 343–77 However, I ought to add that besides the candidates considered there, class nominalism with primitive naturalness or a sparse theory of immanent universals, there is a third strong contender a theory of tropes like that of Donald C Williams, On the Elements of Being, *Review of Metaphysics* 7 (1953) 3–18 and 171–92, but with the tropes cut to a minimum, so that the special status of natural properties is built into the ontology itself

[2] On contingent supervenience theses, see the discussion of materialism in New Work for a Theory of Universals On inner and outer spheres of possibility, see Brian Skyrms, Tractarian Nominalism, *Philosophical Studies* 40 (1981) 199–206, and D M Armstrong, Metaphysics and Supervenience, *Critica* 14 (1982) 3–17

supervenience is true at all, it is true in more or less the way that present physics would suggest

I have conceded that Humean supervenience is a contingent, therefore an empirical, issue Then why should I, as philosopher rather than physics fan, care about it? Isn't my professional business more with the whole expanse of logical space than with the question which of its districts happens to be ours?—Fair enough Really, what I uphold is not so much the truth of Humean supervenience as the *tenability* of it If physics itself were to teach me that it is false, I wouldn't grieve

That might happen maybe the lesson of Bell's theorem is exactly that there are *physical* entities which are unlocalized, and which might therefore make a difference between worlds—worlds in the inner sphere—that match perfectly in their arrangements of local qualities Maybe so I'm ready to believe it But I am not ready to take lessons in ontology from quantum physics as it now is First I must see how it looks when it is purified of instrumentalist frivolity, and dares to say something not just about pointer readings but about the constitution of the world, and when it is purified of doublethinking deviant logic, and—most of all—when it is purified of supernatural tales about the power of the observant mind to make things jump If, after all that, it still teaches nonlocality, I shall submit willingly to the best of authority

What I want to fight are *philosophical* arguments against Humean supervenience When philosophers claim that one or another commonplace feature of the world cannot supervene on the arrangement of qualities, I make it my business to resist Being a commonsensical fellow (except where unactualized possible worlds are concerned) I will seldom deny that the features in question exist I grant their existence, and do my best to show how they can, after all, supervene on the arrangement of qualities The plan of battle is as follows

First, laws of nature Few would deny that laws of nature, whatever else they may be, are at least exceptionless regularities Not all regularities are laws, of course But, following the lead of (a short temporal segment of) Ramsey, I suggest that the laws are the ones that buy into those systems of truths that achieve an unexcelled combination of simplicity and strength That serves the Humean cause For what it is to be simple and strong is safely noncontingent, and what regularities there are, or more generally what candidate systems of truths, seems to supervene safely on the arrangement of qualities I stated such a theory of lawhood in my book *Counterfactuals*,[3] and here I discuss it further

[3] (Oxford Blackwell, 1973)

in Postscript C to "A Subjectivist's Guide to Objective Chance."

I am prepared at this point to take the offensive against alleged unHumean lawmakers. I say there is no point believing in them, because they would be unfit for their work. Here I have in mind the theory that laws are made by a lawmaking second-order relation of universals, a theory most fully presented by D. M. Armstrong in *What is a Law of Nature?*[4] Let N be the supposed lawmaker relation, the idea, in its simplest form, is that it is a contingent matter, and one not supervenient on the arrangement of qualities, which universals stand in the relation N, but it is somehow necessary that if $N(F, G)$, then we have the regularity that all F's are G's. I ask: how can the alleged lawmaker impose a regularity? Why can't we have $N(F, G)$, and still have F's that are not G's? What prevents it? Don't try *defining* N in terms of there being a law and hence a regularity—we're trying to *explain* lawhood. And it's no good just giving the lawmaker a name that presupposes that somehow it does its stuff, as when Armstrong calls it "necessitation." If you find it hard to ask why there can't be F's that are not G's when F "necessitates" G, you should ask instead how any N can do what it must do to deserve that name.

Next, counterfactuals. I take them to be governed by similarity of worlds, according to the analysis given in "Counterfactuals and Comparative Possibility," in this volume. To the extent that this similarity consists of perfect match in matters of particular fact, it supervenes easily on the arrangement of qualities, and to the extent that it consists of (perfect or imperfect) conformity by one world to the laws of the other, it supervenes if the laws do. In "Counterfactual Dependence and Time's Arrow," I argue that one important sort of counterfactual, at least, will work properly if it is governed by just these respects of similarity.

Next, causation. In "Causation" and its postscripts, I defend an analysis of causation in terms of counterfactual dependence between events. The counterfactuals are discussed here in the two papers just mentioned, and since counterfactual dependence only seems causal when it is between events, my treatment of causation requires "Events" before it is done. Causation draws the arrow from past to future, that arrow exists only as an asymmetric pattern in the arrangement of qualities, so causal counterfactuals must somehow be sensitive

[4] (Cambridge: Cambridge University Press, 1983). See also Fred I. Dretske, Laws of Nature, *Philosophy of Science* 44 (1977): 248–68; and Michael Tooley, The Nature of Laws, *Canadian Journal of Philosophy* 4 (1977): 667–98.

to the asymmetry In "Counterfactual Dependence and Time's Arrow" I offer an account of that sensitivity Given causation, or rather causal dependence, we can proceed to causal analyses of various things, for instance seeing, in "Veridical Hallucination and Prosthetic Vision," or what else you can do if you can freely raise your hand, in "Are We Free To Break the Laws?"

Next, persistence through time I take the view that nothing endures identically through time (Except universals, if such there be, their loci would coincide with relations of qualitative match, would indeed constitute these relations, so they would commit no violations of Humean supervenience) Persisting particulars consist of temporal parts, united by various kinds of continuity To the extent that the continuity is spatiotemporal and qualitative, of course it supervenes on the arrangement of qualities But the continuity that often matters most is causal continuity the thing stays more or less the same because of the way its later temporal parts depend causally for their existence and character on the ones just before So the spatiotemporal boundaries of persisting things, for instance people, can supervene on the arrangement of qualities, provided that causation does I discuss lines of causal continuity, not ruling out zigzag or broken lines, in "The Paradoxes of Time Travel " In "Survival and Identity," in Volume I of these *Papers*, I reply to some paradoxes brought against the idea that our survival is a matter of continuities that unite our temporal parts [5]

Next, mind and language Several papers in the previous volume concern the thesis that mental states, indexed with content when appropriate, are definable as the occupants of causal roles Some of these states are people's beliefs, and some of their beliefs are their

[5] It is at this point that Humean supervenience has come under direct attack Saul Kripke, in Identity through Time, given at the 1979 conference of the American Philosophical Association, Eastern Division, has argued that if a disk is made of homogeneous matter, then whether the disk is spinning or not is a feature of the world that does not supervene on the arrangement of qualities We might have two worlds, just alike in their arrangements of qualities, one with a spinning disk and one with a stationary disk (My Humean supervenience corresponds roughly to the attenuated holographic hypothesis, which was one of Kripke's targets) Whether the disk spins is, of course, definable in terms of the persistence of its parts through time, so in the first instance it is persistence that fails to supervene But that might be because causation fails to supervene, and persistence requires causal continuity

I reply by conceding, as I have, that Humean supervenience is contingent The two worlds with their differing disks must (one or both) be worlds where there is something extra to make the difference That does not show that any feature of *this* world fails to supervene on the arrangement of qualities (Here I am indebted to Mark Johnston)

expectations about other people In *Convention*,[6] and in "Languages and Language" and "Radical Interpretation" in Volume I, I suggested how semantic facts could obtain in virtue of the mutual expectations that prevail in a linguistic community

And so it goes There is room for endless argument over the details, but I remain confident that at every step mentioned the connection is something like what I have said—enough like it, anyway, to allow the cumulative Humean supervenience of one thing after another At every step mentioned—but there is one that I passed over

There is one big bad bug chance It is here, and here alone, that I fear defeat But if I'm beaten here, then the entire campaign goes kaput For chances enter at the very beginning A law, I said with Ramsey, is a regularity that enters into the best systems But what sort of systems? If there are chances—single-case objective probabilities, for instance, that a certain atom will decay this week—then some regularities have to do with chances, and the best true systems will be those that do best, *inter alia*, at systematizing the truth about chances So bestness of true systems, and hence lawhood, and hence counterfactuals and causation and occupancy of causal roles and all the rest, will not supervene just on the actual arrangement of qualities, but on that plus all the chances there are, at various times, of that arrangement continuing in one way or another Therefore the only hope for Humean supervenience is that the chances themselves might somehow supervene on the arrangement of qualities

How could they? It is easy to go partway The chances for alternative futures that obtain at a moment surely depend on just how things actually are at that moment We might as well throw in the way things are at all previous times, that might help, and it's no harm including too much So far, so good We have a conditional if history is so-and-so then the chances are such-and-such And the antecedent of that conditional—history up to the moment in question—surely does supervene on the arrangement of qualities

But what is the status of the history-to-chance conditional itself? Is it necessary or contingent? If contingent, does it supervene or not on the arrangement of qualities?

If history-to-chance conditionals are necessary truths, no worries Then the chances at any moment supervene on the arrangement of qualities, in fact on just the part of it up to that moment Sometimes, we can see how the conditional might be necessary suppose it says

[6] (Cambridge, Massachusetts Harvard University Press, 1968)

that when we have prominent symmetry in the present set-up and its alternative futures, then those futures have equal chances But sometimes not How can an equality of chances based on symmetries, or any such necessary principle, give us the connections we need between, say, the exact height of a potential barrier and the exact chance of tunnelling through it? I hope there is a way, given the trouble I'm in if not, but I can't see what it is

If the conditionals are contingent, but themselves supervene on the arrangement of qualities, then again no worries That would be so if they hold in virtue of relevant actual frequencies throughout the world, for instance Or they could supervene in some fancier way, for instance by means of the "package deal" for chances and laws that I consider in Postscript C to "A Subjectivist's Guide to Objective Chance " Alas, I fear it cannot be so The trouble is that whatever pattern it is in the arrangement of qualities that makes the conditionals true will itself be something that has some chance of coming about, and some chance of not coming about What happens if there is some chance of getting a pattern that would undermine that very chance? The Principal Principle of "A Subjectivist's Guide to Objective Chance" affords a way of turning this vague worry into a proper argument, hence the dismal ending to that paper

Why not give in? I could admit that the history-to-chance conditionals, and so the chances themselves, are contingent and do not supervene on the arrangement of qualities I could insist for consolation that at any rate all else supervenes on the arrangement of qualities and the chances together Why not? I am not moved just by loyalty to my previous opinions That answer works no better than the others Here again the unHumean candidate for the job turns out to be unfit for its work The distinctive thing about chances is their place in the 'Principal Principle,' which compellingly demands that we conform our credences about outcomes to our credences about their chances Roughly, he who is certain the coin is fair must give equal credence to heads and tails, being less rough is the main business of "A Subjectivist's Guide to Objective Chance " I can see, dimly, how it might be rational to conform my credences about outcomes to my credences about history, symmetries, and frequencies I haven't the faintest notion how it might be rational to conform my credences about outcomes to my credences about some mysterious unHumean magnitude Don't try to take the mystery away by saying that this unHumean magnitude is none other than *chance*! I say that I haven't the faintest notion how an unHumean magnitude can possibly do what it must do

to deserve that name—namely, fit into the principle about rationality of credences—so don't just stipulate that it bears that name Don't say here's chance, now is it Humean or not? Ask is there any way that any Humean magnitude could fill the chance-role? Is there any way that an unHumean magnitude could? What I fear is that the answer is "no" both times! Yet how can I reject the very idea of chance, when I know full well that each tritium atom has a certain chance of decaying at any moment?[7]

I thank all those who have helped me to think about these matters Those who have helped me most are listed in the footnotes to the papers and the postscripts Also I thank the University of California at Los Angeles, Princeton University, St Catherine's College, Oxford, the American Council of Learned Societies, The University of Adelaide and the Australian–American Education Foundation, the National Science Foundation, Victoria University of Wellington and the New Zealand–United States Educational Foundation, Monash University, The Australian National University, La Trobe University, and all those universities that have given me opportunities to try these papers out on critical audiences

For advice and assistance in planning these two volumes, and in seeing the project through difficult times, I am most grateful to Jim Anderson, Jonathan Bennett, Adam Hodgkin, Ruth Marcus, Tom Nagel, and Robert Stalnaker I thank Nancy Etchemendy for the diagrams in Postscript E to "Causation "

D L

Princeton
October 1984

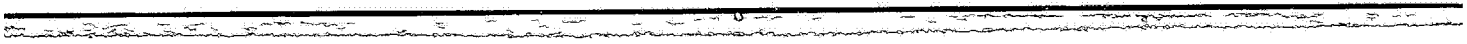
[7] D M Armstrong has pointed out (in discussion) that matters are still worse if we grant that chances may take extreme values, one or zero exactly Let H specify some course of history up to a certain moment and let F specify some course of history after that moment Assume that H and F are contingent (We need not assume that they are maximally specific) Let T be a history-to-chance conditional which says that after history H, the chance of F would be exactly one To grant that chances may take extreme values is to grant that some such H and T might both hold Then is there any possibility that H and T might hold without F? I say not Any genuine possibility deserves at least some small share of credence, perhaps infinitesimal but not zero but to give nonzero credence to this alleged possibility would violate the Principal Principle So H and T strictly imply F Now consider our three hypotheses about the status of history-to chance conditionals

1 Are they noncontingent? If so *T* is necessary, since *ex hypothesi* it is at least possible Then *H* by itself strictly implies *F* How can that be? What prevents us having *H* without *F*, when they specify the character of wholly distinct parts of the world? This necessary connection between distinct existences is unintelligible

2 Are they contingent, but supervenient on the arrangement of qualities? Then what would make *T* true is some pattern in the arrangement of qualities and it is open to say that part of that pattern is simply that *H* does not hold or that *F* does If so, we know how *H* and *T* can strictly imply *F*, so this second hypothesis gives no special problem about the case of extreme chances But it still has its general problem apart from the extreme case, how can a chancemaking pattern not give itself some chance of failing to obtain?

3 Are they contingent, and not supervenient upon the arrangement of qualities? Then if *T* is true, there is some unHumean feature of the world that makes it true Call this unHumean chancemaker *X* Now *X* and *H* strictly imply *F* How can that be? How could *X* manage to impose this constraint on the arrangement of qualities? If we reject strict implication of *F* by *H* alone, as we should, then we grant that there are arrangements of qualities which make *H* hold without *F* How does *X* prevent us from having any of these arrangements? Compare this unHumean chancemaker with Armstrong s unHumean lawmaker, denounced above Armstrong has a fair *tu quoque* against anyone who accepts the one and balks at the other For the two are alike in their supposed power to constrain the course of events, except that one imposes a connection in the single case and the other imposes a general regularity (Indeed the chancemaker might just be the lawmaker at work in one particular instance) Either way, it s unintelligible how the unHumean constrainer can possibly do its stuff

None of these three alternatives seems at all good The escape routes from the trilemma—doubting that chances really can take the extreme values doubting that every genuine possibility deserves some slight credence, or doubting the Principal Principle—seem just as bad But so far as I can see, we must choose one evil or another



PART FOUR

Counterfactuals and Time

SIXTEEN

Counterfactuals and Comparative Possibility*

In the last dozen years or so, our understanding of modality has been much improved by means of possible-world semantics the project of analyzing modal language by systematically specifying the conditions under which a modal sentence is true at a possible world I hope to do the same for counterfactual conditionals I write $A \mathbin{\Box\!\!\rightarrow} C$ for the counterfactual conditional with antecedent A and consequent C It may be read as "If it were the case that A, then it would be the case that C" or some more idiomatic paraphrase thereof

1 ANALYSES

I shall lead up by steps to an analysis I believe to be satisfactory

ANALYSIS 0 $A \mathbin{\Box\!\!\rightarrow} C$ *is true at world* i *iff* C *holds at every* A*-world such that* — "A-world", of course, means "world where A holds"

The blank is to be filled in with some sort of condition restricting the

* The theory presented in this paper is discussed more fully in my book *Counterfactuals* My research on counterfactuals was supported by a fellowship from the American Council of Learned Societies

A-worlds to be considered The condition may depend on i but not on A For instance, we might consider only those A-worlds that agree with i in certain specified respects On this analysis, the counterfactual is some fixed strict conditional

No matter what condition we put into the blank, Analysis 0 cannot be correct For it says that if $A \boxarrow \bar{B}$ is true at i, $\bar{B}$ holds at every A-world such that — In other words, there are no AB-worlds such that — Then $AB \boxarrow \bar{C}$ and $AB \boxarrow C$ are alike vacuously true, and $-(AB \boxarrow C)$ and $-(AB \boxarrow \bar{C})$ are alike false, for any C whatever On the contrary it can perfectly well happen that $A \boxarrow \bar{B}$ is true, yet $AB \boxarrow \bar{C}$ is non-vacuous, and $AB \boxarrow C$ is false In fact, we can have an arbitrarily long sequence like this of non-vacuously true counterfactuals and true denials of their opposites

$A \boxarrow \bar{B}$ and $-(A \boxarrow B)$,
$AB \boxarrow \bar{C}$ and $-(AB \boxarrow C)$,
$ABC \boxarrow \bar{D}$ and $-(ABC \boxarrow D)$,
etc

Example if Albert had come to the party, he would not have brought Betty, for, as he knows, if he had come and had brought Betty, Carl would not have stayed, for, as Carl knows, if Albert had come and had brought Betty and Carl had stayed, Daisy would not have danced with him, Each step of the sequence is a counterexample to Analysis 0 The counterfactual is not any strict conditional whatever

Analysis 0 also says that $A \boxarrow C$ implies $AB \boxarrow C$ If C holds at every A-world such that —, then C holds at such of those worlds as are B-worlds On the contrary we can have an arbitrarily long sequence like this of non-vacuously true counterfactuals and true denials of their opposites

$A \boxarrow \bar{Z}$ and $-(A \boxarrow Z)$,
$AB \boxarrow Z$ and $-(AB \boxarrow \bar{Z})$,
$ABC \boxarrow \bar{Z}$ and $-(ABC \boxarrow Z)$,
etc

Example if I had shirked my duty, no harm would have ensued, but if I had and you had too, harm would have ensued, but if I had and you had too and a third person had done far more than his duty, no harm would have ensued For this reason also the counterfactual is not any strict conditional whatever

More precisely, it is not any one, fixed strict conditional But this much of Analysis 0 is correct (1) to assess the truth of a counterfactual

we must consider whether the consequent holds at certain antecedent-worlds, (2) we should not consider all antecedent-worlds, but only some of them We may ignore antecedent-worlds that are gratuitously remote from actuality

Rather than any fixed strict conditional, we need a *variably strict conditional* Given a far-fetched antecedent, we look perforce at antecedent-worlds remote from actuality There are no others to look at But given a less far-fetched antecedent, we can afford to be more fastidious and ignore the very same worlds In considering the supposition "if I had just let go of my pen " I will go wrong if I consider bizarre worlds where the law of gravity is otherwise than it actually is, whereas in considering the supposition "if the planets traveled in spirals " I will go just as wrong if I ignore such worlds

It is this variable strictness that accounts for our counterexample sequences It may happen that we can find an A-world that meets some stringent restriction, before we can find any AB-world we must relax the restriction, before we can find any ABC-world we must relax it still more, and so on If so a counterexample sequence of the first kind definitely will appear, and one of the second kind will appear also if there is a suitable Z

We dream of considering a world where the antecedent holds but everything else is just as it actually is, the truth of the antecedent being the one difference between that world and ours No hope Differences never come singly, but in infinite multitudes Take, if you can, a world that differs from ours *only* in that Caesar did not cross the Rubicon Are his predicament and ambitions there just as they actually are? The regularities of his character? The psychological laws exemplified by his decision? The orders of the day in his camp? The preparation of the boats? The sound of splashing oars? Hold *everything* else fixed after making one change, and you will not have a possible world at all

If we cannot have an antecedent-world that is otherwise just like our world, what can we have? This, perhaps an antecedent-world that does not differ gratuitously from ours, one that differs only as much as it must to permit the antecedent to hold, one that is closer to our world in similarity, all things considered, than any other antecedent-world Here is a first analysis of the counterfactual as a variably strict conditional

ANALYSIS 1 $A \mathrel{\square\!\!\rightarrow} C$ *is true at* i *iff* C *holds at the closest (accessible)* A*-world to* i*, if there is one* This is Stalnaker's proposal in "A Theory of Conditionals", and elsewhere

It may be objected that Analysis 1 is founded on comparative similarity—"closeness"—of worlds, and that comparative similarity is hopelessly imprecise unless some definite respect of comparison has been specified Imprecise it may be, but that is all to the good Counterfactuals are imprecise too Two imprecise concepts may be rigidly fastened to one another, swaying together rather than separately, and we can hope to be precise about their connection Imprecise though comparative similarity may be, we *do* judge the comparative similarity of complicated things like cities or people or philosophies—and we do it often without benefit of any definite respect of comparison stated in advance We balance off various similarities and dissimilarities according to the importances we attach to various respects of comparison and according to the degrees of similarity in the various respects Conversational context, of course, greatly affects our weighting of respects of comparison, and even in a fixed context we have plenty of latitude Still, not anything goes We have concordant mutual expectations, mutual expectations of expectations, etc , about the relative importances we will attach to respects of comparison Often these are definite and accurate and firm enough to resolve the imprecision of comparative similarity to the point where we can converse without misunderstanding Such imprecision we can live with Still, I grant that a counterfactual based on comparative similarity has no place in the language of the exact sciences

I imposed a restriction to A-worlds "accessible" from i In this I follow Stalnaker, who in turn is following the common practice in modal logic We might think that there are some worlds so very remote from i that they should always be ignored (at i) even if some of them happen to be A-worlds and there are no closer A-worlds If so, we have the wherewithal to ignore them by deeming them *inaccessible* from i I can think of no very convincing cases, but I prefer to remain neutral on the point If we have no need for accessibility restrictions, we can easily drop them by stipulating that all worlds are mutually interaccessible

Unfortunately, Analysis 1 depends on a thoroughly implausible assumption that there will never be more than one closest A-world So fine are the gradations of comparative similarity that despite the infinite number and variety of worlds every tie is broken

Example A is "Bizet and Verdi are compatriots", F is "Bizet and Verdi are French", I is "Bizet and Verdi are Italian" Grant for the sake of argument that we have the closest F-world and the closest I-world, that these are distinct (dual citizenships would be a gratuitous difference from actuality), and that these are the two finalists in the

competition for closest A-world It might be that something favors one over the other—for all I know, Verdi narrowly escaped settling in France and Bizet did not narrowly escape settling in Italy But we can count on no such luck The case may be perfectly balanced between respects of comparison that favor the F-world and respects that favor the I-world It is out of the question, on Analysis 1, to leave the tie unbroken That means there is no such thing as *the* closest A-world Then anything you like holds at the closest A-world if there is one, because there isn't one If Bizet and Verdi had been compatriots they would have been Ukranian

ANALYSIS 2 $A \boxrightarrow C$ *is true at i iff C holds at every closest (accessible) A-world to i, if there are any* This is the obvious revision of Stalnaker's analysis to permit a tie in comparative similarity between several equally close closest A-worlds

Under Analysis 2 unbreakable ties are no problem The case of Bizet and Verdi comes out as follows $A \boxrightarrow F$, $A \boxrightarrow \overline{F}$, $A \boxrightarrow I$, and $A \boxrightarrow \overline{I}$ are all false $A \boxrightarrow (F \vee I)$ and $A \boxrightarrow (\overline{F} \vee \overline{I})$ are both true $A \boxrightarrow FI$ and $A \boxrightarrow \overline{F}\,\overline{I}$ are both false These conclusions seem reasonable enough

This reasonable settlement, however, does not sound so good in words $A \boxrightarrow F$ and $A \boxrightarrow \overline{F}$ are both false, so we want to assert their negations But negate their English readings in any straightforward and natural way, and we do not get $-(A \boxrightarrow F)$ and $-(A \boxrightarrow \overline{F})$ as desired Rather the negation moves in and attaches only to the consequent, and we get sentences that seem to mean $A \boxrightarrow \overline{F}$ and $A \boxrightarrow \overline{\overline{F}}$—a pair of falsehoods, together implying the further falsehood that Bizet and Verdi could not have been compatriots, and exactly the opposite of what we meant to say

Why is it so hard to negate a whole counterfactual, as opposed to negating the consequent? The defender of Analysis 1 is ready with an explanation Except when A is impossible, he says, there is a unique closest A-world Either C is false there, making $-(A \boxrightarrow C)$ and $A \boxrightarrow \overline{C}$ alike true, or C is true there, making them alike false Either way, the two agree We have no need of a way to say $-(A \boxrightarrow C)$ because we might as well say $A \boxrightarrow \overline{C}$ instead (except when A is impossible, in which case we have no need of a way to say $-(A \boxrightarrow C)$ because it is false)

There is some appeal to the view that $-(A \boxrightarrow C)$ and $A \boxrightarrow \overline{C}$ are equivalent (except when A is impossible) and we might be tempted thereby to return to Analysis 1 We might do better to return only part

way, using Bas van Fraassen's method of supervaluations to construct a compromise between Analyses 1 and 2

ANALYSIS $1\frac{1}{2}$ *$A \boxrightarrow C$ is true at i iff C holds at a certain arbitrarily chosen one of the closest (accessible) A-worlds to i, if there are any A sentence is super-true iff it is true no matter how the arbitrary choices are made, super-false iff false no matter how the arbitrary choices are made Otherwise it has no super-truth value Unless a particular arbitrary choice is under discussion, we abbreviate "super-true" as "true", and so on* Something of this kind is mentioned at the end of Thomason, "A Fitch-Style Formulation of Conditional Logic"

Analysis $1\frac{1}{2}$ agrees with Analysis 1 about the equivalence (except when A is impossible) of $-(A \boxrightarrow C)$ and $A \boxrightarrow \bar{C}$ If there are accessible A-worlds, the two agree in truth (i e super-truth) value, and further their biconditional is (super-)true On the other hand, Analysis $1\frac{1}{2}$ tolerates ties in comparative similarity as happily as Analysis 2 Indeed a counterfactual is (super-)true under Analysis $1\frac{1}{2}$ iff it is true under Analysis 2 On the other hand, a counterfactual false under Analysis 2 may either be false or have no (super-)truth value under Analysis $1\frac{1}{2}$ The case of Bizet and Verdi comes out as follows $A \boxrightarrow F$, $A \boxrightarrow I$, $A \boxrightarrow \bar{F}$, $A \boxrightarrow \bar{I}$, and their negations have no truth value $A \boxrightarrow (F \vee I)$ and $A \boxrightarrow (\bar{F} \vee \bar{I})$ are (super-)true $A \boxrightarrow FI$ and $A \boxrightarrow \bar{F}\bar{I}$ are (super-)false

This seems good enough For all I have said yet, Analysis $1\frac{1}{2}$ solves the problem of ties as well as Analysis 2, provided we're not too averse to (super-)truth value gaps But now look again at the question how to deny a counterfactual We have a way after all to deny a "would" counterfactual, use a "might" counterfactual with the same antecedent and negated consequent In reverse likewise to deny a "might" counterfactual, use a "would" counterfactual with the same antecedent and negated consequent Writing $A \diamondrightarrow C$ for "If it were the case that A, then it might be the case that C" or some more idiomatic paraphrase, we have these valid-sounding equivalences

(1) $-(A \boxrightarrow C)$ is equivalent to $A \diamondrightarrow \bar{C}$,
(2) $-(A \diamondrightarrow C)$ is equivalent to $A \boxrightarrow \bar{C}$

The two equivalences yield an explicit definition of "might" from "would" counterfactuals

$$A \diamondrightarrow C =^{df} -(A \boxrightarrow \bar{C}),$$

or, if we prefer, the dual definition of "would" from "might" Accord-

ing to this definition and Analysis 2, $A \diamond\!\!\rightarrow C$ is true at i iff C holds at some closest (accessible) A-world to i In the case of Bizet and Verdi, $A \diamond\!\!\rightarrow F$, $A \diamond\!\!\rightarrow \overline{F}$, $A \diamond\!\!\rightarrow I$, $A \diamond\!\!\rightarrow \overline{I}$ are all true, so are $A \diamond\!\!\rightarrow (F \vee I)$ and $A \diamond\!\!\rightarrow (\overline{F} \vee \overline{I})$, but $A \diamond\!\!\rightarrow FI$ and $A \diamond\!\!\rightarrow FI$ are false

According to the definition and Analysis 1 or $1\frac{1}{2}$, on the other hand, $A \diamond\!\!\rightarrow C$ and $A \Box\!\!\rightarrow C$ are equivalent except when A is impossible That should put the defender of those analyses in an uncomfortable spot He cannot very well claim that "would" and "might" counterfactuals do not differ except when the antecedent is impossible He must therefore reject my definition of the "might" counterfactual, and with it, the equivalences (1) and (2), uncontroversial though they sound He then owes us some other account of the might" counterfactual, which I do not think he can easily find Finally, once we see that we do have a way to negate a whole counterfactual, we no longer appreciate his explanation of why we don't need one I conclude that he would be better off moving at least to Analysis 2

Unfortunately, Analysis 2 is not yet satisfactory Like Analysis 1, it depends on an implausible assumption Given that some A-world is accessible from i, we no longer assume that there must be *exactly* one closest A-world to i, but we still assume that there must be *at least* one I call this the *Limit Assumption* It is the assumption that as we proceed to closer and closer A-worlds we eventually hit a limit and can go no farther But why couldn't it happen that there are closer and closer A-worlds without end—for each one, another even closer to i? Example A is "I am over 7 feet tall" If there are closest A-worlds to ours, pick one of them how tall am I there? I must be $7 + \varepsilon$ feet tall, for some positive ε, else it would not be an A-world But there are A-worlds where I am only $7 + \varepsilon/2$ feet tall Since that is closer to my actual height, why isn't one of these worlds closer to ours than the purportedly closest A-world where I am $7 + \varepsilon$ feet tall? And why isn't a suitable world where I am only $7 + \varepsilon/4$ feet even closer to ours, and so ad infinitum? (In special cases, but not in general, there may be a good reason why not Perhaps $7 + \varepsilon$ could have been produced by a difference in one gene, whereas any height below that but still above 7 would have taken differences in many genes) If there are A-worlds closer and closer to i without end, then any consequent you like holds at every closest A-world to i, because there aren't any If I were over 7 feet tall I would bump my head on the sky

ANALYSIS 3 *$A \boxtimes\!\!\rightarrow C$ is true at i iff some (accessible) AC-world is closer to i than any $A\bar{C}$-world, if there are any (accessible) A-worlds*
This is my final analysis

Analysis 3 looks different from Analysis 1 or 2, but it is similar in principle Whenever there are closest (accessible) A-worlds to a given world, Analyses 2 and 3 agree on the truth value there of $A \boxtimes\!\!\rightarrow C$ They agree also, of course, when there are no (accessible) A-worlds When there are closer and closer A-worlds without end, $A \boxtimes\!\!\rightarrow C$ is true iff, as we proceed to closer and closer A-worlds, we eventually leave all the $A\bar{C}$-worlds behind and find only AC-worlds

Using the definition of $A \Diamond\!\!\rightarrow C$ as $-(A \boxtimes\!\!\rightarrow \bar{C})$, we have this derived truth condition for the "might" counterfactual $A \Diamond\!\!\rightarrow C$ is true at i iff for every (accessible) $A\bar{C}$-world there is some AC-world at least as close to i, and there are (accessible) A-worlds

We have discarded two assumptions about comparative similarity in going from Analysis 1 to Analysis 3 first Stalnaker's assumption of uniqueness, then the Limit Assumption What assumptions remain?

First the *Ordering Assumption* that for each world i, comparative similarity to i yields a *weak ordering* of the worlds accessible from i That is, writing $j \leqslant_i k$ to mean that k is not closer to i than j, each $\leqslant_i$ is *connected* and *transitive* Whenever j and k are accessible from i either $j \leqslant_i k$ or $k \leqslant_i j$, whenever $h \leqslant_i j$ and $j \leqslant_i k$, then $h \leqslant_i k$ It is convenient, if somewhat artificial, to extend the comparative similarity orderings to encompass also the inaccessible worlds, if any we stipulate that each $\leqslant_i$ is to be a weak ordering of *all* the worlds, and that j is closer to i than k whenever j is accessible from i and k is not (Equivalently whenever $j \leqslant_i k$, then if k is accessible from i so is j)

Second, the *Centering Assumption* that each world i is accessible from itself, and closer to itself than any other world is to it

2 REFORMULATIONS

Analysis 3 can be given several superficially different, but equivalent, reformulations

2 1 *Comparative Possibility*

Introduce a connective $\prec$ $A \prec B$ is read as "It is less remote from actuality that A than that B" or "It is more possible that A than that B" and is true at a world i iff some (accessible) A-world is closer to i than

is any B-world First a pair of modalities and then the counterfactual can be defined from this new connective of comparative possibility, as follows (Let $\perp$ be a sentential constant false at every world, or an arbitrarily chosen contradiction, later, let $\top =^{df} - \perp$)

$$\Diamond A =^{df} A < \perp, \quad \Box A =^{df} - \Diamond - A,$$
$$A \mathrel{\Box\!\!\rightarrow} C =^{df} \Diamond A \supset (AC < A\bar{C})$$

The modalities so defined are interpreted by means of accessibility in the usual way $\Diamond A$ is true at i iff some A-world is accessible from i, and $\Box A$ is true at i iff A holds throughout all the worlds accessible from i If accessibility restrictions are discarded, so that all worlds are mutually interaccessible, they become the ordinary "logical" modalities (We might rather have defined the two modalities and comparative possibility from the counterfactual

$$\Box A =^{df} \bar{A} \mathrel{\Box\!\!\rightarrow} \perp, \quad \Diamond A =^{df} - \Box - A,$$
$$A < B =^{df} \Diamond A \,\&\, ((A \vee B) \mathrel{\Box\!\!\rightarrow} A\bar{B})$$

Either order of definitions is correct according to the given truth conditions)

Not only is comparative possibility technically convenient as a primitive, it is of philosophical interest for its own sake It sometimes seems true to say It is possible that A but not that B, it is possible that B but not that C, C but not D, etc Example A is "I speak English", B is "I speak German" (a language I know), C is "I speak Finnish", D is "A dog speaks Finnish", E is "A stone speaks Finnish", F is "A number speaks Finnish" Perhaps if I say all these things, as I would like to, I am equivocating—shifting to weaker and weaker noncomparative senses of "possible" from clause to clause It is by no means clear that there are enough distinct senses to go around As an alternative hypothesis, perhaps the clauses are compatible comparisons of possibility without equivocation $A < B < C < D < E < F$ (Here and elsewhere, I compress conjunctions in the obvious way)

2 2 *Cotenability*

Call B *cotenable* at i with the supposition that A iff some A-world accessible from i is closer to i than any $\bar{B}$-world, or if there are no A-worlds accessible from i In other words iff, at i, the supposition that A is either more possible than the falsity of B, or else impossible Then $A \mathrel{\Box\!\!\rightarrow} C$ is true at i iff C follows from A together with auxiliary

premises B_1, , each true at i and cotenable at i with the supposition that A

There is less to this definition than meets the eye A conjunction is cotenable with a supposition iff its conjuncts all are, so we need only consider the case of a single auxiliary premise B That single premise may always be taken either as $\overline{A}$ (if A is impossible) or as $A \supset C$ (otherwise), so "follows" may be glossed as "follows by truth-functional logic"

Common opinion has it that laws of nature are cotenable with any supposition unless they are downright inconsistent with it What can we make of this? Whatever else laws may be, they are generalizations that we deem especially important If so, then conformity to the prevailing laws of a world i should weigh heavily in the similarity of other worlds to i Laws should therefore tend to be cotenable, unless inconsistent, with counterfactual suppositions Yet I think this tendency may be overridden when conformity to laws carries too high a cost in differences of particular fact Suppose, for instance, that i is a world governed (in all respects of the slightest interest to us) by deterministic laws Let A pertain to matters of particular fact at time t, let A be false at i, and determined at all previous times to be false There are some A-worlds where the laws of i are never violated, all of these differ from i in matters of particular fact at all times before t (Nor can we count on the difference approaching zero as we go back in time) There are other A-worlds exactly like i until very shortly before t when a small, local, temporary, imperceptible suspension of the laws permits A to come true I find it highly plausible that one of the latter resembles i on balance more than any of the former

2 3 *Degrees of Similarity*

Roughly, $A \mathrel{\Box\!\!\rightarrow} C$ is true at i iff either (1) there is some degree of similarity to i within which there are A-worlds and C holds at all of them, or (2) there are no A-worlds within any degree of similarity to i To avoid the questionable assumption that similarity of worlds admits somehow of numerical measurement, it seems best to identify each "degree of similarity to i" with a set of worlds regarded as the set of all worlds within that degree of similarity to i Call a set S of worlds a *sphere* around i iff every S-world is accessible from i and is closer to i than is any $\overline{S}$-world Call a sphere *A-permitting* iff it contains some A-world Letting spheres represent degrees of similarity, we have this reformulation $A \mathrel{\Box\!\!\rightarrow} C$ is true at i iff $A \supset C$ holds throughout some A-permitting sphere around i, if such there be

To review our other operators $A \Diamond\!\!\rightarrow C$ is true at i iff AC holds somewhere in every A-permitting sphere around i, and there are such $\Box A$ is true at i iff A holds throughout every sphere around i $\Diamond A$ is true at i iff A holds somewhere in some sphere around i $A \prec B$ is true at i iff some sphere around i permits A but not B Finally, B is cotenable at i with the supposition that A iff B holds throughout some A-permitting sphere around i, if such there be

Restated in terms of spheres, the Limit Assumption says that if there is any A-permitting sphere around i, then there is a smallest one—the intersection of all A-permitting spheres is then itself an A-permitting sphere We can therefore reformulate Analysis 2 as $A \Box\!\!\rightarrow C$ is true at i iff $A \supset C$ holds throughout the smallest A-permitting sphere around i, if such there be

These systems of spheres may remind one of neighborhood systems in topology, but that would be a mistake The topological concept of closeness captured by means of neighborhoods is purely local and qualitative, not comparative adjacent vs separated, no more Neighborhoods do not capture comparative closeness to a point because arbitrary supersets of neighborhoods of the point are themselves neighborhoods of a point The spheres around a world, on the other hand, are nested, wherefore they capture comparative closeness j is closer to i than k is (according to the definition of spheres and the Ordering Assumption) iff some sphere around i includes j but excludes k

2 4 *Higher-Order Quantification*

The formulation just given as a metalinguistic truth condition can also be stated, with the help of auxiliary apparatus, as an explicit definition in the object language

$$A \Box\!\!\rightarrow C =^{\text{df}} \Diamond A \supset \exists S(\Phi S \,\&\, \Diamond SA \,\&\, \Box(SA \supset C))$$

Here the modalities are as before, "S" is an object-language variable over propositions, and Φ is a higher-order predicate satisfied at a world i by a proposition iff the set of all worlds where that proposition holds is a sphere around i I have assumed that every set of worlds is the truth-set of some—perhaps inexpressible—proposition

We could even quantify over modalities, these being understood as certain properties of propositions Call a modality *spherical* iff for every world i there is a sphere around i such that the modality belongs at i to all and only those propositions that hold throughout that sphere

Letting ■ be a variable over all spherical modalities, and letting ◆ abbreviate – ■ –, we have

$$A \mathrel{\Box\!\!\rightarrow} C =^{\text{df}} \Diamond A \supset \exists \blacksquare (\blacklozenge A \,\&\, \blacksquare (A \supset C))$$

This definition captures explicitly the idea that the counterfactual is a variably strict conditional

To speak of variable strictness, we should be able to compare the strictness of different spherical modalities Call one modality *(locally) stricter* than another at a world i iff the second but not the first belongs to some proposition at i Call two modalities *comparable* iff it does not happen that one is stricter at one world and the other at another Call one modality *stricter* than another iff they are comparable and the first is stricter at some world Call one *uniformly stricter* than another iff it is stricter at every world Comparative strictness is only a partial ordering of the spherical modalities some pairs are incomparable However, we can without loss restrict the range of our variable ■ to a suitable subset of the spherical modalities on which comparative strictness is a linear ordering (Perhaps—iff the inclusion orderings of spheres around worlds all have the same order type—we can do better still, and use a subset linearly ordered by uniform comparative strictness) Unfortunately, these linear sets are not uniquely determined

Example suppose that comparative similarity has only a few gradations Suppose, for instance, that there are only five different (nonempty) spheres around each world Let $\Box_1 A$ be true at i iff A holds throughout the innermost (nonempty) sphere around i let $\Box_2 A$ be true at i iff A holds throughout the innermost-but-one, and likewise for $\Box_3$, $\Box_4$, and $\Box_5$ Then the five spherical modalities expressed by these operators are a suitable linear set Since we have only a finite range, we can replace quantification by disjunction

$$A \mathrel{\Box\!\!\rightarrow} C =^{\text{df}} \Diamond A \supset (\Diamond_1 A \,\&\, \Box_1 (A \supset C)) \vee \dots \vee (\Diamond_5 A \,\&\, \Box_5 (A \supset C))$$

See Goble, "Grades of Modality"

2 5 *Impossible Limit-Worlds*

We were driven from Analysis 2 to Analysis 3 because we had reason to doubt the Limit Assumption It seemed that sometimes there were closer and closer A-worlds to i without limit—that is, without any closest A-worlds None, at least, among the *possible* worlds But we

can find the closest A-worlds instead among certain *impossible* worlds, if we are willing to look there If we count these impossible worlds among the worlds to be considered, the Limit Assumption is rescued and we can safely return to Analysis 2

There are various ways to introduce the impossible limits we need The following method is simplest, but others can be made to seem a little less *ad hoc* Suppose there are closer and closer (accessible, possible) A-worlds to i without limit, and suppose Σ is any maximal set of sentences such that, for any finite conjunction C of sentences in Σ, $A \diamond\!\!\rightarrow C$ holds at i according to Analysis 3 (We can think of such a Σ as a full description of one—possible or impossible—way things might be if it were that A, from the standpoint of i) Then we must posit an impossible limit-world where all of Σ holds It should be accessible from i alone, it should be closer to i than all the possible A-worlds, but it should be no closer to i than any possible world that is itself closer than all the possible A-worlds (Accessibility from, and comparative similarity to, the impossible limit-worlds is undefined Truth of sentences there is determined by the way in which these worlds were introduced as limits, not according to the ordinary truth conditions) Obviously the Limit Assumption is satisfied once these impossible worlds have been added to the worlds under consideration It is easy to verify that the truth values of counterfactuals at possible worlds afterwards according to Analyses 2 and 3 alike agrees with their original truth values according to Analysis 3

The impossible worlds just posited are impossible in the least objectionable way The sentences true there may be *incompatible*, in that not all of them hold together at any possible world, but there is no (correct) way to derive any contradiction from them For a derivation proceeds from finitely many premises, and any finite subset of the sentences true at one of the limit-worlds *is* true together at some possible world Example recall the failure of the Limit Assumption among possible worlds when A is "I am over 7 feet tall" Our limit-worlds will be impossible worlds where A is true but all of "I am at least 7 1 feet tall", "I am at least 7 01 feet tall", "I am at least 7 001 feet tall", etc are false (Do not confuse these with possible worlds where I am infinitesimally more than 7 feet tall For all I know, there are such, but worlds where physical magnitudes can take "non-standard" values differing infinitesimally from a real number presumably differ from ours in a very fundamental way, making them far more remote from actuality than some of the standard worlds where I am, say, 7 1 feet tall If so, "Physical magnitudes never take non-standard values" is

false at any possible world where I am infinitesimally more than 7 feet tall, but true at the impossible closest A-worlds to ours.)

How bad is it to believe in these impossible limit-worlds? Very bad, I think, but there is no reason not to reduce them to something less objectionable, such as sets of propositions or even sentences. I do not like a parallel reduction of possible worlds, chiefly because it is incredible in the case of the possible world *we* happen to live in, and other possible worlds do not differ in kind from ours. But this objection does not carry over to the impossible worlds. We do not live in one of those, and possible and impossible worlds do differ in kind.

2.6 *Selection Functions*

Analysis 2, vindicated either by trafficking in impossible worlds or by faith in the Limit Assumption even for possible worlds, may conveniently be reformulated by introducing a function f that selects, for any antecedent A and possible world i, the set of all closest (accessible) A-worlds to i (the empty set if there are none). $A \mathbin{\Box\!\!\rightarrow} C$ is true at a possible world i iff C holds throughout the selected set $f(A, i)$. Stalnaker formulates Analysis 1 this way, except that his $f(A, i)$ is the unique member of the selected set, if such there be, instead of the set itself.

If we like, we can put the selection function into the object language, but to do this without forgetting that counterfactuals are in general contingent, we must have recourse to *double indexing*. That is, we must think of some special sentences as being true or false at a world i not absolutely, but in relation to a world j. An ordinary sentence is true or false at i, as the case may be, in relation to any j, it will be enough to deal with ordinary counterfactuals compounded out of ordinary sentences. Let $\not A$ (where A is ordinary) be a special sentence true at j in relation to i iff j belongs to $f(A, i)$. Then $\not A \supset C$ (where C is ordinary) is true at j in relation to i iff, if j belongs to $f(A, i)$, C holds at j. Then $\Box(\not A \supset C)$ is true at j in relation to i iff C holds at every world in $f(A, i)$ that is accessible from j. It is therefore true at i in relation to i itself iff C holds throughout $f(A, i)$—that is, iff $A \mathbin{\Box\!\!\rightarrow} C$ holds at i. Introducing an operator $\dagger$ such that $\dagger B$ is true at i in relation to j iff B is true at i in relation to i itself, we can define the counterfactual

$$A \mathbin{\Box\!\!\rightarrow} C =^{\text{df}} \dagger\Box(\not A \supset C)$$

An $\not$-operator without double indexing is discussed in Åqvist, "Modal

Logic with Subjunctive Conditionals and Dispositional Predicates", the †-operator was introduced in Vlach, *"Now" and "Then"*

2.7 *Ternary Accessibility*

If we like, we can reparse counterfactuals as $[A\Box\!\!\rightarrow]C$, regarding $\Box\!\!\rightarrow$ now not as a two-place operator but rather as taking one sentence A to make a one-place operator $[A\Box\!\!\rightarrow]$ If we have closest A-worlds—possible or impossible—whenever A is possible, then each $[A\ \Box\!\!\rightarrow]$ is a necessity operator interpretable in the normal way by means of an accessibility relation Call j *A-accessible* from i (or *accessible from i relative to A*) iff j is a closest (accessible) A-world from i, then $[A\Box\!\!\rightarrow]$ C is true at i iff C holds at every world A-accessible from i See Chellas, "Basic Conditional Logic"

3 FALLACIES

Some familiar argument-forms, valid for certain other conditionals, are invalid for my counterfactuals

Transitivity	Contraposition	Strengthening	Importation
$A \Box\!\!\rightarrow B$ $B \Box\!\!\rightarrow C$ ——— $A \Box\!\!\rightarrow C$	$A \Box\!\!\rightarrow C$ ——— $\overline{C} \Box\!\!\rightarrow \overline{A}$	$A \Box\!\!\rightarrow C$ ——— $AB \Box\!\!\rightarrow C$	$A \Box\!\!\rightarrow (B \supset C)$ ——— $AB \Box\!\!\rightarrow C$

However, there are related valid argument-forms that may often serve as substitutes for these

$A \Box\!\!\rightarrow B$ $AB \Box\!\!\rightarrow C$ ——— $A \Box\!\!\rightarrow C$	$\overline{C}$ $A \Box\!\!\rightarrow C$ ——— $\overline{C} \Box\!\!\rightarrow \overline{A}$	$A \Diamond\!\!\rightarrow B$ $A \Box\!\!\rightarrow C$ ——— $AB \Box\!\!\rightarrow C$	$A \Diamond\!\!\rightarrow B$ $A \Box\!\!\rightarrow (B \supset C)$ ——— $AB \Box\!\!\rightarrow C$

Further valid substitutes for transitivity are these

	$B \Box\!\!\rightarrow A$	$B \Diamond\!\!\rightarrow A$
$A \Box\!\!\rightarrow B$	$A \Box\!\!\rightarrow B$	$A \Box\!\!\rightarrow B$
$\Box (B \supset C)$	$B \Box\!\!\rightarrow C$	$B \Box\!\!\rightarrow C$
———	———	———
$A \Box\!\!\rightarrow C$	$A \Box\!\!\rightarrow C$	$A \Box\!\!\rightarrow C$

4 TRUE ANTECEDENTS

On my analysis, a counterfactual is so called because it is suitable for non-trivial use when the antecedent is presumed false, not because it implies the falsity of the antecedent It is conversationally inappropriate, of course, to use the counterfactual construction unless one supposes the antecedent false, but this defect is not a matter of truth conditions Rather, it turns out that a counterfactual with a true antecedent is true iff the consequent is true, as if it were a material conditional In other words, these two arguments are valid

$$(-)\ \frac{A,\ \bar{C}}{-(A \mathbin{\Box\!\!\rightarrow} C)} \qquad (+)\ \frac{A,\ C}{A \mathbin{\Box\!\!\rightarrow} C}$$

It is hard to study the truth conditions of counterfactuals with true antecedents Their inappropriateness eclipses the question whether they are true However, suppose that someone has unwittingly asserted a counterfactual $A \mathbin{\Box\!\!\rightarrow} C$ with (what you take to be) a true antecedent A Either of these replies would, I think, sound cogent

(−) Wrong, since in fact A and yet not C
(+) Right, since in fact A and indeed C

The two replies depend for their cogency—for the appropriateness of the word "since"—on the validity of the corresponding arguments

I confess that the case for (−) seems more compelling than the case for (+) One who wants to invalidate (+) while keeping (−) can do so if he is prepared to imagine that another world may sometimes be just as similar to a given world as that world is to itself He thereby weakens the Centering Assumption to this each world is self-accessible, and at least as close to itself as any other world is to it Making that change and keeping everything else the same, (−) is valid but (+) is not

5 COUNTERPOSSIBLES

If A is impossible, $A \mathbin{\Box\!\!\rightarrow} C$ is vacuously true regardless of the consequent C Clearly some counterfactuals with impossible antecedents are asserted with confidence, and should therefore come out true "If there were a decision procedure for logic, there woud be one for the halting problem" Others are not asserted by reason of the irrelevance of ante-

cedent to consequent "If there were a decision procedure for logic, there would be a sixth regular solid" or " the war would be over by now" But would these be confidently *denied*? I think not, so I am content to let all of them alike be true Relevance is welcome in the theory of conversation (which I leave to others) but not in the theory of truth conditions

If you do insist on making discriminations of truth value among counterfactuals with impossible antecedents, you might try to do this by extending the comparative similarity orderings of possible worlds to encompass also certain impossible worlds where not-too-blatantly impossible antecedents come true (These are worse than the impossible limit-worlds already considered, where impossible but consistent infinite combinations of possible true sentences come true) See recent work on impossible-world semantics for doxastic logic and for relevant implication, especially Routley, "Ultra-Modal Propositional Functors"

6 POTENTIALITIES

"Had the Emperor not crossed the Rubicon, he would never have become Emperor" does *not* mean that the closest worlds to ours where there is a unique Emperor and he did not cross the Rubicon are worlds where there is a unique Emperor and he never became Emperor Rather, it is *de re* with respect to "the Emperor", and means that he who actually is (or was at the time under discussion) Emperor has a counterfactual property, or *potentiality*, expressed by the formula "if x had not crossed the Rubicon, x would never have become Emperor" We speak of what would have befallen the actual Emperor, not of what would have befallen whoever would have been Emperor Such potentialities may also appear when we quantify into counterfactuals "Any Emperor who would never have become Emperor had he not crossed the Rubicon ends up wishing he hadn't done it" or "Any of these matches would light if it were scratched" We need to know what it is for something to have a potentiality—that is, to satisfy a counterfactual formula $A(x) \mathbin{\Box\!\!\rightarrow} C(x)$

As a first approximation, we might say that something x satisfies the formula $A(x) \mathbin{\Box\!\!\rightarrow} C(x)$ at a world i iff some (accessible) world where x satisfies $A(x)$ and $C(x)$ is closer to i than any world where x satisfies $A(x)$ and $\bar{C}(x)$, if there are (accessible) worlds where x satisfies $A(x)$

The trouble is that this depends on the assumption that one and the same thing can exist—can be available to satisfy formulas—at various worlds I reject this assumption, except in the case of certain abstract entities that inhabit no particular world, and think it better to say that concrete things are confined each to its own single world He who actually is Emperor belongs to our world alone, and is not available to cross the Rubicon or not, become Emperor or not, or do anything else at any other world But although he himself is not present elsewhere, he may have *counterparts* elsewhere inhabitants of other worlds who resemble him closely, and more closely than do the other inhabitants of the same world What he cannot do in person at other worlds he may do vicariously, through his counterparts there So, for instance, I might have been a Republican not because I myself am a Republican at some other world than this—I am not—but because I have Republican counterparts at some worlds See my "Counterpart Theory and Quantified Modal Logic"

Using the method of counterparts, we may say that something x satisfies the formula $A(x) \mathbin{\Box\!\!\rightarrow} C(x)$ at a world i iff some (accessible) world where some counterpart of x satisfies $A(x)$ and $C(x)$ is closer to i than any world where any counterpart of x satisfies $A(x)$ and $\overline{C}(x)$, if there are (accessible) worlds where a counterpart of x satisfies $A(x)$ This works also for abstract entities that inhabit no particular world but exist equally at all, if we say that for these things the counterpart relation is simply identity

A complication it seems that when we deal with relations expressed by counterfactual formulas with more than one free variable, we may need to mix different counterpart relations "If I were you I'd give up" seems to mean that some world where a character-counterpart of me is a predicament-counterpart of you and gives up is closer than any where a character-counterpart of me is a predicament-counterpart of you and does not give up (I omit provision for vacuity and for accessibility restrictions) The difference between Goodman's sentences

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

may be explained by the hypothesis that both are *de re* with respect to both "New York City" and "Georgia", and that a less stringent counterpart relation is used for the subject terms "New York City" in (1) and "Georgia" in (2) than for the object terms "Georgia" in (1) and

"New York City" in (2) I cannot say in general how grammar and context control which counterpart relation is used where

An independent complication since closeness of worlds and counterpart relations among their inhabitants are alike matters of comparative similarity, the two are interdependent At a world close to ours, the inhabitants of our world will mostly have close counterparts, at a world very different from ours, nothing can be a very close counterpart of anything at our world We might therefore wish to fuse closeness of worlds and closeness of counterparts, allowing these to balance off Working with comparative similarity among *pairs* of a concrete thing and the world it inhabits (and ignoring provision for vacuity and for accessibility restrictions), we could say that an inhabitant x of a world i satisfies $A(x) \mathbin{\Box\!\!\rightarrow} C(x)$ at i iff some such thing-world pair $\langle y, j\rangle$ such that y satisfies $A(x)$ and $C(x)$ at j is more similar to the pair $\langle x, i\rangle$ than is any pair $\langle z, k\rangle$ such that z satisfies $A(x)$ and $\bar{C}(x)$ at k To combine this complication and the previous one seems laborious but routine

7 COUNTERCOMPARATIVES

"If my yacht were longer than it is, I would be happier than I am" might be handled by quantifying into a counterfactual formula $\exists x, y$ (my yacht is x feet long & I enjoy y hedons & (my yacht is more than x feet long $\mathbin{\Box\!\!\rightarrow}$ I enjoy more than y hedons)) But sometimes, perhaps in this very example, comparison makes sense when numerical measurement does not An alternative treatment of countercomparatives is available using double indexing (Double indexing has already been mentioned in connection with the $\mathcal{A}$-operator, but if we wanted it both for that purpose and for this, we would need triple indexing) Let A be true at j in relation to i iff my yacht is longer at j than at i (more precisely if my counterpart at j has a longer yacht than my counterpart at i (to be still more precise, decide what to do when there are multiple counterparts or multiple yachts)), let C be true at j in relation to i iff I am happier at j than at i (more precisely if my counterpart) Then $A \mathbin{\Box\!\!\rightarrow} C$ is true at j in relation to i iff some world (accessible from j) where A and C both hold in relation to i is closer to j than any world where A and $\bar{C}$ both hold in relation to i So far, the relativity to i just tags along Our countercomparative is therefore true at i (in relation to any world) iff $A \mathbin{\Box\!\!\rightarrow} C$ is true at i in relation to i itself It is therefore $\dagger(A \mathbin{\Box\!\!\rightarrow} C)$

8 COUNTERFACTUAL PROBABILITY

"The probability that C, if it were the case that A, would be r" cannot be understood to mean any of

(1) $\text{Prob}(A \mathbin{\Box\!\!\rightarrow} C) = r$,
(2) $\text{Prob}(C|A) = r$, or
(3) $A \mathbin{\Box\!\!\rightarrow} \text{Prob}(C) = r$

Rather, it is true at a world i (with respect to a given probability measure) iff for any positive ε there exists an A-permitting sphere T around i such that for any A-permitting sphere S around i within T, $\text{Prob}(C|AS)$, unless undefined, is within ε of r

Example A is "The sample contained abracadabrene", C is "The test for abracadabrene was positive", Prob is my present subjective probability measure after watching the test come out negative and tentatively concluding that abracadabrene was absent I consider that the probability of a positive result, had abracadabrene been present, would have been 97% (1) I know that false negatives occur because of the inherently indeterministic character of the radioactive decay of the tracer used in the test, so I am convinced that no matter what the actual conditions were, there might have been a false negative even if abracadabrene had been present $\text{Prob}(A \mathbin{\Diamond\!\!\rightarrow} \bar{C}) \approx 1$, $\text{Prob}(A \mathbin{\Box\!\!\rightarrow} C) \approx 0$ (2) Having seen that the test was negative, I disbelieve C much more strongly than I disbelieve A, $\text{Prob}(AC)$ is much less than $\text{Prob}(A)$, $\text{Prob}(C|A) \approx 0$ (3) Unknown to me, the sample was from my own blood, and abracadabrene is a powerful hallucinogen that makes white things look purple Positive tests are white, negatives are purple So had abracadabrene been present, I would have strongly disbelieved C no matter what the outcome of the test really was $A \mathbin{\Box\!\!\rightarrow} \text{Prob}(C) \approx 0$ (Taking (3) *de re* with respect to "Prob" is just as bad since actually $\text{Prob}(C) \approx 0$, $A \mathbin{\Box\!\!\rightarrow} \text{Prob}(C) \approx 0$ also) My suggested definition seems to work, however, provided that the outcome of the test at a close A-world does not influence the closeness of that world to ours

9 ANALOGIES

The counterfactual as I have analyzed it is parallel in its semantics to operators in other branches of intensional logic, based on other comparative relations There is one difference in the case of these anal-

ogous operators, it seems best to omit the provision for vacuous truth They correspond to a doctored counterfactual $\Box\!\!\Rightarrow$ that is automatically false instead of automatically true when the antecedent is impossible $A \Box\!\!\Rightarrow C \stackrel{df}{=} \Diamond A \,\&\, (A \Box\!\!\rightarrow C)$

Deontic We have the operator $A \Box\!\!\Rightarrow_d C$, read as "Given that A, it ought to be that C", true at a world i iff some AC-world evaluable from the standpoint of i is better, from the standpoint of i, than any $A\bar{C}$-world Roughly (under a Limit Assumption), iff C holds at the best A-worlds See the operator of "conditional obligation" discussed in Hansson, "An Analysis of Some Deontic Logics"

Temporal We have $A \Box\!\!\Rightarrow_f C$, read as "When next A, it will be that C", true at a time t iff some AC-time after t comes sooner after t than any $A\bar{C}$-time, roughly, iff C holds at the next A-time We have also the past mirror image $A \Box\!\!\Rightarrow_p C$, read as "When last A, it was that C"

Egocentric (in the sense of Prior, "Egocentric Logic") We have $A \Box\!\!\Rightarrow_e C$, read as "The A is C", true for a thing x iff some AC-thing in x's ken is more salient to x than any $A\bar{C}$-thing, roughly, iff the most salient A-thing is C

To motivate the given truth conditions, we may note that these operators all permit sequences of truths of the two forms

$A \Box\!\!\Rightarrow \bar{B}$,		$A \Box\!\!\Rightarrow Z$,
$AB \Box\!\!\Rightarrow \bar{C}$,	and	$AB \Box\!\!\Rightarrow \bar{Z}$,
$ABC \Box\!\!\Rightarrow \bar{D}$,		$ABC \Box\!\!\Rightarrow Z$,
etc ,		etc

It is such sequences that led us to treat the counterfactual as a variably strict conditional The analogous operators here are likewise variably strict conditionals Each is based on a binary relation and a family of comparative relations in just the way that the (doctored) counterfactual is based on accessibility and the family of comparative similarity orderings In each case, the Ordering Assumption holds The Centering Assumption, however, holds only in the counterfactual case New assumptions hold in some of the other cases

In the deontic case, we may or may not have different comparative orderings from the standpoint of different worlds If we evaluate worlds according to their conformity to the edicts of the god who reigns at a given world, then we will get different orderings, and no worlds will be evaluable from the standpoint of a godless world If rather we evaluate worlds according to their total yield of hedons, then evaluability and comparative goodness of worlds will be absolute

In the temporal case, both the binary relation and the families of

comparative relations, both for "when next" and for "when last' , are based on the single underlying linear order of time

The sentence $(A \vee \bar{B}) \Box\!\!\Rightarrow_f AB$ is true at time t iff some A-time after t precedes any $\bar{B}$-time after t It thus approximates the sentence "Until A, B", understood as being true at t iff some A-time after t is not preceded by any $\bar{B}$-time after t Likewise $(A \vee \bar{B}) \Box\!\!\Rightarrow_p AB$ approximates "Since A, B", with "since" understood as the past mirror image of "until" Kamp has shown that "since" and "until" suffice to define all possible tense operators, provided that the order of time is a complete linear order, see his *Tense Logic and the Theory of Order* Do my approximations have the same power? No, consider "Until $\top$, $\bot$", true at t iff there is a next moment after t This sentence cannot be translated using my operators For if the order of time is a complete linear order with discrete stretches and dense stretches, then the given sentence will vary in truth value, but if in addition there is no beginning or end of time, and if there are no atomic sentences that vary in truth value, then no sentences that vary in truth value can be built up by means of truth-functional connectives, $\Box\!\!\Rightarrow_f$, and $\Box\!\!\Rightarrow_p$

Starting from any of our various $\Box\!\!\Rightarrow$-operators, we can introduce one-place operators I shall call the *inner modalities*

$$\Box A =^{\text{df}} \top \Box\!\!\Rightarrow A,$$

$$\Diamond A =^{\text{df}} - \Box - A,$$

and likewise in the analogous cases The inner modalities in the counterfactual case are of no interest (unless Centering is weakened), since $\Box A$ and $\Diamond A$ are both equivalent to A itself Nor are they anything noteworthy in the egocentric case In the deontic case, however, they turn out to be slightly improved versions of the usual so-called obligation and permission operators $\Box_d A$ is true at i iff some (evaluable) A-world is better, from the standpoint of i, than any $\bar{A}$-world, that is, iff either (1) there are best (evaluable) worlds, and A holds throughout them, or (2) there are better and better (evaluable) worlds without end, and A holds throughout all sufficiently good ones In the temporal case, $\Box_f A$ is true at t iff some A-time after t comes sooner than any $\bar{A}$-time, that is, iff either (1) there is a next moment, and A holds then, or (2) there is no next moment, and A holds throughout some interval beginning immediately and extending into the future $\Box_f A$ may thus be read "Immediately, A", as may $\Diamond_f A$, but in a somewhat different sense

If no worlds are evaluable from the standpoint of a given world—

say, because no god reigns there—it turns out that $\Box_d A$ is false and $\Diamond_d A$ is true for any A whatever Nothing is obligatory, everything is permitted Similarly for $\Box_f A$ and $\Diamond_f A$ at the end of time, if such there be, and for $\Box_p A$ and $\Diamond_p A$ at its beginning Modalities that behave in this way are called *abnormal*, and it is interesting to find these moderately natural examples of abnormality

10 AXIOMATICS

The set of all sentences valid under my analysis may be axiomatized taking the counterfactual connective as primitive One such axiom system—not the neatest—is the system **C1** of my paper "Completeness and Decidability of Three Logics of Counterfactual Conditionals", essentially as follows

Rules

If A and $A \supset B$ are theorems, so is B

If $(B_1 \;\&\; \quad) \supset C$ is a theorem, so is

$$((A \mathbin{\Box\!\!\rightarrow} B_1) \;\&\; \quad) \supset (A \mathbin{\Box\!\!\rightarrow} C)$$

Axioms

All truth-functional tautologies are axioms

$A \mathbin{\Box\!\!\rightarrow} A$

$(A \mathbin{\Box\!\!\rightarrow} B) \;\&\; (B \mathbin{\Box\!\!\rightarrow} A) \supset (A \mathbin{\Box\!\!\rightarrow} C) \equiv (B \mathbin{\Box\!\!\rightarrow} C)$

$((A \vee B) \mathbin{\Box\!\!\rightarrow} A) \vee ((A \vee B) \mathbin{\Box\!\!\rightarrow} B) \vee (((A \vee B) \mathbin{\Box\!\!\rightarrow} C) \equiv (A \mathbin{\Box\!\!\rightarrow} C) \;\&\; (B \mathbin{\Box\!\!\rightarrow} C))$

$A \mathbin{\Box\!\!\rightarrow} B \supset A \supset B$

$AB \supset A \mathbin{\Box\!\!\rightarrow} B$

(Rules and axioms here and henceforth should be taken as schematic) Recall that modalities and comparative possibility may be introduced via the following definitions $\Box A =^{\text{df}} \overline{A} \mathbin{\Box\!\!\rightarrow} \bot$, $\Diamond A =^{\text{df}} \text{—}\Box\text{—}A$, $A < B =^{\text{df}} \Diamond A \;\&\; ((A \vee B) \mathbin{\Box\!\!\rightarrow} A\overline{B})$

A more intuitive axiom system, called **VC**, is obtained if we take comparative possibility instead of the counterfactual as primitive Let $A \leqslant B =^{\text{df}} \text{—}(B < A)$

Rules

If A and $A \supset B$ are theorems, so is B

If $A \supset B$ is a theorem, so is $B \leqslant A$

Basic Axioms

All truth-functional tautologies are basic axioms
$A \preccurlyeq B \preccurlyeq C \supset A \preccurlyeq C$
$A \preccurlyeq B \vee B \preccurlyeq A$
$A \preccurlyeq (A \vee B) \vee B \preccurlyeq (A \vee B)$

Axiom **C**

$A\overline{B} \supset A \prec B$

Recall that modalities and the counterfactual may be introduced via the following definitions $\Diamond A =^{df} A \prec \bot$, $\Box A =^{df} -\Diamond -A$, $A \mathrel{\Box\!\!\rightarrow} C =^{df} \Diamond A \supset (AC \prec A\overline{C})$

VC and **C1** turn out to be definitionally equivalent That is, their respective definitional extensions (via the indicated definitions) yield exactly the same theorems It may now be verified that these theorems are exactly the ones we ought to have Since the definitions are correct (under my truth conditions) it is sufficient to consider sentences in the primitive notation of **VC**

In general, we may define a *model* as any quadruple $< I, R, \preccurlyeq, [\![\ ]\!] >$ such that

(1) I is a nonempty set (regarded as playing the role of the set of worlds),
(2) R is a binary relation over I (regarded as the accessibility relation),
(3) $\preccurlyeq$ assigns to each i in I a weak ordering $\preccurlyeq_i$ of I (regarded as the comparative similarity ordering of worlds from the standpoint of i) such that whenever $j \preccurlyeq_i k$, if iRk then iRj,
(4) $[\![\ ]\!]$ assigns to each sentence A a subset $[\![A]\!]$ of I (regarded as the set of worlds where A is true),
(5) $[\![-A]\!]$ is $I - [\![A]\!]$, $[\![A \,\&\, B]\!]$ is $[\![A]\!] \cap [\![B]\!]$, and so on,
(6) $[\![A \prec B]\!]$ is $\{i \varepsilon I$ for some j in $[\![A]\!]$ such that iRj, there is no k in $[\![B]\!]$ such that $k \preccurlyeq_i j\}$

The *intended models*, for the counterfactual case, are those in which I, R, $\preccurlyeq$, and $[\![\ ]\!]$ really are what we regarded them as being the set of worlds, some reasonable accessibility relation, some reasonable family of comparative similarity orderings, and an appropriate assignment to sentences of truth sets The Ordering Assumption has been written into the very definition of a model (clause 3) since it is common to the counterfactual case and the analogous cases as well As for the Centering Assumption, we must impose it on the intended models as a further condition

(C) R is reflexive on I, and $j \leqslant_i i$ only if $j = i$

It seems impossible to impose other purely mathematical conditions on the intended models (with the possible exception of (U), discussed below) We therefore hope that **VC** yields as theorems exactly the sentences valid—true at all worlds—in all models that meet condition (C) This is the case

VC is sound for models meeting (C), for the basic axioms are valid, and the rules preserve validity, in all models, and Axiom **C** is valid in any model meeting (C)

VC is complete for models meeting (C) for there is a certain such model in which only theorems of **VC** are valid This model is called the *canonical model* for **VC**, and is as follows

(1) I is the set of all maximal **VC**-consistent sets of sentences,
(2) iRj iff, for every sentence A in j, $\Diamond A$ is in i,
(3) $j \leqslant_i k$ iff there is no set Σ of sentences that overlaps j but not k, such that whenever $A \preccurlyeq B$ is in i and A is in Σ then B also is in Σ,
(4) i is in $[\![A]\!]$ iff A is in i

In the same way, we can prove that the system consisting of the rules, the basic axioms, and *any* combination of the axioms listed below is sound and complete for models meeting the corresponding combination of conditions Nomenclature the system generated by the rules, the basic axioms, and the listed axioms — is called V— (Note that the conditions are not independent (C) implies (W), which implies (T), which implies (N) (S) implies (L) (A –) implies (U –) (W) and (S) together imply (C) (C) and (A –) together imply (S) by implying the stronger, trivializing condition that no world is accessible from any other Accordingly, many combinations of the listed axioms are redundant)

Axioms

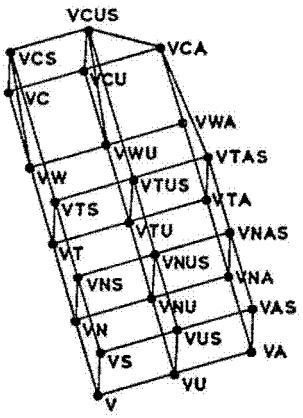
N $\Box \mathrm{T}$
T $\Box A \supset A$
W $AB \supset \Diamond A \,\&\, A \preccurlyeq B$
C $A\overline{B} \supset A \prec B$
L (no further axiom, or some tautology)
S $A \mathrel{\Box\!\!\rightarrow} C \vee A \mathrel{\Box\!\!\rightarrow} \overline{C}$
U $\Box A \supset \Box\Box A$ and $\Diamond A \supset \Box\Diamond A$
A $A \preccurlyeq B \supset \Box(A \preccurlyeq B)$ and $A \prec B \supset \Box(A \prec B)$

Conditions

(N) (normality) For any i in I there is some j in i such that iRj
(T) (total reflexivity) R is reflexive on I
(W) (weak centering) R is reflexive on I, for any i and j in I, $i \leqslant_i j$
(C) (centering) R is reflexive on I, and $j \leqslant_i i$ only if $j = i$
(L) (Limit Assumption) Whenever iRj for some j in $[\![A]\!]$, $[\![A]\!]$ has at least one $\leqslant_i$-minimal element
(S) (Stalnaker's Assumption) Whenever iRj for some j in $[\![A]\!]$, $[\![A]\!]$ has exactly one $\leqslant_i$-minimal element
(U –) (local uniformity) If iRj, then jRk iff iRk
(A –) (local absoluteness) If iRj, then jRk iff iRk and $h \leqslant_j k$ iff $h \leqslant_i k$

The Limit Assumption (L) corresponds to no special axiom Any one of our systems is sound and complete both for a combination of conditions without (L) and for that combination plus (L) The reason is that our canonical models always are rich enough to satisfy the Limit Assumption, but our axioms are sound without it (Except S, for which the issue does not arise because (S) implies (L)) Moral the Limit Assumption is irrelevant to the logical properties of the counterfactual Had our interest been confined to logic, we might as well have stopped with Analysis 2

Omitting redundant combinations of axioms, we have the 26 distinct systems shown in the diagram



The general soundness and completeness result still holds if we

replace the local conditions (U –) and (A –) by the stronger global conditions (U) and (A)

(U) (uniformity) For any i, j, k in I, jRk iff iRk
(A) (absoluteness) For any h, i, j, k in I, jRk iff iRk and $h \leqslant_j k$ iff $h \leqslant_i k$

Any model meeting (U –) or (A –) can be divided up into models meeting (U) or (A) The other listed conditions hold in the models produced by the division if they held in the original model Therefore a sentence is valid under a combination of conditions including (U) or (A) iff it is valid under the combination that results from weakening (U) to (U –), or (A) to (A –)

In the presence of (C), (W), or (T), condition (U) is equivalent to the condition for any i and j in I, iRj **VCU** is thus the correct system to use if we want to drop accessibility restrictions **VW**, or perhaps **VWU**, is the correct system for anyone who wants to invalidate the implication from A and C to $A \mathbin{\Box\!\!\rightarrow} C$ by allowing that another world might be just as close to a given world as that world is to itself **VCS**, or **VCUS** if we drop accessibility restrictions, is the system corresponding to Analysis 1 or $1\frac{1}{2}$ **VCS** is definitionally equivalent to Stalnaker's system **C2**

The systems given by various combinations of **N**, **T**, **U**, and **A** apply, under various assumptions, to the deontic case **VN** is definitionally equivalent to a system **CD** given by van Fraassen in "The Logic of Conditional Obligation", and shown there to be sound and complete for the class of what we may call *multi-positional models* meeting (N) These differ from models in my sense in that a world may occur at more than one position in an ordering $\leqslant_i$ (Motivation different positions may be assigned to one world *qua* realizer of different kinds of value) Technically, we no longer have a direct ordering of the worlds themselves, rather, we have for each i in I a linear ordering of some set V_i and an assignment to each world j such that iRj of one or more members of V_i, regarded as giving the positions of j in the ordering from the standpoint of i $A < B$ is true at i iff some position assigned to some A-world j (such that iRj) is better according to the given ordering than any position assigned to any B-world My models are essentially the same as those multi-positional models in which no world does have more than one assigned position in any of the orderings Hence **CD** is at least as strong as **VN**, but no stronger, since **VN** is already sound for all multi-positional models meeting (N)

All the systems are decidable To decide whether a given sentence A

is a theorem of a given system, it is enough to decide whether the validity of A under the corresponding combination of conditions can be refuted by a *small* countermodel—one with at most 2^n worlds, where n is the number of subsentences of A (Take (U) and (A), rather than (U–) and (A–), as the conditions corresponding to **U** and **A**) That can be decided by examining finitely many cases, since it is unnecessary to consider two models separately if they are isomorphic, or if they have the same I, R, $\leqslant$, and the same $[\![P]\!]$ whenever P is a sentence letter of A If A is a theorem, then by soundness there is no countermodel and *a fortiori* no small countermodel If A is not a theorem, then by completeness there is a countermodel $< I, R, \leqslant, [\![\ ]\!] >$ We derive thence a small countermodel, called a *filtration* of the original countermodel, as follows Let D_i, for each i in I, be the conjunction in some definite arbitrary order of all the subsentences of A that are true at i in the original countermodel, together with the negations of all the subsentences of A that are false at i in the original countermodel Now let $< I', R', \leqslant', [\![\ ]\!] >$ be as follows

(1) I' is a subset of I containing exactly one member of each nonempty $[\![D_i]\!]$,
(2) for any i and j in I , $iR'j$ iff i is in $[\![\Diamond D_j]\!]$,
(3) for any i, j, k in I', $j \leqslant_i' k$ iff i is in $[\![D_j \leqslant D_k]\!]$,
(4) for any sentence letter P, $[\![P]\!]$ is $[\![P]\!] \cap I$, for any compound sentence B, $[\![B]\!]$ is such that $<I$, R', $\leqslant$, $[\![\ ]\!]'>$ meets conditions (5) and (6) in the definition of a model

Then it may easily be shown that $<I$, R , $\leqslant$, $[\![\ ]\!] >$ is a small countermodel to the validity of A under the appropriate combination of conditions, and thereby to the theoremhood of A in the given system

REFERENCES

Lennart Åqvist, Modal Logic with Subjunctive Conditionals and Dispositional Predicates," *Journal of Philosophical Logic* 2 (1973) 1–76

Brian F Chellas, Basic Conditional Logic," *Journal of Philosophical Logic* 4 (1975) 133–53

Louis F Goble, "Grades of Modality," *Logique et Analyse* 51 (1970) 323–34

Bengt Hansson, "An Analysis of Some Deontic Logics," *Nous* 3 (1969) 373–98

Hans Kamp, *Tense Logic and the Theory of Order* (Ph D dissertation, University of California at Los Angeles, 1968)

David Lewis, "Counterpart Theory and Quantified Modal Logic," *Journal of Philosophy* 65 (1968) 113–26, reprinted in *Philosophical Papers*, Volume I

David Lewis, "Completeness and Decidability of Three Logics of Counterfactual Conditionals," *Theoria* 37 (1971) 74–85

David Lewis, *Counterfactuals* (Oxford Blackwell, 1973)

A N Prior, Egocentric Logic,' *Noûs* 2 (1968) 191–207

Richard Routley, "Ultra-Modal Propositional Functors," presented at the 1971 conference of the Australasian Association of Philosophy, Brisbane

Robert Stalnaker, ' A Theory of Conditionals," in Nicholas Rescher, ed , *Studies in Logical Theory* (Oxford Blackwell, 1968)

Richmond Thomason, "A Fitch-Style Formulation of Conditional Logic, *Logique et Analyse* 52 (1970) 397–412

Bas van Fraassen, "The Logic of Conditional Obligation,' *Journal of Philosophical Logic* 1 (1972) 417–38

Frank Vlach, *"Now" and "Then" A Formal Study in the Logic of Tense and Anaphora* (Ph D dissertation, University of California at Los Angeles, 1973)

SEVENTEEN

Counterfactual Dependence and Time's Arrow

THE ASYMMETRY OF COUNTERFACTUAL DEPENDENCE

Today I am typing words on a page Suppose today were different Suppose I were typing different words Then plainly tomorrow would be different also, for instance, different words would appear on the page Would yesterday also be different? If so, how? Invited to answer, you will perhaps come up with something But I do not think there is anything you can say about how yesterday would be that will seem clearly and uncontroversially true

The way the future is depends counterfactually on the way the present is If the present were different, the future would be different, and there are counterfactual conditionals, many of them as unquestionably true as counterfactuals ever get, that tell us a good deal about how the future would be different if the present were different in various ways Likewise the present depends counterfactually on the past, and in general the way things are later depends on the way things were earlier

Not so in reverse Seldom, if ever, can we find a clearly true counterfactual about how the past would be different if the present were somehow different Such a counterfactual, unless clearly false, normally is not clear one way or the other It is at best doubtful whether the past depends counterfactually on the present, whether the present depends

on the future, and in general whether the way things are earlier depends on the way things will be later

Often, indeed, we seem to reason in a way that takes it for granted that the past is counterfactually independent of the present that is, that even if the present were different, the past would be just as it actually is In reasoning from a counterfactual supposition, we use auxiliary premises drawn from (what we take to be) our factual knowledge But not just anything we know may be used, since some truths would not be true under the given supposition If the supposition concerns the present, we do not feel free to use all we know about the future If the supposition were true, the future would be different and some things we know about the actual future might not hold in this different counterfactual future But we do feel free, ordinarily, to use whatever we know about the past We evidently assume that even if our supposition about the present were true, the past would be no different If I were acting otherwise just now, I would revenge a wrong done me last year—it is absurd even to raise the question whether that past wrong would have taken place if I were acting otherwise now! More generally, in reasoning from a counterfactual supposition about any time, we ordinarily assume that facts about earlier times are counterfactually independent of the supposition and so may freely be used as auxiliary premises

I would like to present a neat contrast between counterfactual dependence in one direction of time and counterfactual independence in the other direction But until a distinction is made, the situation is not as neat as that There are some special contexts that complicate matters We know that present conditions have their past causes We can persuade ourselves, and sometimes do, that if the present were different then these past causes would have to be different, else they would have caused the present to be as it actually is Given such an argument—call it a *back-tracking argument*—we willingly grant that if the present were different, the past would be different too I borrow an example from Downing ([5]) Jim and Jack quarreled yesterday, and Jack is still hopping mad We conclude that if Jim asked Jack for help today, Jack would not help him But wait Jim is a prideful fellow He never would ask for help after such a quarrel, if Jim were to ask Jack for help today, there would have to have been no quarrel yesterday In that case Jack would be his usual generous self So if Jim asked Jack for help today, Jack would help him after all

At this stage we may be persuaded (and rightly so, I think) that if Jim asked Jack for help today, there would have been no quarrel yes-

terday But the persuasion does not last We very easily slip back into our usual sort of counterfactual reasoning, and implicitly assume once again that facts about earlier times are counterfactually independent of facts about later times Consider whether pride is costly In this case, at least, it costs Jim nothing It would be useless for Jim to ask Jack for help, since Jack would not help him We rely once more on the premise we recently doubted if Jim asked Jack for help today, the quarrel would nevertheless have taken place yesterday

What is going on, I suggest, can best be explained as follows (1) Counterfactuals are infected with vagueness, as everyone agrees Different ways of (partly) resolving the vagueness are appropriate in different contexts Remember the case of Caesar in Korea had he been in command, would he have used the atom bomb? Or would he have used catapults? It is right to say either, though not to say both together Each is true under a resolution of vagueness appropriate to some contexts (2) We ordinarily resolve the vagueness of counterfactuals in such a way that counterfactual dependence is asymmetric (except perhaps in cases of time travel or the like) Under this standard resolution, back-tracking arguments are mistaken if the present were different the past would be the same, but the same past causes would fail somehow to cause the same present effects If Jim asked Jack for help today, somehow Jim would have overcome his pride and asked despite yesterday's quarrel (3) Some special contexts favor a different resolution of vagueness, one under which the past depends counterfactually on the present and some back-tracking arguments are correct If someone propounds a back-tracking argument, for instance, his cooperative partners in conversation will switch to a resolution that gives him a chance to be right (This sort of accommodating shift in abstract features of context is common, see Lewis ([14])) But when the need for a special resolution of vagueness comes to an end, the standard resolution returns (4) A counterfactual saying that the past would be different if the present were somehow different may come out true under the special resolution of its vagueness, but false under the standard resolution If so, call it a *back-tracking counterfactual* Taken out of context, it will not be clearly true or clearly false Although we tend to favor the standard resolution, we also charitably tend to favor a resolution which gives the sentence under consideration a chance of truth

(Back-tracking counterfactuals, used in a context that favors their truth, are marked by a syntactic peculiarity They are the ones in which the usual subjunctive conditional constructions are readily

replaced by more complicated constructions "If it were that then it would have to be that " or the like A suitable context may make it acceptable to say "If Jim asked Jack for help today, there would have been no quarrel yesterday", but it would be more natural to say " there would have to have been no quarrel yesterday " Three paragraphs ago, I used such constructions to lure you into a context that favors back-tracking)

I have distinguished the standard resolution of vagueness from the sort that permits back-tracking only so that I can ask you to ignore the latter Only under the standard resolution do we have the clear-cut asymmetry of counterfactual dependence that interests me

I do not claim that the asymmetry holds in all possible, or even all actual, cases It holds for the sorts of familiar cases that arise in everyday life But it well might break down in the different conditions that might obtain in a time machine, or at the edge of a black hole, or before the Big Bang, or after the Heat Death, or at a possible world consisting of one solitary atom in the void It may also break down with respect to the immediate past We shall return to these matters later

Subject to these needed qualifications, what I claim is as follows Consider those counterfactuals of the form "If it were that A, then it would be that C" in which the supposition A is indeed false, and in which A and C are entirely about the states of affairs at two times t_A and t_C respectively Many such counterfactuals are true in which C also is false, and in which t_C is later than t_A These are counterfactuals that say how the way things are later depends on the way things were earlier But if t_C is earlier than t_A, then such counterfactuals are true if and only if C is true These are the counterfactuals that tell us how the way things are earlier does not depend on the way things will be later

ASYMMETRIES OF CAUSATION AND OPENNESS

The asymmetry of counterfactual dependence has been little discussed (However, see Downing [5], Bennett [2], and Slote [19]) Some of its consequences are better known It is instructive to see how the asymmetry of counterfactual dependence serves to explain these more familiar asymmetries

Consider the temporal asymmetry of causation Effects do not precede their causes, or at least not ordinarily Elsewhere ([12]) I have advocated a counterfactual analysis of causation (1) the relation of

cause to effect consists in their being linked by a causal chain, (2) a causal chain is a certain kind of chain of counterfactual dependences, and (3) the counterfactuals involved are to be taken under the standard resolution of vagueness If anything of the sort is right, there can be no backward causation without counterfactual dependence of past on future Only where the asymmetry of counterfactual dependence breaks down can there possibly be exceptions to the predominant futureward direction of causation

Consider also what I shall call the *asymmetry of openness* the obscure contrast we draw between the "open future" and the "fixed past" We tend to regard the future as a multitude of alternative possibilities, a "garden of forking paths" in Borges' phrase, whereas we regard the past as a unique, settled, immutable actuality These descriptions scarcely wear their meaning on their sleeves, yet do seem to capture some genuine and important difference between past and future What can it be? Several hypotheses do not seem quite satisfactory

Hypothesis 1 Asymmetry of Epistemic Possibility Is it just that we know more about the past than about the future, so that the future is richer in epistemic possibilities? I think that's not it The epistemic contrast is a matter of degree, not a difference in kind, and sometimes is not very pronounced There is a great deal we know about the future, and a great deal we don't know about the past Ignorance of history has not the least tendency to make (most of) us think of the past as somewhat future-like, multiple, open, or unfixed

Hypothesis 2 Asymmetry of Multiple Actuality Is it that all our possible futures are equally actual? It is possible, I think, to make sense of multiple actuality Elsewhere I have argued for two theses (in [9] and [8]) (1) any inhabitant of any possible world may truly call his own world actual, (2) we ourselves inhabit this one world only, and are not identical with our other-worldly counterparts Both theses are controversial, so perhaps I am right about one and wrong about the other If (1) is true and (2) is false, here we are inhabiting several worlds at once and truly calling all of them actual (Adams argues contrapositively in [1], arguing from the denial of multiple actuality and the denial of (2) to the denial of (1)) That makes sense, I think, but it gives us no asymmetry For in some sufficiently broad sense of possibility, we have alternative possible pasts as well as alternative possible futures But if (1) is true and (2) is false, that means that *all* our possibilities are equally actual, past as well as future

Hypothesis 3 Asymmetry of Indeterminism Is it that we think of our world as governed by indeterministic laws of nature, so that the actual past and present are nomically compossible with various alternative future continuations? I think this hypothesis also fails

For one thing, it is less certain that our world is indeterministic than that there is an asymmetry between open future and fixed past—whatever that may turn out to be Our best reason to believe in indeterminism is the success of quantum mechanics, but that reason is none too good until quantum mechanics succeeds in giving a satisfactory account of processes of measurement

For another thing, such reason as we have to believe in indeterminism is reason to believe that the laws of nature are indeterministic in both directions, so that the actual future and present are nomically compossible with various alternative pasts If there is a process of reduction of the wave packet in which a given superposition may be followed by any of many eigenstates, equally this is a process in which a given eigenstate may have been preceded by any of many superpositions Again we have no asymmetry

I believe that indeterminism is neither necessary nor sufficient for the asymmetries I am discussing Therefore I shall ignore the possibility of indeterminism in the rest of this paper, and see how the asymmetries might arise even under strict determinism A *deterministic* system of laws is one such that, whenever two possible worlds both obey the laws perfectly, then either they are exactly alike throughout all of time, or else they are not exactly alike through any stretch of time They are alike always or never They do not diverge, matching perfectly in their initial segments but not thereafter, neither do they converge Let us assume, for the sake of the argument, that the laws of nature of our actual world are in this sense deterministic

(My definition of determinism derives from Montague ([15]), but with modifications (1) I prefer to avoid his use of mathematical constructions as *ersatz* possible worlds But should you prefer *ersatz* worlds to the real thing, that will not matter for the purposes of this paper (2) I take exact likeness of worlds at times as a primitive relation, Montague instead uses the relation of having the same complete description in a certain language, which he leaves unspecified

My definition presupposes that we can identify stretches of time from one world to another That presupposition is questionable, but it could be avoided at the cost of some complication)

Hypothesis 4 Asymmetry of Mutability Is it that we can change the future, but not the past? Not so, if "change" has its literal meaning

It is true enough that if t is any past time, then we cannot bring about a difference between the state of affairs at t at time t_1 and the (supposedly changed) state of affairs at t at a later time t_2. But the pastness of t is irrelevant, the same would be true if t were present or future. Past, present, and future are alike immutable. What explains the impossibility is that such phrases as "the state of affairs at t at t_1" or "the state of affairs at t at t_2", if they mean anything, just mean the state of affairs at t. Of course we cannot bring about a difference between that and itself.

Final Hypothesis: Asymmetry of Counterfactual Dependence. Our fourth hypothesis was closer to the truth than the others. What we *can* do by way of "changing the future" (so to speak) is to bring it about that the future is the way it actually will be, rather than any of the other ways it would have been if we acted differently in the present. That is something like change. We make a difference. But it is not literally change, since the difference we make is between actuality and other possibilities, not between successive actualities. The literal truth is just that the future depends counterfactually on the present. It depends, partly, on what we do now.

Likewise, something we ordinarily *cannot* do by way of "changing the past" is to bring it about that the past is the way it actually was, rather than some other way it would have been if we acted differently in the present. The past would be the same, however we acted now. The past does not at all depend on what we do now. It is counterfactually independent of the present.

In short, I suggest that the mysterious asymmetry between open future and fixed past is nothing else than the asymmetry of counterfactual dependence. The forking paths into the future—the actual one and all the rest—are the many alternative futures that would come about under various counterfactual suppositions about the present. The one actual, fixed past is the one past that would remain actual under this same range of suppositions.

TWO ANALYSES OF COUNTERFACTUALS

I hope I have now convinced you that an asymmetry of counterfactual dependence exists, that it has important consequences, and therefore that it had better be explained by any satisfactory semantic analysis of counterfactual conditionals. In the rest of this paper, I shall consider how that explanation ought to work.

It might work by fiat It is an easy matter to build the asymmetry into an analysis of counterfactuals, for instance as follows

ANALYSIS 1 *Consider a counterfactual "If it were that A, then it would be that C" where A is entirely about affairs in a stretch of time* t_A *Consider all those possible worlds w such that*

(1) A is true at w,
(2) w is exactly like our actual world at all times before a transition period beginning shortly before t_A*,*
(3) w conforms to the actual laws of nature at all times after t_A*, and*
(4) during t_A *and the preceding transition period, w differs no more from our actual world than it must to permit A to hold*

The counterfactual is true if and only if C holds at every such world w

In short, take the counterfactual present (if t_A is now), avoiding gratuitous difference from the actual present, graft it smoothly onto the actual past, let the situation evolve according to the actual laws, and see what happens An analysis close to Analysis 1 has been put forward by Jackson ([7]) Bennett ([2]), Bowie ([3]), and Weiner ([21]) have considered, but not endorsed, similar treatments

Analysis 1 guarantees the asymmetry of counterfactual dependence, with an exception for the immediate past Let C be entirely about a stretch of time t_C If t_C is later than t_A, then C may very well be false at our world, yet true at the worlds that meet the conditions listed in Analysis 1 We have the counterfactuals whereby the affairs of later times depend on those of earlier times But if t_C is before t_A, and also before the transition period, then C holds at worlds that meet condition (2) if and only if C is true at our actual world Since C is entirely about something that does not differ at all from one of these worlds to another, its truth value cannot vary Therefore, except for cases in which t_C falls in the transition period, we have the counterfactuals whereby the affairs of earlier times are independent of those of later times

We need the transition period, and should resist any temptation to replace (2) by the simpler and stronger

(2) w is exactly like our actual world at all times before t_A

(2) makes for abrupt discontinuities Right up to t, the match was stationary and a foot away from the striking surface If it had been

struck at t, would it have travelled a foot in no time at all? No, we should sacrifice the independence of the immediate past to provide an orderly transition from actual past to counterfactual present and future. That is not to say, however, that the immediate past depends on the present in any very definite way. There may be a variety of ways the transition might go, hence there may be no true counterfactuals that say in any detail how the immediate past would be if the present were different. I hope not, since if there were a definite and detailed dependence, it would be hard for me to say why some of this dependence should not be interpreted—wrongly, of course—as backward causation over short intervals of time in cases that are not at all extraordinary.

Analysis 1 seems to fit a wide range of counterfactuals, and it explains the asymmetry of counterfactual dependence, though with one rather plausible exception. Should we be content? I fear not, for two reasons.

First, Analysis 1 is built for a special case. We need a supposition about a particular time, and we need a counterfactual taken under the standard resolution of vagueness. What shall we do with suppositions such as

If kangaroos had no tails.

If gravity went by the inverse cube of distance.

If Collett had ever designed a Pacific.

which are not about particular times? Analysis 1 cannot cope as it stands, nor is there any obvious way to generalize it. At most we could give separate treatments of other cases, drawing on the cases handled by Analysis 1. (Jackson ([7]) does this to some extent.) Analysis 1 is not much of a start toward a uniform treatment of counterfactuals in general.

Second, Analysis 1 gives us more of an asymmetry than we ought to want. No matter how special the circumstances of the case may be, no provision whatever is made for actual or possible exceptions to the asymmetry (except in the transition period). That is too inflexible. Careful readers have thought they could make sense of stories of time travel (see my [13] for further discussion), hard-headed psychical researchers have believed in precognition, speculative physicists have given serious consideration to tachyons, advanced potentials, and cosmological models with closed timelike curves. Most or all of these phenomena would involve special exceptions to the normal asymmetry

of counterfactual dependence It will not do to declare them impossible *a priori*

The asymmetry-by-fiat strategy of Analysis 1 is an instructive error, not a dead loss Often we do have the right sort of supposition, the standard resolution of vagueness, and no extraordinary circumstances Then Analysis 1 works as well as we could ask The right analysis of counterfactuals needs to be both more general and more flexible But also it needs to agree with Analysis 1 over the wide range of cases for which Analysis 1 succeeds

The right general analysis of counterfactuals, in my opinion, is one based on comparative similarity of possible worlds Roughly, a counterfactual is true if every world that makes the antecedent true without gratuitous departure from actuality is a world that also makes the consequent true Such an analysis is given in my [10] and [11], here is one formulation

ANALYSIS 2 *A counterfactual "If it were that A, then it would be that C" is (non-vacuously) true if and only if some (accessible) world where both A and C are true is more similar to our actual world, overall, than is any world where A is true but C is false*

This analysis is fully general *A* can be a supposition of any sort It is also extremely vague Overall similarity among worlds is some sort of resultant of similarities and differences of many different kinds, and I have not said what system of weights or priorities should be used to squeeze these down into a single relation of overall similarity I count that a virtue Counterfactuals are both vague and various Different resolutions of the vagueness of overall similarity are appropriate in different contexts

Analysis 2 (plus some simple observations about the formal character of comparative similarity) is about all that can be said in full generality about counterfactuals While not devoid of testable content—it settles some questions of logic—it does little to predict the truth values of particular counterfactuals in particular contexts The rest of the study of counterfactuals is not fully general Analysis 2 is only a skeleton It must be fleshed out with an account of the appropriate similarity relation, and this will differ from context to context Our present task is to see what sort of similarity relation can be combined with Analysis 2 to yield what I have called the standard resolution of vagueness one that invalidates back-tracking arguments, one that yields an asymmetry of counterfactual dependence except

perhaps under special circumstances, one that agrees with Analysis 1, our asymmetry-by-fiat analysis, whenever it ought to

But first, a word of warning! Do not assume that just any respect of similarity you can think of must enter into the balance of overall similarity with positive weight The point is obvious for some respects of similarity, if such they be It contributes nothing to the similarity of two gemstones that both are grue (To be *grue* is to be green and first examined before 2000 A D or blue and not first examined before 2000 A D) But even some similarities in less gruesome respects may count for nothing They may have zero weight, at least under some reasonable resolutions of vagueness To what extent are the philosophical writings of Wittgenstein similar, overall, to those of Heidegger? I don't know But here is one respect of comparison that does not enter into it at all, not even with negligible weight the ratio of vowels to consonants

(Bowie ([3]) has argued that if some respects of comparison counted for nothing, my assumption of "centering" in [10] and [11] would be violated worlds differing from ours only in the respects that don't count would be as similar to our world as our world is to itself I reply that there may not be any worlds that differ from ours only in the respects that don't count, even if there are some respects that don't count Respects of comparison may not be entirely separable If the writings of two philosophers were alike in every respect that mattered, they would be word-for-word the same, then they would have the same ratio of vowels to consonants)

And next, another word of warning! It is all too easy to make offhand similarity judgments and then assume that they will do for all purposes But if we respect the extreme shiftiness and context-dependence of similarity, we will not set much store by offhand judgments We will be prepared to distinguish between the similarity relations that guide our offhand explicit judgments and those that govern our counterfactuals in various contexts

Indeed, unless we are prepared so to distinguish, Analysis 2 faces immediate refutation Sometimes a pair of counterfactuals of the following form seem true "If A, the world would be very different, but if A and B, the world would not be very different " Only if the similarity relation governing counterfactuals disagrees with that governing explicit judgments of what is "very different" can such a pair be true under Analysis 2 (I owe this argument to Pavel Tichy and, in a slightly different form, to Richard J Hall) It seems to me no surprise, given the instability even of explicit judgments of similarity, that two different

comparative similarity relations should enter into the interpretation of a single sentence

The thing to do is not to start by deciding, once and for all, what we think about similarity of worlds, so that we can afterwards use these decisions to test Analysis 2 What that would test would be the combination of Analysis 2 with a foolish denial of the shiftiness of similarity Rather, we must use what we know about the truth and falsity of counterfactuals to see if we can find some sort of similarity relation—not necessarily the first one that springs to mind—that combines with Analysis 2 to yield the proper truth conditions It is this combination that can be tested against our knowledge of counterfactuals, not Analysis 2 by itself In looking for a combination that will stand up to the test, we must use what we know about counterfactuals to find out about the appropriate similarity relation—not the other way around

THE FUTURE SIMILARITY OBJECTION

Several people have raised what they take to be a serious objection against Analysis 2 (It was first brought to my attention by Michael Slote, it occurs, in various forms, in [2], [3], [4], [6], [7], [17], [18], and [19] Kit Fine ([6] 452) states it as follows

> The counterfactual "If Nixon had pressed the button there would have been a nuclear holocaust" is true or can be imagined to be so Now suppose that there never will be a nuclear holocaust Then that counterfactual is, on Lewis's analysis, very likely false For given any world in which antecedent and consequent are both true it will be easy to imagine a closer world in which the antecedent is true and the consequent false For we need only imagine a change that prevents the holocaust but that does not require such a great divergence from reality

The presence or absence of a nuclear holocaust surely does contribute with overwhelming weight to some prominent similarity relations (For instance, to one that governs the explicit judgment of similarity in the consequent of "If Nixon had pressed the button, the world would be very different ") But the relation that governs the counterfactual may not be one of these It may nevertheless be a relation of overall similarity—not because it is likely to guide our explicit judgments of similarity, but rather because it is a resultant, under some system of weights or priorities, of a multitude of relations of similarity in particular respects

Let us take the supposition that Nixon pressed the button as implicitly referring to a particular time t—let it be the darkest moment of the final days Consider w_0, a world that may or may not be ours At w_0, Nixon does not press the button at t and no nuclear holocaust ever occurs Let w_0 also be a world with deterministic laws, since we have confined our attention here to counterfactual dependence under determinism Let w_0 also be a world that fits our worst fantasies about the button there is such a button, it is connected to a fully automatic command and control system, the wired-in war plan consists of one big salvo, everything is in faultless working order, there is no way for anyone to stop the attack, and so on Then I agree that Fine's counterfactual is true at w_0 if Nixon had pressed the button, there would have been a nuclear holocaust

There are all sorts of worlds where Nixon (or rather, a counterpart of Nixon) presses the button at t We must consider which of these differ least, under the appropriate similarity relation, from w_0 Some are non-starters Those where the payload of the rockets consists entirely of confetti depart gratuitously from w_0 by any reasonable standards The more serious candidates fall into several classes

One class is typified by the world w_1 Until shortly before t, w_1 is exactly like w_0 The two match perfectly in every detail of particular fact, however minute Shortly before t, however, the spatio-temporal region of perfect match comes to an end as w_1 and w_0 begin to diverge The deterministic laws of w_0 are violated at w_1 in some simple, localized, inconspicuous way A tiny miracle takes place Perhaps a few extra neurons fire in some corner of Nixon's brain As a result of this, Nixon presses the button With no further miracles events take their lawful course and the two worlds w_1 and w_0 go their separate ways The holocaust takes place From that point on, at least so far as the surface of this planet is concerned, the two worlds are not even approximately similar in matters of particular fact In short, the worlds typified by w_1 are the worlds that meet the conditions listed in Analysis 1, our asymmetry-by-fiat analysis What is the case throughout these worlds is just what we think would have been the case if Nixon had pressed the button (assuming that we are at w_0, and operating under the standard resolution of vagueness) Therefore, the worlds typified by w_1 should turn out to be more similar to w_0, under the similarity relation we seek, than any of the other worlds where Nixon pressed the button

(When I say that a miracle takes place at w_1, I mean that there is a violation of the laws of nature But note that the violated laws are not laws of the same world where they are violated That is impossible,

whatever else a law may be, it is at least an exceptionless regularity I am using "miracle" to express a relation between different worlds A miracle at w_1, relative to w_0, is a violation at w_1 of the laws of w_0, which are at best the almost-laws of w_1 The laws of w_1 itself, if such there be, do not enter into it)

A second class of candidates is typified by w_2 This is a world completely free of miracles the deterministic laws of w_0 are obeyed perfectly However, w_2 differs from w_0 in that Nixon pressed the button By definition of determinism, w_2 and w_0 are alike always or alike never, and they are not alike always Therefore, they are not exactly alike through any stretch of time They differ even in the remote past What is worse, there is no guarantee whatever that w_2 can be chosen so that the differences diminish and eventually become negligible in the more and more remote past Indeed, it is hard to imagine how two deterministic worlds anything like ours could possibly remain just a little bit different for very long There are altogether too many opportunities for little differences to give rise to bigger differences

Certainly such worlds as w_2 should not turn out to be the most similar worlds to w_0 where Nixon pressed the button That would lead to back-tracking unlimited (And as Bennett observes in [2], it would make counterfactuals useless, we know far too little to figure out which of them are true under a resolution of vagueness that validates very much back-tracking) The lesson we learn by comparing w_1 and w_2 is that under the similarity relation we seek, a lot of perfect match of particular fact is worth a little miracle

A third class of candidates is typified by w_3 This world begins like w_1 Until shortly before t, w_3 is exactly like w_0 Then a tiny miracle takes place, permitting divergence Nixon presses the button at t But there is no holocaust, because soon after t a second tiny miracle takes place, just as simple and localized and inconspicuous as the first The fatal signal vanishes on its way from the button to the rockets Thereafter events at w_3 take their lawful course At least for a while, worlds w_0 and w_3 remain very closely similar in matters of particular fact But they are no longer exactly alike The holocaust has been prevented, but Nixon's deed has left its mark on the world w_3 There are his fingerprints on the button Nixon is still trembling, wondering what went wrong—or right His gin bottle is depleted The click of the button has been preserved on tape Light waves that flew out the window, bearing the image of Nixon's finger on the button, are still on their way into outer space The wire is ever so slightly warmed where the signal current passed through it And so on, and on, and on The differences

between w_3 and w_0 are many and varied, although no one of them amounts to much

I should think that the close similarity between w_3 and w_0 could not last Some of the little differences would give rise to bigger differences sooner or later Maybe Nixon's memoirs are more sanctimonious at w_3 than at w_0 Consequently they have a different impact on the character of a few hundred out of the millions who read them A few of these few hundred make different decisions at crucial moments of their lives—and we're off! But if you are not convinced that the differences need increase, no matter My case will not depend on that

If Analysis 2 is to succeed, such worlds as w_3 must not turn out to be the most similar worlds to w_0 where Nixon pressed the button The lesson we learn by comparing w_1 and w_3 is that under the similarity relation we seek, close but approximate match of particular fact (especially if it is temporary) is not worth even a little miracle Taking that and the previous lesson of w_2 together, we learn that perfect match of particular fact counts for much more than imperfect match, even if the imperfect match is good enough to give us similarity in respects that matter very much to us I do not claim that this pre-eminence of perfect match is intuitively obvious I do not claim that it is a feature of the similarity relations most likely to guide our explicit judgments It is not, else the objection we are considering never would have been put forward (See also the opinion survey reported by Bennett in [2]) But the pre-eminence of perfect match is a feature of some relations of overall similarity, and it must be a feature of any similarity relation that will meet our present needs

A fourth class of candidates is typified by w_4 This world begins like w_1 and w_3 There is perfect match with w_0 until shortly before t, there is a tiny divergence miracle, the button is pressed But there is a widespread and complicated and diverse second miracle after t It not only prevents the holocaust but also removes all traces of Nixon's button-pressing The cover-up job is miraculously perfect Of course the fatal signal vanishes, just as at w_3, but there is much more The fingerprint vanishes, and the sweat returns to Nixon's fingertip Nixon's nerves are soothed, his memories are falsified, and so he feels no need of the extra martini The click on the tape is replaced by innocent noises The receding light waves cease to bear their incriminating images The wire cools down, and not by heating its surroundings in the ordinary way And so on, and on, and on Not only are there no traces that any human detective could read, in every detail of particular fact, however minute, it is just as if the button-pressing had never been The worlds

w_4 and w_0 reconverge They are exactly alike again soon after t, and exactly alike forevermore All it takes is enough of a reconvergence miracle one involving enough different sorts of violations of the laws of w_0, in enough different places Because there are many different sorts of traces to be removed, and because the traces spread out rapidly, the cover-up job divides into very many parts Each part requires a miracle at least on a par with the small miracle required to prevent the holocaust, or the one required to get the button pressed in the first place Different sorts of unlawful processes are needed to remove different sorts of traces the miraculous vanishing of a pulse of current in a wire is not like the miraculous rearrangement of magnetized grains on a recording tape The big miracle required for perfect reconvergence consists of a multitude of little miracles, spread out and diverse

Such worlds as w_4 had better not turn out to be the most similar worlds to w_0 where Nixon pressed the button The lesson we learn by comparing w_1 and w_4 is that under the similarity relation we seek, perfect match of particular fact even through the entire future is not worth a big, widespread, diverse miracle Taking that and the lesson of w_2 together, we learn that avoidance of big miracles counts for much more than avoidance of little miracles Miracles are not all equal The all-or-nothing distinction between worlds that do and that do not ever violate the laws of w_0 is not sensitive enough to meet our needs

This completes our survey of the leading candidates There are other candidates, but they teach us nothing new There are some worlds where approximate reconvergence to w_0 is secured by a second small miracle before t, rather than afterward as at w_3 Haig has seen fit to disconnect the button Likewise there are worlds where a diverse and widespread miracle to permit perfect reconvergence takes place mostly before and during t Nixon's fingers leave no prints, the tape recorder malfunctions, and so on

Under the similarity relation we seek, w_1 must count as closer to w_0 than any of w_2, w_3, and w_4 That means that a similarity relation that combines with Analysis 2 to give the correct truth conditions for counterfactuals such as the one we have considered, taken under the standard resolution of vagueness, must be governed by the following system of weights or priorities

(1) It is of the first importance to avoid big, widespread, diverse violations of law
(2) It is of the second importance to maximize the spatio-temporal

region throughout which perfect match of particular fact prevails

(3) It is of the third importance to avoid even small, localized, simple violations of law

(4) It is of little or no importance to secure approximate similarity of particular fact, even in matters that concern us greatly

(It is a good question whether approximate similarities of particular fact should have little weight or none Different cases come out differently, and I would like to know why Tichy ([20]) and Jackson ([7]) give cases which appear to come out right under Analysis 2 only if approximate similarities count for nothing, but Morgenbesser has given a case, reported in Slote ([19]), which appears to go the other way This problem was first brought to my attention by Ernest Loevinsohn)

Plenty of unresolved vagueness remains, of course, even after we have distinguished the four sorts of respect of comparison and ranked them in decreasing order of importance But enough has been said to answer Fine's objection, and I think other versions of the future similarity objection may be answered in the same way

THE ASYMMETRY OF MIRACLES

Enough has been said, also, to explain why there is an asymmetry of counterfactual dependence in such a case as we have just considered If Nixon had pressed the button, the future would have been of the sort found at w_1 a future very different, in matters of particular fact, from that of w_0 The past also would have been of the sort found at w_1 a past exactly like that of w_0 until shortly before t Whence came this asymmetry? It is not built into Analysis 2 It is not built into the standards of similarity that we have seen fit to combine with Analysis 2

It came instead from an asymmetry in the range of candidates We considered worlds where a small miracle permitted divergence from w_0 We considered worlds where a small miracle permitted approximate convergence to w_0 and worlds where a big miracle permitted perfect convergence to w_0 But we did not consider any worlds where a small miracle permitted perfect convergence to w_0 If we had, our symmetric standards of similarity would have favored such worlds no less than w_1

But are there any such worlds to consider? What could they be like

how could one small, localized, simple miracle possibly do all that needs doing? How could it deal with the fatal signal, the fingerprints, the memories, the tape, the light waves, and all the rest? I put it to you that it can't be done! Divergence from a world such as w_0 is easier than perfect convergence to it Either takes a miracle, since w_0 is deterministic, but convergence takes very much more of a miracle The asymmetry of counterfactual dependence arises because the appropriate standards of similarity, themselves symmetric, respond to this asymmetry of miracles

It might be otherwise if w_0 were a different sort of world I do not mean to suggest that the asymmetry of divergence and convergence miracles holds necessary or universally For instance, consider a simple word inhabited by just one atom Consider the worlds that differ from it in a certain way at a certain time You will doubtless conclude that convergence to this world takes no more of a varied and widespread miracle than divergence from it That means, if I am right, that no asymmetry of counterfactual dependence prevails at this world Asymmetry-by-fiat analyses go wrong for such simple worlds The asymmetry of miracles, and hence of counterfactual dependence, rests on a feature of worlds like w_0 which very simple worlds cannot share

ASYMMETRY OF OVERDETERMINATION

Any particular fact about a deterministic world is predetermined throughout the past and postdetermined throughout the future At any time, past or future, it has at least one *determinant* a minimal set of conditions jointly sufficient, given the laws of nature, for the fact in question (Members of such a set may be causes of the fact, or traces of it, or neither) The fact may have only one determinant at a given time, disregarding inessential differences in a way I shall not try to make precise Or it may have two or more essentially different determinants at a given time, each sufficient by itself If so, it is *overdetermined* at that time Overdetermination is a matter of degree there might be two determinants, or there might be very many more than two

I suggest that what makes convergence take so much more of a miracle than divergence, in the case of a world such as w_0, is an asymmetry of overdetermination at such a world How much overdetermination of later affairs by earlier ones is there at our world, or at a deterministic world which might be ours for all we know? We have our stock examples—the victim whose heart is simultaneously pierced by two

bullets, and the like But those cases seem uncommon Moreover, the overdetermination is not very extreme We have more than one determinant, but still not a very great number Extreme overdetermination of earlier affairs by later ones, on the other hand, may well be more or less universal at a world like ours Whatever goes on leaves widespread and varied traces at future times Most of these traces are so minute or so dispersed or so complicated that no human detective could ever read them, but no matter, so long as they exist It is plausible that very many simultaneous disjoint combinations of traces of any present fact are determinants thereof, there is no lawful way for the combination to have come about in the absence of the fact (Even if a trace could somehow have been faked, traces of the absence of the requisite means of fakery may be included with the trace itself to form a set jointly sufficient for the fact in question) If so, the abundance of future traces makes for a like abundance of future determinants We may reasonably expect overdetermination toward the past on an altogether different scale from the occasional case of mild overdetermination toward the future

That would explain the asymmetry of miracles It takes a miracle to break the link between any determinant and that which it determines Consider our example To diverge from w_0, a world where Nixon presses the button need only break the links whereby certain past conditions determine that he does not press it To converge to w_0, a world where Nixon presses the button must break the links whereby a varied multitude of future conditions vastly overdetermine that he does not press it The more overdetermination, the more links need breaking and the more widespread and diverse must a miracle be if it is to break them all

An asymmetry noted by Popper ([16]) is a special case of the asymmetry of overdetermination There are processes in which a spherical wave expands outward from a point source to infinity The opposite processes, in which a spherical wave contracts inward from infinity and is absorbed, would obey the laws of nature equally well But they never occur A process of either sort exhibits extreme overdetermination in one direction Countless tiny samples of the wave each determine what happens at the space-time point where the wave is emitted or absorbed The processes that occur are the ones in which this extreme overdetermination goes toward the past, not those in which it goes toward the future I suggest that the same is true more generally

Let me emphasize, once more, that the asymmetry of overdetermination is a contingent, *de facto* matter Moreover, it may be a local

matter, holding near here but not in remote parts of time and space If so, then all that rests on it—the asymmetries of miracles, of counterfactual dependence, of causation and openness—may likewise be local and subject to exceptions

I regret that I do not know how to connect the several asymmetries I have discussed and the famous asymmetry of entropy [1]

REFERENCES

[1] Robert M Adams, Theories of Actuality," *Nous* 8(1974) 211–31
[2] Jonathan Bennett, review of Lewis ([10]), *The Canadian Journal of Philosophy* 4(1974) 381–402
[3] G Lee Bowie, 'The Similarity Approach to Counterfactuals Some Problems," *Nous* 13(1979) 477–98
[4] Lewis Creary and Christopher Hill, review of Lewis ([10]), *Philosophy of Science* 42(1975) 341–4
[5] P B Downing, 'Subjunctive Conditionals, Time Order, and Causation," *Proceedings of the Aristotelian Society* 59(1959) 125–40
[6] Kit Fine, review of Lewis ([10]), *Mind* 84(1975) 451–8
[7] Frank Jackson, "A Causal Theory of Counterfactuals," *Australasian Journal of Philosophy* 55(1977) 3–21
[8] David Lewis, "Counterpart Theory and Quantified Modal Logic," *Journal of Philosophy* 65(1968) 113–26
[9] ———, "Anselm and Actuality," *Nous* 4(1970) 175–88
[10] ———, *Counterfactuals* (Oxford Blackwell, 1973)
[11] ———, 'Counterfactuals and Comparative Possibility," *Journal of Philosophical Logic* 2(1973) 418–46
[12] ———, "Causation," *Journal of Philosophy* 70(1973) 556–67, reprinted in Ernest Sosa (ed), *Causation and Conditionals* (London Oxford University Press, 1975)
[13] ———, 'The Paradoxes of Time Travel," *American Philosophical Quarterly* 13(1976) 145–52

[1] I am grateful to many friends for discussion of these matters and especially to Jonathan Bennett, Robert Goble, Philip Kitcher Ernest Loevinsohn, John Perry, Michael Slote, and Robert Stalnaker I am grateful to seminar audiences at several universities in New Zealand for comments on an early version of this paper, and to the New Zealand–United States Educational Foundation for making those seminars possible I also thank Princeton University and the American Council of Learned Societies for research support at earlier stages An earlier version of this paper was presented at the 1976 Annual Conference of the Australasian Association of Philosophy

[14] ———, 'Scorekeeping in a Language Game," *Journal of Philosophical Logic*, 8 (1979) 339–59
[15] Richard Montague, Deterministic Theories," in *Decisions, Values and Groups* (Oxford Pergamon Press, 1962), reprinted in Montague, *Formal Philosophy* (New Haven Yale University Press, 1974)
[16] Karl Popper, "The Arrow of Time," *Nature* 177(1956) 538
[17] Tom Richards, The Worlds of David Lewis," *Australasian Journal of Philosophy* 53(1975) 105–118
[18] Eugene Schlossberger, 'Similarity and Counterfactuals," *Analysis* 38(1978) 80–2
[19] Michael A Slote, "Time in Counterfactuals," *Philosophical Review* 87(1978) 3–27
[20] Pavel Tichy, "A Counterexample to the Stalnaker–Lewis Analysis of Counterfactuals," *Philosophical Studies* 29(1976) 271–73
[21] Joan Weiner, ' Counterfactual Conundrum, *Nous* 13(1979) 499–509

Postscripts to

"Counterfactual Dependence and Time's Arrow"

A NEW THEORY AND OLD?

From time to time I have been told, much to my surprise, that this paper presents a "new theory" of counterfactuals, opposed to the "old theory" I had advanced in earlier writings [1]

I would have thought, rather, that the truth of the matter was as follows In the earlier writings I said that counterfactuals were governed in their truth conditions by comparative overall similarity of worlds, but that there was no one precisely fixed relation of similarity that governed all counterfactuals always To the contrary, the governing similarity relation was both vague and context-dependent Different contexts would select different ranges of similarity relations, probably without ever reaching full determinacy In this paper I reiterate all that

[1] Principally *Counterfactuals* (Oxford Blackwell 1973), also Counterfactuals and Comparative Possibility, first published in 1973 and reprinted in this volume

Then I focus attention on some contexts in particular, and on the range of similarity relations that apply in such contexts Thereby I add to my earlier discussion, but do not at all subtract from it Yet not a few readers think I have taken something back Why?

The trouble seems to be that a comparative relation of the sort that I now put forward—one that turns to some extent on the size of regions of perfect match, and to some extent on the scarcity in one world of events that violate the laws of another—is not at all what my earlier writings led these readers to expect But why not? I think the trouble has three sources

One source, I think, is entrenched doubt about the very idea of similarity It is widely thought that *every* shared property, in the most inclusive possible sense of that word, is *prima facie* a respect of similarity that things can be similar in respect of satisfying the same miscellaneously disjunctive formula, or in respect of belonging to the same utterly miscellaneous class If so, then there's little to be said about comparative similarity Any two things, be they two peas in a pod or be they a raven and a writing-desk, are alike in infinitely many respects and unlike in equally many

Against this scepticism, I observed that we undeniably do make judgments of comparative overall similarity And readers took the point—but in far too limited a way "Yes," I think they thought, "there is indeed a comparative relation that is special in the way it governs our explicit snap judgments We can scarcely doubt that—we have an operational test But leave that firm ground, and we're as much at sea as ever Apart from that one special case, we do not understand how one shared property can be more or less of a similarity-maker than another, or how it can be that some orderings are comparisons of similarity and others aren't " And so I speak of similarity, and these sceptics understand me in the only way they can they seize on the one discrimination they regard as unproblematic, since they can understand how to pick out one similarity relation operationally in terms of snap judgments Then they observe, quite rightly, that the "similarity relation" I now put forward as governing counterfactuals isn't *that* one

The right lesson would have been more far-reaching Our ability to make the snap judgment is one reason, among others, to reject the sceptical, egalitarian orthodoxy It just isn't so that all properties (in the most inclusive sense) are equally respects of similarity Then it is by no means empty to say as I do that a relation of overall similarity is any weighted resultant of respects of similarity and dissimilarity (To

which I add that the weighting might be nonarchimedean, that is, we might have a system of priorities rather than trade-offs) Here we have a class of comparative relations that can go far beyond the one that governs the snap judgments, and that yet falls far short of the class delineated just by the formal character of comparative similarity

Once we reject egalitarianism, what shall we put in its place? An analysis, somehow, of the difference between those properties that are respects of similarity and those that aren't? A primitive distinction? A distinction built into our ontology, in the form of a denial of the very existence of the alleged properties that aren't respects of similarity? A fair question, but one it is risky to take up, lest we put the onus on the wrong side What we know best on this subject, I think, is that egalitarianism is *prima facie* incredible We are entitled to reject it without owing any developed alternative [2]

A second source of trouble, I suspect, is that some readers think of imperfect similarity always as imperfect match, and neglect the case of perfect match over a limited region To illustrate, consider three locomotives 2818, 4018, and 6018 2818 and 4018 are alike in this way they have duplicate boilers, smokeboxes, and fireboxes (to the extent that two of a kind from an early 20th century production line ever are duplicates), and various lesser fittings also are duplicated But 2818 is a slow, small-wheeled, two-cylindered 2–8–0 coal hauler—plenty of pull, little speed—whereas 4018 is the opposite, a fast, large-wheeled, four-cylindered 4–6–0 express passenger locomotive So is 6018, but 6018, unlike 2818, has few if any parts that duplicate the corresponding parts of 4018 (6018 is a scaled-up and modernized version of 4018) Anyone can see the way in which 6018 is more similar to 4018 than 2818 is But I would insist that there is another way of comparing similarity, equally deserving of that name, on which the duplicate standard parts make 2818 the stronger candidate

A third source of trouble may be a hasty step from similarity with respect to laws of nature to similarity *of* the laws—or, I might even say, to similarity of the linguistic codifications of laws Consider three worlds The first has some nice, elegant system of uniform laws The second does not the best way to write down its laws would be to write

[2] In this area I am indebted to Michael Slote, I long ago defended the egalitarian ortho doxy against his good sense, not entirely to my own satisfaction More recently I have benefited from extensive discussions with D M Armstrong, which have helped me to distinguish and relate the question of egalitarianism and the traditional problem of universals For further discussion, see my New Work For a Theory of Universals, *Australasian Journal of Philosophy* 61 (1983) 343–77

down the laws of the first world, then to mutilate them by sticking in clauses to permit various exceptions in an unprincipled fashion Yet almost everything that ever happens in the second world conforms perfectly to the laws of the first The third world does have a nice, elegant, uniform system, its laws resemble those of the first world except for a change of sign here, a switch from inverse square to inverse cube there, and a few other such minor changes Consequently, the third world constantly violates the laws of the first, any little thing that goes on in the third would be prohibited by the laws of the first Focus on the linguistic codification of the laws, and it may well seem that the third world resembles the first with respect to laws far more than the second does But I would insist that there is another way of comparing similarity with respect to laws, equally deserving of that name, on which the second world resembles the first very well, and the third resembles the first very badly That is the way that neglects linguistic codifications, and looks instead at the classes of lawful and of outlawed events

B BIG AND LITTLE MIRACLES

It has often been suggested, not often by well-wishers, that I should distinguish big and little miracles thus big miracles are other-worldly events that break many of the laws that actually obtain, whereas little miracles break only a few laws I think this proposal is thoroughly misguided It is a good thing that I never endorsed it, and a bad thing that I am sometimes said to have endorsed it

Consider two cases (1) By "laws" we might mean *fundamental* laws those regularities that would come out as axioms in a system that was optimal among true systems in its combination of simplicity and strength If the hopes of physics come true, there may be only a few of these fundamental laws altogether Then *no* miracle violates many fundamental laws, *any* miracle violates the Grand Unified Field equation, the Schrodinger equation, or another one of the very few, very sweeping fundamental laws

Or (2) by "laws" we might rather mean fundamental *or derived* laws those regularities that would come out as axioms *or theorems* in an optimal system Then any miracle violates infinitely many laws, and again it doesn't seem that big miracles violate more laws than little ones

It's a blind alley to count the violated laws What to do instead?

Take the laws collectively, distinguish lawful events from unlawful ones (For example, lawful pair-annihilations with radiation from unlawful quiet disappearings of single particles without a trace) In whatever way events can be spread out or localized, unlawful events can be spread out or localized In whatever way several events can be alike or varied, several unlawful events can be alike or varied In whatever way we can distinguish one simple event from many simple events, or from one complex event consisting of many simple parts, we can in particular distinguish one simple unlawful event from many, or from one complex event consisting of many simple unlawful parts A big miracle consists of many little miracles together, preferably not all alike What makes the big miracle more of a miracle is not that it breaks more laws, but that it is divisible into many and varied parts, any one of which is on a par with the little miracle

C WORLDS TO WHICH CONVERGENCE IS EASY

Begin with our base world w_0, the deterministic world something like our own Proceed to w_1, the world which starts out just like w_0, diverges from it by a small miracle, and thereafter evolves in accordance with the laws of w_0 Now extrapolate the later part of w_1 backward in accordance with the laws of w_0 to obtain what I shall call a *Bennett world*[3] This Bennett world is free of miracles, relative to w_0 That is, it conforms perfectly to the laws of w_0, and it seems safe to suppose that these are the laws of the Bennett world also From a certain time onward, the Bennett world and world w_1 match perfectly, which is to say that w_1 converges to the Bennett world Further, this convergence is accomplished by a small miracle namely, the very same small miracle whereby w_1 diverges from w_0 For we had already settled that this small divergence miracle was the only violation by w_1 of the laws of w_0, and those are the same as the laws of the Bennett world Thus the Bennett world is a world to which convergence is easy, since w_1 converges to it by only a small miracle

What then becomes of my asymmetry of miracles? I said that

[3] So-called to acknowledge my indebtedness to Jonathan Bennett who first brought the possibility of such worlds to my attention See his Counterfactuals and Temporal Direction, *Philosophical Review* 93 (1984) 57–91, especially pp 63–64 I am indebted also to David Sanford for helpful correspondence on the subject

' divergence from such a world as w_0 is easier than perfect convergence to it Either takes a miracle but convergence takes very much more of a miracle " To be sure, I said that it might be otherwise for a different sort of world But the Bennett world seems to be a world of the same sort as w_0 After all, it has the very same laws

No Same laws are not enough If there are *de facto* asymmetries of time, not written into the laws, they could be just what it takes to make the difference between a world to which the asymmetry of miracles applies and a world to which it does not, that is, between a world like w_0 (or ours) to which convergence is difficult and a Bennett world to which convergence is easy Consider, for instance, Popper's asymmetry [4] That is not a matter of law, so it could obtain in one and not the other of two worlds with exactly the same laws Likewise in general for the asymmetry of overdetermination

A Bennett world is deceptive After the time of its convergence with w_1, it contains exactly the same apparent traces of its past that w_1 does, and the traces to be found in w_1 are such as to record a past exactly like that of the base world w_0 So the Bennett world is full of traces that seem to record a past like that of w_0 But the past of the Bennett world is not like the past of w_0 under the laws that are common to both worlds, the past of the Bennett world predetermines that Nixon presses the button, whereas the past of w_0 predetermines that he does not Further, we cannot suppose that the two pasts are even close As I noted in discussing world w_2, there is no reason to think that two lawful histories can, before diverging, remain very close throughout a long initial segment of time To constrain a history to be lawful in its own right, and to constrain it also to stay very close to a given lawful history for a long time and then swerve off, is to impose two very strong constraints There is not the slightest reason to think the two constraints are compatible

To be sure, any complete cross section of the Bennett world, taken in full detail, is a truthful record of its past, because the Bennett world is lawful, and its laws are *ex hypothesi* deterministic (in both directions), and any complete cross section of such a world is lawfully sufficient for any other But in a world like w_0, one that manifests the ordinary *de facto* asymmetries, we also have plenty of very *incomplete* cross sections that postdetermine incomplete cross sections at earlier times It is these incomplete postdeterminants that are missing from

[4] Karl Popper, The Arrow of Time, *Nature* 177 (1956) 538

the Bennett world Not throughout its history, but the postdetermination across the time of convergence with w_1 is deficient

Popper's pond is deceptive in just the same way Ripples rise around the edge, they contract inward and get higher, when they reach the center a stone flies out of the water—and then the pond is perfectly calm What has happened is the time-reversed mirror image of what ordinarily happens when a stone falls into a pond It is no less lawful, the violated asymmetries are not a matter of law There would be no feasible way to detect what had happened For there would be no trace on the water of its previous agitation, and the rock would be dry, the air would bear no sound of a splash, the nearby light would bear no tell-tale image, In short, a perfect cover-up job—and without any miracle! But not in a world like w_0, and not in a world like ours To be sure, if the laws are deterministic, the event is postdetermined by any complete cross section afterward But we lack the usual abundance of lesser postdeterminants

D THE INDETERMINISTIC CASE

I assumed determinism for the sake of the argument I considered the deterministic case in order to oppose the view that the asymmetries under consideration arise out of one-way indeterministic branching

That is not to say, of course, that I assume determinism *simpliciter* I do not Accepted physics, after all, is not deterministic It is hard to know what to make of the indeterminism in present-day quantum mechanics *Pace* Einstein, indeterminism *per se* is credible enough But the trouble is that the only indeterministic process in nature—reduction of the wave function, as opposed to Schrodinger evolution—is supposed to be special to the phenomenon of measurement [5] And the

[5] For a forthright account of the predicament, see Eugene P Wigner, *Symmetries and Reflections Scientific Essays of Eugene P Wigner* (Bloomington Indiana University Press 1967) Chapters 12–14 The hypothesis that measurement reduces the wave function comes from treatments of measurement by John von Neumann, *Mathematische Grundlagen der Quantenmechanik* (Berlin Springer, 1932) and by Fritz London and Edmond Bauer, *La theorie de l observation en mecanique quantique* (Paris Hermann & Cie , 1939) Prospects for a way out are surveyed in Abner Shimony, Role of the Observer in Quantum Theory, *American Journal of Physics* 31 (1963) 755–73, and in Nancy Cartwright, How the Measurement Problem is an Artefact of the Mathematics, in her *How the Laws of Physics Lie* (Oxford Clarendon Press, 1983)

idea that a unique microphysical process takes place when a person makes a measurement seems about as credible as the idea that a unique kind of vibration takes place when two people fall truly in love. Instrumentalist philosophy among physicists doesn't help matters, though perhaps the quantum theory of measurement is such a disaster that it *deserves* to be dismissed as a mere instrument. Which parts of present theory are fact, which fiction? What will remain when the dust settles?

I can only guess, my guess is not especially well informed, but for what it is worth, I guess as follows. The theoretical foundation of quantum mechanics is probably wrong to say that reduction is brought on when people measure. But the working quantum mechanics of radioactive decay, coherent solids, chemical bonding, and the like can somehow stand on its own. It does not need this unfortunate anthropocentric foundation.[6] Then the laws of nature that govern our world really are indeterministic. Whatever we make of the reduction of the wave function supposedly brought on by measurement, at any rate there are chance processes involved in radioactive decay, in the making and breaking of chemical bonds, in ionization, in the radiation of light and heat, and so on. These processes are pervasive. So much so that not only is the world as a whole indeterministic, but also it can contain few if any deterministic enclaves.

If so, then what becomes of my asymmetries? In one way, the problem is easier. Divergence no longer requires a small miracle, not if there are abundant opportunities for divergence in the outcomes of chance processes. (If indeterministic processes were very scarce, miracles might still be required sometimes. But that is probably not the case for our world.) So in the indeterministic case it does not matter whether I am right to count small miracles as relatively cheap dissimilarities. Our divergences come cheaper still.

The thing to say about approximate convergence remains the same. Even if approximate convergence is cheap—and even if it is cheaper still when it can be had without even a little miracle—still we can say that it counts for little or nothing, so it is not so that if Nixon had

[6] Here I follow the lead of Nancy Cartwright, *How the Laws of Physics Lie*, in suggesting that scepticism about alleged fundamental laws—in this case, the projection postulate, which says that measurement reduces the wave function—may perfectly well be combined with staunch realism about the unobservable objects and processes to which the doubted law is supposed to apply, and with acceptance of less fundamental laws about these objects and processes. However, I do not follow Cartwright in her general scepticism about fundamental laws of physics. I take the projection postulate to be a special case, a sick spot in the midst of general good health.

pressed the button there would have been approximate convergence to our world, and no holocaust

But what about perfect convergence? Here, indeterminism makes my problem harder It is not to be said that the similarity achieved by perfect convergence counts for little or nothing For it is just like the past similarity that has decisive weight in the deterministic case, tilting the balance in favor of last-minute divergence instead of difference throughout the past I said that perfect convergence would take a big, widespread, varied miracle—a miraculously perfect cover-up job But if chance processes are abundant, as I have guessed that they are, why couldn't they accomplish the cover-up? Why couldn't convergence happen without any miracles at all, simply by the right pattern of lawful outcomes of many different chance processes? Call such a pattern a *quasi-miracle* It is extraordinarily improbable, no doubt, but it does not violate the laws of nature that prevail at our world

What must be said, I think, is that a quasi-miracle to accomplish perfect convergence, though it is entirely lawful, nevertheless detracts from similarity in much the same way that a convergence miracle does That seems plausible enough (Though the test of the hypothesis is not in its offhand plausibility, but its success in yielding the right counterfactuals) The quasi-miracle would be such a remarkable coincidence that it would be quite unlike the goings-on we take to be typical of our world Like a big genuine miracle, it makes a tremendous difference from our world Therefore it is not something that happens in the closest worlds to ours where Nixon presses the button These worlds have no convergence miracles, and also no convergence quasi-miracles So the case turns out as it should the closest worlds where Nixon presses the button are worlds where a holocaust ensues

My point is not that quasi-miracles detract from similarity because they are so very improbable They are, but ever so many unremarkable things that actually happen, and ever so many other things that might happen under various counterfactual suppositions, are likewise very improbable What makes a quasi-miracle is not improbability *per se*, but rather the remarkable way in which the chance outcomes seem to conspire to produce a pattern If the monkey at the typewriter produces a 950-page dissertation on the varieties of anti-realism, that is at least somewhat quasi-miraculous, the chance keystrokes happen to simulate the traces that would have been left by quite a different process If the monkey instead types 950 pages of jumbled letters, that is not at all quasi-miraculous But, given suitable assumptions about what sort of chance device the monkey is, the one text is exactly as

improbable as the other (It is irrelevant to compare the probability that there will be *some* dissertation with the probability that there will be *some* jumble—the monkey does not just select one or the other kind of text, but also produces a particular text of the selected kind) The pattern of systematic falsification of traces required for perfect convergence is quasi-miraculous in the same way

(What if, contrary to what we believe, our own world is full of quasi-miracles? Then other-worldly quasi-miracles would not make other worlds dissimilar to ours But if so, we would be very badly wrong about our world, so why should we not turn out to be wrong also about which counterfactuals it makes true? I say that the case needn't worry us Let it fall where it may)

In the deterministic case, the asymmetry of counterfactuals derives from an asymmetry of miracles divergence takes less of a miracle than (perfect) convergence Likewise in the indeterministic case we have an asymmetry of quasi-miracles Convergence to an indeterministic world of the sort that ours might be takes a quasi-miracle, divergence from such a world does not (I do not speak of small quasi-miracles, what corresponds to a small miracle in the deterministic case is a perfectly commonplace chance occurrence) The asymmetry is made plausible by the same thought-experiment as before think, in some detail and without neglecting imperceptible differences, of what would be needed for a perfect cover-up

The trouble with using quasi-miracles as a weighty respect of dissimilarity is that it seems to prove too much, more than is true For if quasi-miracles make enough of a dissimilarity to outweigh perfect match throughout the future, and if I am right that counterfactuals work by similarity, then we can flatly say that if Nixon had pressed the button there would have been no quasi-miracle But if chance processes are abundant, and would have been likewise abundant if Nixon had pressed the button, then in that case there would have been *some* chance of a quasi-miracle To be sure, the probability would have been very low indeed But it would not have been zero

But if there would have been some minute probability of a quasi-miracle, does it not follow that there might have been one? And if there might have been one, then is it not false to say that there would not have been one? True, it would have been overwhelmingly probable that there not be one But may we say flatly that this improbable thing would not have happened?

(Note that I am not talking about probabilities that certain counterfactuals are true Rather, the consequents of the counterfactuals have to

do with probabilities. In particular, they have to do with objective single-case chances, as of the time right after the hypothetical pressing, of the patterns of events that would comprise a suitable quasi-miracle.[7])

Is there, perhaps, an exact balance? Suppose that perfect match throughout the future contributes to similarity exactly as much as the quasi-miracle needed to achieve that match detracts from similarity. Then worlds with a quasi-miraculous convergence have no net advantage, and no net disadvantage. Then they can be some, but not all, of the closest worlds where Nixon pressed the button. That seems to give the right counterfactuals: it is not so that if he had pressed the button then there *would* have been quasi-miraculous convergence, and such convergence would not have been at all probable, but it is so that if Nixon had pressed the button then there *might* have been quasi-miraculous convergence. So far, so good. But this solution (besides seeming artificial) fails to solve the whole problem. What about other quasi-miracles: patterns of outcomes of chance processes that are just as much remarkable coincidences, just as improbable, just as dissimilar from what typically goes on at our world—but do not yield convergence? On the balance hypothesis, these *non-convergence quasi-miracles* detract greatly from similarity and bring no compensating gain. So they, unlike the convergence quasi-miracles, are not to be found at any of the closest worlds where Nixon had pressed the button. And that seems wrong. It seems that we should say the same thing about *any* quasi-miracle, whether or not it yields convergence: if Nixon had pressed the button, it would have had some minute probability of happening, hence if so it might have happened, hence we should not say flatly that it would not have happened. So the hypothesis of exact balance does not save the day and I am still in trouble.

The line of retreat, of course, is asymmetry by fiat. Analysis 1, which drops the whole idea that counterfactuals work by similarity, is still available. It has no need of determinism. Or we could complicate the weighting of respects of similarity so that perfect match in the past weighs heavily but perfect match in the future counts for nothing. (More precisely: perfect match before and after the time relevant to the counterfactual supposition in question—as it might be, the time of Nixon's supposed pressing of the button.) Either way, we build an asymmetry between the directions of time into our very analysis of counterfactual, and hence on my view causal, dependence: just *consists*

[7] See "A Subjectivist's Guide to Objective Chance", "Causal Decision Theory", and Postscript B to "Causation", all in this volume.

in part of temporal order I still say that won't do It imposes *a priori* answers on questions that ought to be empirical No, the asymmetry of counterfactual dependence should come from a symmetrical analysis and an asymmetrical world

What is to be done? Our trouble was caused by an apparent logical connection between counterfactuals about what would happen, counterfactuals about what might happen, and counterfactuals about what the chances would be One escape route is to reconsider that connection Indeed, the connection seemed intuitively right, and I would be reluctant to challenge it just as a cure for my present trouble But it needs challenging also for other reasons

Recall the problem By treating quasi-miracles as a weighty respect of dissimilarity, I make it turn out that there are no quasi-miracles of any kind, and hence there is no quasi-miraculous convergence, at any of the most similar worlds where Nixon pressed the button That means that

(1) If Nixon had pressed the button, there would not have been a quasi-miracle

But quasi-miracles are just certain special patterns of outcomes of chance processes, and the chances would have been much the same if Nixon had pressed the button That means that

(2) If Nixon had pressed the button, there would have been some minute chance of a quasi-miracle

We had better accept both (1) and (2) But they seem to conflict Or do they? Considered by themselves, there is no very clear impression of conflict Above, to create a semblance of conflict, I went in two steps, by way of

(3) If Nixon had pressed the button, there might have been a quasi-miracle

Whether or not (1) and (2) conflict, it certainly seems that (1) and (3) conflict, and it also seems that (2) implies (3) But perhaps we are being fooled by an ambiguity in (3)

I have hitherto advocated a "not-would-not" reading of "might" counterfactuals, on which (3) comes out as

(3-nwn) It is not the case that if Nixon had pressed the button, there would not have been a quasi-miracle

But perhaps there is also a "would-be-possible" reading, on which (3) comes out as

(3-wbp) If Nixon had pressed the button, it would be that a quasi-miracle is possible

The readings differ as follows (3-nwn) means that some of the most similar worlds where Nixon pressed the button are worlds where a quasi-miracle happens, whereas (3-wbp) means that all of them are worlds where it is possible for a quasi-miracle to happen If all of them are worlds where there is an unfulfilled possibility of a quasi-miracle, that makes (3-nwn) false and (3-wbp) true And if we take possibility to mean non-zero chance (as of the time of the pressing), then that is exactly the situation that makes (1) and (2) both true together Indeed, (1) conflicts with (3-nwn), indeed, (2) implies (3-wbp), but (1) and (2) are compatible [8]

I note that the same problem arises in consequence of my treatment of counterfactuals with true antecedents Suppose that our world is an A-world with an unfulfilled non-zero chance of B Then, since a counterfactual with a true antecedent is true iff its consequent is,[9] we have a pair of true counterfactuals that parallel (1) and (2)

(4) If it were that A, then it would be that not B

(5) If it were that A, then there would be some chance that B

Thus (4) and (5), on my account, are compatible Yet they may appear to conflict if we consider

(6) If it were that A, then it might be that B

This counterfactual seems to conflict with (4) and to be implied by (5) I reply that on the "not-would-not" reading (6) conflicts with (4) and is false, whereas on the "would-be-possible" reading it is implied by (5) and is true

[8] Compare Robert Stalnaker s discussion of might counterfactuals in his A Defense of Conditional Excluded Middle in *Ifs* ed by William Harper, Robert Stalnaker and Glenn Pearce (Dordrecht Reidel 1980) He recognizes a range of different senses for might counterfactuals Some of these are epistemic senses irrelevant to our present concerns But he does admit my not would not reading, though as quasi epistemic, and it seems that he would also admit my would-be-possible reading though as doubly abnormal because the might neither is epistemic nor has wide scope

[9] This follows from my assumption of centering see Section 1 of Counterfactuals and Comparative Possibility in this volume

In fact, our problem is more far-reaching still. If we want any kind of similarity theory of counterfactuals, we dare not treat "there would be some chance of it" and "it would not happen" in general as incompatible. Suppose for *reductio* that counterfactuals of these two kinds are in general incompatible. Let C be any proposition that might obtain or not as a matter of chance, let u and v be a C-world and a not-C-world, respectively, but let them both be worlds that have a chance of going either way, let A be the proposition that holds at these two worlds, and no others, and let w be any third world. It is true at w that if A, there would be some chance that C, so by the supposed incompatibility, it is false at w that if A, it would be that not-C, so u must be at least as close to w as v is. Likewise, putting not-C in place of C, v must be at least as close to w as u is. That is, worlds u and v are tied in closeness to any third world. But u and v were *any* two worlds that differ in respect of the outcome of a matter of chance—no matter how much they may differ in other ways as well! This completes the *reductio*.[10]

We can have a simpler *reductio* if we suppose that it is legitimate to mention chances in the antecedent of a counterfactual—and how can that fail to be legitimate, if chances are indeed an objective feature of the world? What would be the case if there were an unfulfilled chance of C? If so, then there would be some chance that C. But if so, then also it would not be that C. So here we have a counterexample to the supposed incompatibility, just on the principle that a counterfactual holds when the antecedent implies the consequent.

So the supposed incompatibility had better be rejected. The reconciliation of (1) with (2), (4) with (5), and the like is by no means just a dodge to defend my controversial views about time's arrow and about counterfactuals with true antecedents. But it does serve, *inter alia*, to defend them. We can count quasi-miracles as weighty dissimilarities from actuality, we can persuade ourselves by examples that perfect reconvergence to a world like ours would require, if not a big miracle, at least a quasi-miracle, we can conclude that if Nixon had pressed the button, there would have been no perfect convergence, and still we can say, as we should, that there would have been some minute chance of perfect convergence.

[10] Pavel Tichy raises the problem of chance and future similarity in "A Counterexample to the Stalnaker-Lewis Theory of Counterfactuals," *Philosophical Studies* 29 (1976): 271–73. His example is ineffective however. It can be met simply by denying that imperfect match counts toward similarity, and thus it serves to support that denial.

E UBIQUITOUS TRACES AND COMMON KNOWLEDGE

My argument for an asymmetry of miracles (or of quasi-miracles) relied on an empirical premise: at a world like ours, everything that happens leaves many and varied traces, so that it would take a big miracle—equivalently, many and varied small miracles working together—to eradicate those traces and achieve reconvergence. But I need more than merely the truth of that premise. I need common knowledge of it. For if the premise were true but generally disbelieved, and if our counterfactuals work as I say they do, then we ought to find people often accepting the counterfactuals that would be true on my account if that premise were false. We ought to find them saying that if Nixon had pressed the button, the future would have been no different, there would have been convergence and no holocaust. In illustrating the multitude of traces that the pressing would have left, and the difficulty of a perfect cover-up, I relied on a certain amount of scientific knowledge that many people do not share. I may have explained why the right counterfactuals come out true according to my beliefs. But I have done nothing to explain why ignorant folk accept those same counterfactuals.

I reply that everyone believes in ubiquity of traces. Maybe not everyone can illustrate the point in the way I did (though I must say that I did not use anything very esoteric) but they can still think that *somehow* everything leaves many and varied traces.

Consider detective stories. Seldom are they written by, or for, expert scientists. The background against which they are to be read is common knowledge, not expert knowledge. And part of that background is the assumption that events leave many and varied traces. Else the plots would not make sense. We are supposed to marvel at the skill of the detective in spotting and reading the traces. We are not supposed to marvel that the traces are there at all. Ignorant or expert, anyone knows better than to read the tale as a hard-luck story: how the criminal was caught because he was especially unfortunate in leaving traces. And anyone knows better than to read the tale as science fiction: how things would be in a bizarre world where things leave far more traces than they do in ours. No, it is supposed to be a tale of a world like ours, and the ubiquity of traces is part of the likeness.

EIGHTEEN

The Paradoxes of Time Travel

Time travel, I maintain, is possible The paradoxes of time travel are oddities, not impossibilities They prove only this much, which few would have doubted that a possible world where time travel took place would be a most strange world, different in fundamental ways from the world we think is ours

I shall be concerned here with the sort of time travel that is recounted in science fiction Not all science fiction writers are clear-headed, to be sure, and inconsistent time travel stories have often been written But some writers have thought the problems through with great care, and their stories are perfectly consistent [1]

If I can defend the consistency of some science fiction stories of time travel, then I suppose parallel defenses might be given of some controversial physical hypotheses, such as the hypothesis that time is circular or the hypothesis that there are particles that travel faster than light But I shall not explore these parallels here

What is time travel? Inevitably, it involves a discrepancy between time and time Any traveler departs and then arrives at his destination,

[1] I have particularly in mind two of the time travel stories of Robert A Heinlein By His Bootstraps, in R A Heinlein, *The Menace from Earth* (Hicksville, N Y , 1959), and —All You Zombies—, in R A Heinlein, *The Unpleasant Profession of Jonathan Hoag* (Hicksville, N Y , 1959)

the time elapsed from departure to arrival (positive, or perhaps zero) is the duration of the journey But if he is a time traveler, the separation in time between departure and arrival does not equal the duration of his journey He departs, he travels for an hour, let us say, then he arrives The time he reaches is not the time one hour after his departure It is later, if he has traveled toward the future, earlier, if he has traveled toward the past If he has traveled far toward the past, it is earlier even than his departure How can it be that the same two events, his departure and his arrival, are separated by two unequal amounts of time?

It is tempting to reply that there must be two independent time dimensions, that for time travel to be possible, time must be not a line but a plane [2] Then a pair of events may have two unequal separations if they are separated more in one of the time dimensions than in the other The lives of common people occupy straight diagonal lines across the plane of time, sloping at a rate of exactly one hour of $time_1$ per hour of $time_2$ The life of the time traveler occupies a bent path, of varying slope

On closer inspection, however, this account seems not to give us time travel as we know it from the stories When the traveler revisits the days of his childhood, will his playmates be there to meet him? No, he has not reached the part of the plane of time where they are He is no longer separated from them along one of the two dimensions of time, but he is still separated from them along the other I do not say that two-dimensional time is impossible, or that there is no way to square it with the usual conception of what time travel would be like Nevertheless I shall say no more about two-dimensional time Let us set it aside, and see how time travel is possible even in one-dimensional time

The world—the time traveler's world, or ours—is a four-dimensional manifold of events Time is one dimension of the four, like the spatial dimensions except that the prevailing laws of nature discriminate between time and the others—or rather, perhaps, between various timelike dimensions and various spacelike dimensions (Time remains one-dimensional, since no two timelike dimensions are orthogonal) Enduring things are timelike streaks wholes composed of temporal parts, or *stages*, located at various times and places Change is qualitative difference between different stages—different temporal parts—of some enduring thing, just as a "change" in scenery from east to west is

[2] Accounts of time travel in two-dimensional time are found in Jack W Meiland A Two-Dimensional Passage Model of Time for Time Travel, *Philosophical Studies*, vol 26 (1974), pp 153–173 and in the initial chapters of Isaac Asimov *The End of Eternity* (Garden City, N Y , 1955) Asimov s denouement, however, seems to require some different conception of time travel

a qualitative difference between the eastern and western spatial parts of the landscape If this paper should change your mind about the possibility of time travel, there will be a difference of opinion between two different temporal parts of you, the stage that started reading and the subsequent stage that finishes

If change is qualitative difference between temporal parts of something, then what doesn't have temporal parts can't change For instance, numbers can't change, nor can the events of any moment of time, since they cannot be subdivided into dissimilar temporal parts (We have set aside the case of two-dimensional time, and hence the possibility that an event might be momentary along one time dimension but divisible along the other) It is essential to distinguish change from "Cambridge change," which can befall anything Even a number can "change" from being to not being the rate of exchange between pounds and dollars Even a momentary event can "change" from being a year ago to being a year and a day ago, or from being forgotten to being remembered But these are not genuine changes Not just any old reversal in truth value of a time-sensitive sentence about something makes a change in the thing itself

A time traveler, like anyone else, is a streak through the manifold of space-time, a whole composed of stages located at various times and places But he is not a streak like other streaks If he travels toward the past he is a zig-zag streak, doubling back on himself If he travels toward the future, he is a stretched-out streak And if he travels either way instantaneously, so that there are no intermediate stages between the stage that departs and the stage that arrives and his journey has zero duration, then he is a broken streak

I asked how it could be that the same two events were separated by two unequal amounts of time, and I set aside the reply that time might have two independent dimensions Instead I reply by distinguishing time itself, *external time* as I shall also call it, from the *personal time* of a particular time traveler roughly, that which is measured by his wristwatch His journey takes an hour of his personal time, let us say, his wristwatch reads an hour later at arrival than at departure But the arrival is more than an hour after the departure in external time, if he travels toward the future, or the arrival is before the departure in external time (or less than an hour after), if he travels toward the past

That is only rough I do not wish to define personal time operationally, making wristwatches infallible by definition That which is measured by my own wristwatch often disagrees with external time, yet I am no time traveler, what my misregulated wristwatch measures

is neither time itself nor my personal time Instead of an operational definition, we need a functional definition of personal time it is that which occupies a certain role in the pattern of events that comprise the time traveler's life If you take the stages of a common person, they manifest certain regularities with respect to external time Properties change continuously as you go along, for the most part, and in familiar ways First come infantile stages Last come senile ones Memories accumulate Food digests Hair grows Wristwatch hands move If you take the stages of a time traveler instead, they do not manifest the common regularities with respect to external time But there is one way to assign coordinates to the time traveler's stages, and one way only (apart from the arbitrary choice of a zero point), so that the regularities that hold with respect to this assignment match those that commonly hold with respect to external time With respect to the correct assignment properties change continuously as you go along, for the most part, and in familiar ways First come infantile stages Last come senile ones Memories accumulate Food digests Hair grows Wristwatch hands move The assignment of coordinates that yields this match is the time traveler's personal time It isn't really time, but it plays the role in his life that time plays in the life of a common person It's enough like time so that we can—with due caution—transplant our temporal vocabulary to it in discussing his affairs We can say without contradiction, as the time traveler prepares to set out, "Soon he will be in the past " We mean that a stage of him is slightly later in his personal time, but much earlier in external time, than the stage of him that is present as we say the sentence

We may assign locations in the time traveler's personal time not only to his stages themselves but also to the events that go on around him Soon Caesar will die, long ago, that is, a stage slightly later in the time traveler's personal time than his present stage, but long ago in external time, is simultaneous with Caesar's death We could even extend the assignment of personal time to events that are not part of the time traveler's life, and not simultaneous with any of his stages If his funeral in ancient Egypt is separated from his death by three days of external time and his death is separated from his birth by three score years and ten of his personal time, then we may add the two intervals and say that his funeral follows his birth by three score years and ten and three days of *extended personal time* Likewise a bystander might truly say, three years after the last departure of another famous time traveler, that "he may even now—if I may use the phrase—be wandering on some plesiosaurus-haunted oolitic coral reef, or beside the

lonely saline seas of the Triassic Age "[3] If the time traveler does wander on an oolitic coral reef three years after his departure in his personal time, then it is no mistake to say with respect to his extended personal time that the wandering is taking place "even now"

We may liken intervals of external time to distances as the crow flies, and intervals of personal time to distances along a winding path The time traveler's life is like a mountain railway The place two miles due east of here may also be nine miles down the line, in the west-bound direction Clearly we are not dealing here with two independent dimensions Just as distance along the railway is not a fourth spatial dimension, so a time traveler's personal time is not a second dimension of time How far down the line some place is depends on its location in three-dimensional space, and likewise the location of events in personal time depend on their locations in one-dimensional external time

Five miles down the line from here is a place where the line goes under a trestle, two miles further is a place where the line goes over a trestle, these places are one and the same The trestle by which the line crosses over itself has two different locations along the line, five miles down from here and also seven In the same way, an event in a time traveler's life may have more than one location in his personal time If he doubles back toward the past, but not too far, he may be able to talk to himself The conversation involves two of his stages, separated in his personal time but simultaneous in external time The location of the conversation in personal time should be the location of the stage involved in it But there are two such stages, to share the locations of both, the conversation must be assigned two different locations in personal time

The more we extend the assignment of personal time outwards from the time traveler's stages to the surrounding events, the more will such events acquire multiple locations It may happen also, as we have already seen, that events that are not simultaneous in external time will be assigned the same location in personal time—or rather, that at least one of the locations of one will be the same as at least one of the locations of the other So extension must not be carried too far, lest the location of events in extended personal time lose its utility as a means of keeping track of their roles in the time traveler's history

A time traveler who talks to himself, on the telephone perhaps,

[3] H G Wells, *The Time Machine An Invention* (London 1895), epilogue The passage is criticized as contradictory in Donald C Williams, The Myth of Passage *The Journal of Philosophy*, vol 48 (1951) p 463

looks for all the world like two different people talking to each other It isn't quite right to say that the whole of him is in two places at once, since neither of the two stages involved in the conversation is the whole of him, or even the whole of the part of him that is located at the (external) time of the conversation What's true is that he, unlike the rest of us, has two different complete stages located at the same time at different places What reason have I, then, to regard him as one person and not two? What unites his stages, including the simultaneous ones, into a single person? The problem of personal identity is especially acute if he is the sort of time traveler whose journeys are instantaneous, a broken streak consisting of several unconnected segments Then the natural way to regard him as more than one person is to take each segment as a different person No one of them is a time traveler, and the peculiarity of the situation comes to this all but one of these several people vanish into thin air, all but another one appear out of thin air, and there are remarkable resemblances between one at his appearance and another at his vanishing Why isn't that at least as good a description as the one I gave, on which the several segments are all parts of one time traveler?

I answer that what unites the stages (or segments) of a time traveler is the same sort of mental, or mostly mental, continuity and connectedness that unites anyone else The only difference is that whereas a common person is connected and continuous with respect to external time, the time traveler is connected and continuous only with respect to his own personal time Taking the stages in order, mental (and bodily) change is mostly gradual rather than sudden, and at no point is there sudden change in too many different respects all at once (We can include position in external time among the respects we keep track of, if we like It may change discontinuously with respect to personal time if not too much else changes discontinuously along with it) Moreover, there is not too much change altogether Plenty of traits and traces last a lifetime Finally, the connectedness and the continuity are not accidental They are explicable, and further, they are explained by the fact that the properties of each stage depend causally on those of the stages just before in personal time, the dependence being such as tends to keep things the same [4]

To see the purpose of my final requirement of causal continuity, let

[4] I discuss the relation between personal identity and mental connectedness and continuity at greater length in Survival and Identity in *The Identities of Persons* ed by Amelie Rorty (Berkeley and Los Angeles, 1976)

us see how it excludes a case of counterfeit time travel Fred was created out of thin air, as if in the midst of life, he lived a while, then died He was created by a demon, and the demon had chosen at random what Fred was to be like at the moment of his creation Much later someone else, Sam, came to resemble Fred as he was when first created At the very moment when the resemblance became perfect, the demon destroyed Sam Fred and Sam together are very much like a single person a time traveler whose personal time starts at Sam's birth, goes on to Sam's destruction and Fred's creation, and goes on from there to Fred's death Taken in this order, the stages of Fred-*cum*-Sam have the proper connectedness and continuity But they lack causal continuity, so Fred-*cum*-Sam is not one person and not a time traveler Perhaps it was pure coincidence that Fred at his creation and Sam at his destruction were exactly alike, then the connectedness and continuity of Fred-*cum*-Sam across the crucial point are accidental Perhaps instead the demon remembered what Fred was like, guided Sam toward perfect resemblance, watched his progress, and destroyed him at the right moment Then the connectedness and continuity of Fred-*cum*-Sam has a causal explanation, but of the wrong sort Either way, Fred's first stages do not depend causally for their properties on Sam's last stages So the case of Fred and Sam is rightly disqualified as a case of personal identity and as a case of time travel

We might expect that when a time traveler visits the past there will be reversals of causation You may punch his face before he leaves, causing his eye to blacken centuries ago Indeed, travel into the past necessarily involves reversed causation For time travel requires personal identity—he who arrives must be the same person who departed That requires causal continuity, in which causation runs from earlier to later stages in the order of personal time But the orders of personal and external time disagree at some point, and there we have causation that runs from later to earlier stages in the order of external time Elsewhere I have given an analysis of causation in terms of chains of counterfactual dependence, and I took care that my analysis would not rule out causal reversal *a priori* [5] I think I can argue (but not here) that under my analysis the direction of counterfactual dependence and causation is governed by the direction of other *de facto* asymmetries of time If so, then reversed causation and time travel are not excluded altogether, but can occur only where there are local exceptions to these

[5] Causation, *The Journal of Philosophy*, vol 70 (1973), pp 556–567, the analysis relies on the analysis of counterfactuals given in my *Counterfactuals* (Oxford, 1973)

asymmetries As I said at the outset, the time traveler's world would be a most strange one

Stranger still, if there are local—but only local—causal reversals, then there may also be causal loops closed causal chains in which some of the causal links are normal in direction and others are reversed (Perhaps there must be loops if there is reversal, I am not sure) Each event on the loop has a causal explanation, being caused by events elsewhere on the loop That is not to say that the loop as a whole is caused or explicable It may not be Its inexplicability is especially remarkable if it is made up of the sort of causal processes that transmit information Recall the time traveler who talked to himself He talked to himself about time travel, and in the course of the conversation his older self told his younger self how to build a time machine That information was available in no other way His older self knew how because his younger self had been told and the information had been preserved by the causal processes that constitute recording, storage, and retrieval of memory traces His younger self knew, after the conversation, because his older self had known and the information had been preserved by the causal processes that constitute telling But where did the information come from in the first place? Why did the whole affair happen? There is simply no answer The parts of the loop are explicable, the whole of it is not Strange! But not impossible, and not too different from inexplicabilities we are already inured to Almost everyone agrees that God, or the Big Bang, or the entire infinite past of the universe or the decay of a tritium atom, is uncaused and inexplicable Then if these are possible, why not also the inexplicable causal loops that arise in time travel?

I have committed a circularity in order not to talk about too much at once, and this is a good place to set it right In explaining personal time, I presupposed that we were entitled to regard certain stages as comprising a single person Then in explaining what united the stages into a single person, I presupposed that we were given a personal time order for them The proper way to proceed is to define personhood and personal time simultaneously, as follows Suppose given a pair of an aggregate of person-stages, regarded as a candidate for personhood, and an assignment of coordinates to those stages, regarded as a candidate for his personal time Iff the stages satisfy the conditions given in my circular explanation with respect to the assignment of coordinates, then both candidates succeed the stages do comprise a person and the assignment is his personal time

I have argued so far that what goes on in a time travel story may be a

possible pattern of events in four-dimensional space-time with no extra time dimension, that it may be correct to regard the scattered stages of the alleged time traveler as comprising a single person, and that we may legitimately assign to those stages and their surroundings a personal time order that disagrees sometimes with their order in external time Some might concede all this, but protest that the impossibility of time travel is revealed after all when we ask not what the time traveler *does*, but what he *could do* Could a time traveler change the past? It seems not the events of a past moment could no more change than numbers could Yet it seems that he would be as able as anyone to do things that would change the past if he did them If a time traveler visiting the past both could and couldn't do something that would change it, then there cannot possibly be such a time traveler

Consider Tim He detests his grandfather, whose success in the munitions trade built the family fortune that paid for Tim's time machine Tim would like nothing so much as to kill Grandfather, but alas he is too late Grandfather died in his bed in 1957, while Tim was a young boy But when Tim has built his time machine and traveled to 1920, suddenly he realizes that he is not too late after all He buys a rifle, he spends long hours in target practice, he shadows Grandfather to learn the route of his daily walk to the munitions works, he rents a room along the route, and there he lurks, one winter day in 1921, rifle loaded, hate in his heart, as Grandfather walks closer, closer,

Tim can kill Grandfather He has what it takes Conditions are perfect in every way the best rifle money could buy, Grandfather an easy target only twenty yards away, not a breeze, door securely locked against intruders, Tim a good shot to begin with and now at the peak of training, and so on What's to stop him? The forces of logic will not stay his hand! No powerful chaperone stands by to defend the past from interference (To imagine such a chaperone, as some authors do, is a boring evasion, not needed to make Tim's story consistent) In short, Tim is as much able to kill Grandfather as anyone ever is to kill anyone Suppose that down the street another sniper, Tom, lurks waiting for another victim, Grandfather's partner Tom is not a time traveler, but otherwise he is just like Tim same make of rifle, same murderous intent, same everything We can even suppose that Tom, like Tim, believes himself to be a time traveler Someone has gone to a lot of trouble to deceive Tom into thinking so There's no doubt that Tom can kill his victim, and Tim has everything going for him that Tom does By any ordinary standards of ability, Tim can kill Grandfather

Tim cannot kill grandfather Grandfather lived, so to kill him

would be to change the past But the events of a past moment are not subdivisible into temporal parts and therefore cannot change Either the events of 1921 timelessly do include Tim's killing of Grandfather, or else they timelessly don't We may be tempted to speak of the "original" 1921 that lies in Tim's personal past, many years before his birth, in which Grandfather lived, and of the "new" 1921 in which Tim now finds himself waiting in ambush to kill Grandfather But if we do speak so, we merely confer two names on one thing The events of 1921 are doubly located in Tim's (extended) personal time, like the trestle on the railway, but the "original" 1921 and the "new" 1921 are one and the same If Tim did not kill Grandfather in the "original" 1921, then if he does kill Grandfather in the "new" 1921, he must both kill and not kill Grandfather in 1921—in the one and only 1921, which is both the "new" and the "original" 1921 It is logically impossible that Tim should change the past by killing Grandfather in 1921 So Tim cannot kill Grandfather

Not that past moments are special, no more can anyone change the present or the future Present and future momentary events no more have temporal parts than past ones do You cannot change a present or future event from what it was originally to what it is after you change it What you *can* do is to change the present or the future from the unactualized way they would have been without some action of yours to the way they actually are But that is not an actual change not a difference between two successive actualities And Tim can certainly do as much, he changes the past from the unactualized way it would have been without him to the one and only way it actually is To "change" the past in this way, Tim need not do anything momentous, it is enough just to be there, however unobtrusively

You know, of course, roughly how the story of Tim must go on if it is to be consistent he somehow fails Since Tim didn't kill Grandfather in the "original" 1921, consistency demands that neither does he kill Grandfather in the "new" 1921 Why not? For some commonplace reason Perhaps some noise distracts him at the last moment, perhaps he misses despite all his target practice, perhaps his nerve fails, perhaps he even feels a pang of unaccustomed mercy His failure by no means proves that he was not really able to kill Grandfather We often try and fail to do what we are able to do Success at some tasks requires not only ability but also luck, and lack of luck is not a temporary lack of ability Suppose our other sniper, Tom, fails to kill Grandfather's partner for the same reason, whatever it is, that Tim fails to kill Grandfather It does not follow that Tom was unable to No more does it follow in Tim's case that he was unable to do what he did not succeed in doing

We have this seeming contradiction "*Tim doesn't, but can, because he has what it takes*" versus "*Tim doesn't, and can't, because it's logically impossible to change the past*" I reply that there is no contradiction Both conclusions are true, and for the reasons given They are compatible because "can" is equivocal

To say that something can happen means that its happening is compossible with certain facts *Which* facts? That is determined, but sometimes not determined well enough, by context An ape can't speak a human language—say, Finnish—but I can Facts about the anatomy and operation of the ape's larynx and nervous system are not compossible with his speaking Finnish The corresponding facts about my larynx and nervous system are compossible with my speaking Finnish But don't take me along to Helsinki as your interpreter I can't speak Finnish My speaking Finnish is compossible with the facts considered so far, but not with further facts about my lack of training What I can do, relative to one set of facts, I cannot do, relative to another, more inclusive, set Whenever the context leaves it open which facts are to count as relevant, it is possible to equivocate about whether I can speak Finnish It is likewise possible to equivocate about whether it is possible for me to speak Finnish, or whether I am able to, or whether I have the ability or capacity or power or potentiality to Our many words for much the same thing are little help since they do not seem to correspond to different fixed delineations of the relevant facts

Tim's killing Grandfather that day in 1921 is compossible with a fairly rich set of facts the facts about his rifle, his skill and training, the unobstructed line of fire, the locked door and the absence of any chaperone to defend the past, and so on Indeed it is compossible with all the facts of the sorts we would ordinarily count as relevant in saying what someone can do It is compossible with all the facts corresponding to those we deem relevant in Tom's case Relative to these facts, Tim can kill Grandfather But his killing Grandfather is not compossible with another, more inclusive set of facts There is the simple fact that Grandfather was not killed Also there are various other facts about Grandfather's doings after 1921 and their effects Grandfather begat Father in 1922 and Father begat Tim in 1949 Relative to these facts, Tim cannot kill Grandfather He can and he can't, but under different delineations of the relevant facts You can reasonably choose the narrower delineation, and say that he can, or the wider delineation, and say that he can't But choose What you mustn't do is waver, say in the same breath that he both can and can't, and then claim that this contradiction proves that time travel is impossible

Exactly the same goes for Tom's parallel failure For Tom to kill Grandfather's partner also is compossible with all facts of the sorts we ordinarily count as relevant, but not compossible with a larger set including, for instance, the fact that the intended victim lived until 1934 In Tom's case we are not puzzled We say without hesitation that he can do it, because we see at once that the facts that are not compossible with his success are facts about the future of the time in question and therefore not the sort of facts we count as relevant in saying what Tom can do

In Tim's case it is harder to keep track of which facts are relevant We are accustomed to exclude facts about the future of the time in question, but to include some facts about its past Our standards do not apply unequivocally to the crucial facts in this special case Tim's failure, Grandfather's survival, and his subsequent doings If we have foremost in mind that they lie in the external future of that moment in 1921 when Tim is almost ready to shoot, then we exclude them just as we exclude the parallel facts in Tom's case But if we have foremost in mind that they precede that moment in Tim's extended personal time, then we tend to include them To make the latter be foremost in your mind, I chose to tell Tim's story in the order of his personal time, rather than in the order of external time The fact of Grandfather's survival until 1957 had already been told before I got to the part of the story about Tim lurking in ambush to kill him in 1921 We must decide, if we can, whether to treat these personally past and externally future facts as if they were straightforwardly past or as if they were straightforwardly future

Fatalists—the best of them—are philosophers who take facts we count as irrelevant in saying what someone can do, disguise them somehow as facts of a different sort that we count as relevant, and thereby argue that we can do less than we think—indeed, that there is nothing at all that we don't do but can I am not going to vote Republican next fall The fatalist argues that, strange to say, I not only won't but can't, for my voting Republican is not compossible with the fact that it was true already in the year 1548 that I was not going to vote Republican 428 years later My rejoinder is that this is a fact, sure enough, however, it is an irrelevant fact about the future masquerading as a relevant fact about the past, and so should be left out of account in saying what, in any ordinary sense, I can do We are unlikely to be fooled by the fatalist's methods of disguise in this case, or other ordinary cases But in cases of time travel, precognition, or the like, we're on less familiar ground, so it may take less of a disguise to fool us Also, new methods of disguise are available, thanks to the device of personal time

Here's another bit of fatalist trickery Tim, as he lurks, already knows that he will fail At least he has the wherewithal to know it if he thinks, he knows it implicitly For he remembers that Grandfather was alive when he was a boy, he knows that those who are killed are thereafter not alive, he knows (let us suppose) that he is a time traveler who has reached the same 1921 that lies in his personal past, and he ought to understand—as we do—why a time traveler cannot change the past What is known cannot be false So his success is not only not compossible with facts that belong to the external future and his personal past, but also is not compossible with the present fact of his knowledge that he will fail I reply that the fact of his foreknowledge, at the moment while he waits to shoot, is not a fact entirely about that moment It may be divided into two parts There is the fact that he then believes (perhaps only implicitly) that he will fail, and there is the further fact that his belief is correct, and correct not at all by accident, and hence qualifies as an item of knowledge It is only the latter fact that is not compossible with his success, but it is only the former that is entirely about the moment in question In calling Tim's state at that moment knowledge, not just belief, facts about personally earlier but externally later moments were smuggled into consideration

I have argued that Tim's case and Tom's are alike, except that in Tim's case we are more tempted than usual—and with reason—to opt for a semi-fatalist mode of speech But perhaps they differ in another way In Tom's case, we can expect a perfectly consistent answer to the counterfactual question what if Tom had killed Grandfather's partner? Tim's case is more difficult If Tim had killed Grandfather, it seems offhand that contradictions would have been true The killing both would and wouldn't have occurred No Grandfather, no Father, no Father, no Tim, no Tim, no killing And for good measure no Grandfather, no family fortune, no fortune, no time machine, no time machine, no killing So the supposition that Tim killed Grandfather seems impossible in more than the semi-fatalistic sense already granted

If you suppose Tim to kill Grandfather and hold all the rest of his story fixed, of course you get a contradiction But likewise if you suppose Tom to kill Grandfather's partner and hold the rest of his story fixed—including the part that told of his failure—you get a contradiction If you make *any* counterfactual supposition and hold all else fixed you get a contradiction The thing to do is rather to make the counterfactual supposition and hold all else as close to fixed as you consistently can That procedure will yield perfectly consistent answers to

the question what if Tim had not killed Grandfather? In that case, some of the story I told would not have been true Perhaps Tim might have been the time-traveling grandson of someone else Perhaps he might have been the grandson of a man killed in 1921 and miraculously resurrected Perhaps he might have been not a time traveler at all, but rather someone created out of nothing in 1920 equipped with false memories of a personal past that never was It is hard to say what is the least revision of Tim's story to make it true that Tim kills Grandfather, but certainly the contradictory story in which the killing both does and doesn't occur is not the least revision Hence it is false (according to the unrevised story) that if Tim had killed Grandfather then contradictions would have been true

What difference would it make if Tim travels in branching time? Suppose that at the possible world of Tim's story the space-time manifold branches, the branches are separated not in time, and not in space, but in some other way Tim travels not only in time but also from one branch to another In one branch Tim is absent from the events of 1921, Grandfather lives, Tim is born, grows up, and vanishes in his time machine The other branch diverges from the first when Tim turns up in 1921, there Tim kills Grandfather and Grandfather leaves no descendants and no fortune, the events of the two branches differ more and more from that time on Certainly this is a consistent story, it is a story in which Grandfather both is and isn't killed in 1921 (in the different branches), and it is a story in which Tim, by killing Grandfather, succeeds in preventing his own birth (in one of the branches) But it is not a story in which Tim's killing of Grandfather both does occur and doesn't it simply does, though it is located in one branch and not in the other And it is not a story in which Tim changes the past 1921 and later years contain the events of both branches, coexisting somehow without interaction It remains true at all the personal times of Tim's life, even after the killing, that Grandfather lives in one branch and dies in the other [6]

[6] The present paper summarizes a series of lectures of the same title, given as the Gavin David Young Lectures in Philosophy at the University of Adelaide in July, 1971 I thank the Australian-American Educational Foundation and the American Council of Learned Societies for research support I am grateful to many friends for comments on earlier versions of this paper, especially Philip Kitcher William Newton-Smith, J J C Smart, and Donald Williams

PART FIVE

Probability

NINETEEN

A Subjectivist's Guide to Objective Chance*

INTRODUCTION

We subjectivists conceive of probability as the measure of reasonable partial belief But we need not make war against other conceptions of probability, declaring that where subjective credence leaves off, there nonsense begins Along with subjective credence we should believe also in objective chance The practice and the analysis of science require both concepts Neither can replace the other Among the propositions that deserve our credence we find, for instance, the proposition that (as a matter of contingent fact about our world) any tritium atom that now exists has a certain chance of decaying within a year Why should we subjectivists be less able than other folk to make sense of that?

Carnap (1945) did well to distinguish two concepts of probability, insisting that both were legitimate and useful and that neither was at fault because it was not the other I do not think Carnap chose quite the right two concepts, however In place of his "degree of confirmation" I would put *credence* or *degree of belief*, in place of his "relative

* I am grateful to several people for valuable discussions of this material, especially John Burgess, Nancy Cartwright, Richard Jeffrey, Peter Railton, and Brian Skyrms I am also much indebted to Mellor (1971) which presents a view very close to mine, exactly how close I am not prepared to say

frequency in the long run" I would put *chance* or *propensity*, understood as making sense in the single case The division of labor between the two concepts will be little changed by these replacements Credence is well suited to play the role of Carnap's probability$_1$, and chance to play the role of probability$_2$

Given two kinds of probability, credence and chance, we can have hybrid probabilities of probabilities (Not "second order probabilities", which suggests one kind of probability self-applied) Chance of credence need not detain us It may be partly a matter of chance what one comes to believe, but what of it? Credence about chance is more important To the believer in chance, chance is a proper subject to have beliefs about Propositions about chance will enjoy various degrees of belief, and other propositions will be believed to various degrees conditionally upon them

As I hope the following questionnaire will show, we have some very firm and definite opinions concerning reasonable credence about chance These opinions seem to me to afford the best grip we have on the concept of chance Indeed, I am led to wonder whether anyone *but* a subjectivist is in a position to understand objective chance!

QUESTIONNAIRE

First question A certain coin is scheduled to be tossed at noon today You are sure that this chosen coin is fair it has a 50% chance of falling heads and a 50% chance of falling tails You have no other relevant information Consider the proposition that the coin tossed at noon today falls heads To what degree would you now believe that proposition?

Answer 50%, of course

(Two comments (1) It is abbreviation to speak of the coin as fair Strictly speaking, what you are sure of is that the entire "chance setup" is fair coin, tosser, landing surface, air, and surroundings together are such as to make it so that the chance of heads is 50% (2) Is it reasonable to think of coin-tossing as a genuine chance process, given present-day scientific knowledge? I think so consider, for instance, that air resistance depends partly on the chance making and breaking of chemical bonds between the coin and the air molecules it encounters What is less clear is that the toss could be designed so that you could reasonably be sure that the chance of heads is 50% exactly

If you doubt that such a toss could be designed, you may substitute an example involving radioactive decay.)

Next question. As before, except that you have plenty of seemingly relevant evidence tending to lead you to expect that the coin will fall heads. This coin is known to have a displaced center of mass, it has been tossed 100 times before with 86 heads, and many duplicates of it have been tossed thousands of times with about 90% heads. Yet you remain quite sure, despite all this evidence, that the chance of heads this time is 50%. To what degree should you believe the proposition that the coin falls heads this time?

Answer. Still 50%. Such evidence is relevant to the outcome by way of its relevance to the proposition that the chance of heads is 50%, not in any other way. If the evidence somehow fails to diminish your certainty that the coin is fair, then it should have no effect on the distribution of credence about outcomes that accords with that certainty about chance. To the extent that uncertainty about outcomes is based on certainty about their chances, it is a stable, resilient sort of uncertainty—new evidence won't get rid of it. (The term "resiliency" comes from Skyrms (1977), see also Jeffrey (1965), §12.5.)

Someone might object that you could not reasonably remain sure that the coin was fair, given such evidence as I described and no contrary evidence that I failed to mention. That may be so, but it doesn't matter. Canons of reasonable belief need not be counsels of perfection. A moral code that forbids all robbery may also prescribe that if one nevertheless robs, one should rob only the rich. Likewise it is a sensible question what it is reasonable to believe about outcomes if one is unreasonably stubborn in clinging to one's certainty about chances.

Next question. As before, except that now it is afternoon and you have evidence that became available after the coin was tossed at noon. Maybe you know for certain that it fell heads, maybe some fairly reliable witness has told you that it fell heads, maybe the witness has told you that it fell heads in nine out of ten tosses of which the noon toss was one. You remain as sure as ever that the chance of heads, just before noon, was 50%. To what degree should you believe that the coin tossed at noon fell heads?

Answer. Not 50%, but something not far short of 100%. Resiliency has its limits. If evidence bears in a direct enough way on the outcome—a way which may nevertheless fall short of outright implication—then it may bear on your beliefs about outcomes otherwise than by way of your beliefs about the chances of the outcomes. Resiliency under all evidence whatever would be extremely unreasonable.

We can only say that degrees of belief about outcomes that are based on certainty about chances are resilient under *admissible* evidence The previous question gave examples of admissible evidence, this question gave examples of inadmissible evidence

Last question You have no inadmissible evidence, if you have any relevant admissible evidence, it already has had its proper effect on your credence about the chance of heads But this time, suppose you are not sure that the coin is fair You divide your belief among three alternative hypotheses about the chance of heads, as follows

You believe to degree 27% that the chance of heads is 50%

You believe to degree 22% that the chance of heads is 35%

You believe to degree 51% that the chance of heads is 80%

Then to what degree should you believe that the coin falls heads?

Answer (27% × 50%) + (22% × 35%) + (51% × 80%), that is, 62% Your degree of belief that the coin falls heads, conditionally on any one of the hypotheses about the chance of heads, should equal your unconditional degree of belief if you were sure of that hypothesis That in turn should equal the chance of heads according to the hypothesis 50% for the first hypothesis, 35% for the second, and 80% for the third Given your degrees of belief that the coin falls heads, conditionally on the hypotheses, we need only apply the standard multiplicative and additive principles to obtain our answer

THE PRINCIPAL PRINCIPLE

I have given undefended answers to my four questions I hope you found them obviously right, so that you will be willing to take them as evidence for what follows If not, do please reconsider If so, splendid—now read on

It is time to formulate a general principle to capture the intuitions that were forthcoming in our questionnaire It will resemble familiar principles of direct inference except that (1) it will concern chance, not some sort of actual or hypothetical frequency, and (2) it will incorporate the observation that certainty about chances—or conditionality on propositions about chances—makes for resilient degrees of belief about outcomes Since this principle seems to me to capture all we know about chance, I call it

THE PRINCIPAL PRINCIPLE Let C be any reasonable initial credence function Let t be any time Let x be any real number in the unit interval Let X be the proposition that the chance, at time t, of A's holding equals x Let E be any proposition compatible with X that is admissible at time t Then

$$C(A/XE) = x$$

That will need a good deal of explaining But first I shall illustrate the principle by applying it to the cases in our questionnaire

Suppose your present credence function is $C(-/E)$, the function that comes from some reasonable initial credence function C by conditionalizing on your present total evidence E Let t be the time of the toss, noon today, and let A be the proposition that the coin tossed today falls heads Let X be the proposition that the chance at noon (just before the toss) of heads is x (In our questionnaire, we mostly considered the case that x is 50%) Suppose that nothing in your total evidence E contradicts X, suppose also that it is not yet noon, and you have no foreknowledge of the outcome, so everything that is included in E is entirely admissible The conditions of the Principal Principle are met Therefore $C(A/XE)$ equals x That is to say that x is your present degree of belief that the coin falls heads, conditionally on the proposition that its chance of falling heads is x If in addition you are sure that the chance of heads is x—that is, if $C(X/E)$ is one—then it follows also that x is your present unconditional degree of belief that the coin falls heads More generally, whether or not you are sure about the chance of heads, your unconditional degree of belief that the coin falls heads is given by summing over alternative hypotheses about chance

$$C(A/E) = \Sigma_x C(X_x/E)C(A/X_xE) = \Sigma_x C(X_x/E)x,$$

where X_x, for any value of x, is the proposition that the chance at t of A equals x

Several parts of the formulation of the Principal Principle call for explanation and comment Let us take them in turn

THE INITIAL CREDENCE FUNCTION C

I said let C be any reasonable initial credence function By that I meant, in part, that C was to be a probability distribution over (at least) the space whose points are possible worlds and whose regions

(sets of worlds) are propositions C is a non-negative, normalized, finitely additive measure defined on all propositions

The corresponding conditional credence function is defined simply as a quotient of unconditional credences

$$C(A/B) =_{df} C(AB)/C(B)$$

I should like to assume that it makes sense to conditionalize on any but the empty proposition Therefore, I require that C is *regular* $C(B)$ is zero, and $C(A/B)$ is undefined, only if B is the empty proposition, true at no worlds You may protest that there are too many alternative possible worlds to permit regularity But that is so only if we suppose, as I do not, that the values of the function C are restricted to the standard reals Many propositions must have infinitesimal C-values, and $C(A/B)$ often will be defined as a quotient of infinitesimals, each infinitely close but not equal to zero (See Bernstein and Wattenberg (1969)) The assumption that C is regular will prove convenient, but it is not justified only as a convenience Also it is required as a condition of reasonableness one who started out with an irregular credence function (and who then learned from experience by conditionalizing) would stubbornly refuse to believe some propositions no matter what the evidence in their favor

In general, C is to be reasonable in the sense that if you started out with it as your initial credence function, and if you always learned from experience by conditionalizing on your total evidence, then no matter what course of experience you might undergo your beliefs would be reasonable for one who had undergone that course of experience I do not say what distinguishes a reasonable from an unreasonable credence function to arrive at after a given course of experience We do make the distinction, even if we cannot analyze it, and therefore I may appeal to it in saying what it means to require that C be a reasonable initial credence function

I have assumed that the method of conditionalizing is *one* reasonable way to learn from experience, given the right initial credence function I have not assumed something more controversial that it is the *only* reasonable way The latter view may also be right (the cases where it seems wrong to conditionalize may all be cases where one departure from ideal rationality is needed to compensate for another) but I shall not need it here

(I said that C was to be a probability distribution over *at least* the space of worlds, the reason for that qualification is that sometimes one's credence might be divided between different possibilities within

a single world That is the case for someone who is sure what sort of world he lives in, but not at all sure who and when and where in the world he is In a fully general treatment of credence it would be well to replace the worlds by something like the "centered worlds" of Quine (1969), and the propositions by something corresponding to properties But I shall ignore these complications here)

THE REAL NUMBER x

I said let x be any real number in the unit interval I must emphasize that "x" is a quantified variable, it is not a schematic letter that may freely be replaced by terms that designate real numbers in the unit interval For fixed A and t, "the chance, at t, of A's holding" is such a term, suppose we put it in for the variable x It might seem that for suitable C and E we have the following if X is the proposition that the chance, at t, of A's holding equals the chance, at t, of A's holding—in other words, if X is the necessary proposition—then

$$C(A/XE) = \text{the chance, at } t, \text{ of } A\text{'s holding}$$

But that is absurd It means that if E is your present total evidence and $C(-/E)$ is your present credence function, then if the coin is in fact fair—whether or not you think it is!—then your degree of belief that it falls heads is 50% Fortunately, that absurdity is not an instance of the Principal Principle The term "the chance, at t, of A's holding" is a non-rigid designator, chance being a matter of contingent fact, it designates different numbers at different worlds The context "the proposition that ", within which the variable "x" occurs, is intensional Universal instantiation into an intensional context with a non-rigid term is a fallacy It is the fallacy that takes you, for instance, from the true premise "For any number x, the proposition that x is nine is non-contingent" to the false conclusion "The proposition that the number of planets is nine is non-contingent" See Jeffrey (1970) for discussion of this point in connection with a relative of the Principal Principle

I should note that the values of "x" are not restricted to the standard reals in the unit interval The Principal Principle may be applied as follows you are sure that some spinner is fair, hence that it has infinitesimal chance of coming to rest at any particular point, therefore (if your total evidence is admissible) you should believe only to an infinitesimal degree that it will come to rest at any particular point

THE PROPOSITION *X*

I said let *X* be the proposition that the chance, at time *t*, of *A*'s holding equals *x*. I emphasize that I am speaking of objective, single-case chance—not credence, not frequency. Like it or not, we have this concept. We think that a coin about to be tossed has a certain chance of falling heads, or that a radioactive atom has a certain chance of decaying within the year, quite regardless of what anyone may believe about it and quite regardless of whether there are any other similar coins or atoms. As philosophers we may well find the concept of objective chance troublesome, but that is no excuse to deny its existence, its legitimacy, or its indispensability. If we can't understand it, so much the worse for us.

Chance and credence are distinct, but I don't say they are unrelated. What is the Principal Principle but a statement of their relation? Neither do I say that chance and frequency are unrelated, but they are distinct. Suppose we have many coin-tosses with the same chance of heads (not zero or one) in each case. Then there is some chance of getting any frequency of heads whatever, and hence some chance that the frequency and the uniform single-case chance of heads may differ, which could not be so if these were one and the same thing. Indeed the chance of difference may be infinitesimal if there are infinitely many tosses, but that is still not zero. Nor do hypothetical frequencies fare any better. There is no such thing as *the* infinite sequence of outcomes, or *the* limiting frequency of heads, that *would* eventuate if some particular coin-toss were somehow repeated forever. Rather there are countless sequences, and countless frequencies, that *might* eventuate and would have some chance (perhaps infinitesimal) of eventuating. (See Jeffrey (1977), Skyrms (1977), and the discussion of "might" counterfactuals in Lewis (1973).)

Chance is not the same thing as credence or frequency, this is not yet to deny that there might be some roundabout way to analyze chance in terms of credence or frequency. I would only ask that no such analysis be accepted unless it is compatible with the Principal Principle. We shall consider how this requirement bears on the prospects for an analysis of chance, but without settling the question of whether such an analysis is possible.

I think of chance as attaching in the first instance to propositions. The chance of an event, an outcome, etc. is the chance of truth of the proposition that holds at just those worlds where that event, outcome, or whatnot occurs. (Here I ignore the special usage of "event" to

simply mean "proposition") I have foremost in mind the chances of truth of propositions about localized matters of particular fact—a certain toss of a coin, the fate of a certain tritium atom on a certain day—but I do not say that those are the only propositions to which chance applies Not only does it make sense to speak of the chance that a coin will fall heads on a particular occasion, equally it makes sense to speak of the chance of getting exactly seven heads in a particular sequence of eleven tosses It is only caution, not any definite reason to think otherwise, that stops me from assuming that chance of truth applies to any proposition whatever I shall assume, however, that the broad class of propositions to which chance of truth applies is closed under the Boolean operations of conjunction (intersection), disjunction (union), and negation (complementation)

We ordinarily think of chance as time-dependent, and I have made that dependence explicit Suppose you enter a labyrinth at 11 00 a m , planning to choose your turn whenever you come to a branch point by tossing a coin When you enter at 11 00, you may have a 42% chance of reaching the center by noon But in the first half hour you may stray into a region from which it is hard to reach the center, so that by 11 30 your chance of reaching the center by noon has fallen to 26% But then you turn lucky, by 11 45 you are not far from the center and your chance of reaching it by noon is 78% At 11 49 you reach the center, then and forevermore your chance of reaching it by noon is 100%

Sometimes, to be sure, we omit reference to a time I do not think this means that we have some timeless notion of chance Rather, we have other ways to fix the time than by specifying it explicitly In the case of the labyrinth we might well say (before, after, or during your exploration) that your chance of reaching the center by noon is 42% The understood time of reference is the time when your exploration begins Likewise we might speak simply of the chance of a certain atom's decaying within a certain year, meaning the chance at the beginning of that year In general, if A is the proposition that something or other takes place within a certain interval beginning at time t, then we may take a special interest in what I shall call the *endpoint chance* of A's holding the chance at t, the beginning of the interval in question If we speak simply of the chance of A's holding, not mentioning a time, it is this endpoint chance—the chance at t of A's holding—that we are likely to mean

Chance also is world-dependent Your chance at 11 00 of reaching the center of the labyrinth by noon depends on all sorts of contingent features of the world the structure of the labyrinth and the speed with

which you can walk through it, for instance Your chance at 11 30 of reaching the center by noon depends on these things, and also on where in the labyrinth you then are Since these things vary from world to world, so does your chance (at either time) of reaching the center by noon Your chance at noon of reaching the center by noon is one at the worlds where you have reached the center, zero at all others, including those worlds where you do not explore the labyrinth at all, perhaps because you or it do not exist (Here I am speaking loosely, as if I believed that you and the labyrinth could inhabit several worlds at once See Lewis (1968) for the needed correction)

We have decided this much about chance, at least it is a function of three arguments To a proposition, a time, and a world it assigns a real number Fixing the proposition A, the time t, and the number x, we have our proposition X it is the proposition that holds at all and only those worlds w such that this function assigns to A, t, and w the value x This is the proposition that the chance, at t, of A's holding is x

THE ADMISSIBLE PROPOSITION *E*

I said let E be any proposition that is admissible at time t Admissible propositions are the sort of information whose impact on credence about outcomes comes entirely by way of credence about the chances of those outcomes Once the chances are given outright, conditionally or unconditionally, evidence bearing on them no longer matters (Once it is settled that the suspect fired the gun, the discovery of his fingerprint on the trigger adds nothing to the case against him) The power of the Principal Principle depends entirely on how much is admissible If nothing is admissible it is vacuous If everything is admissible it is inconsistent Our questionnaire suggested that a great deal is admissible, but we saw examples also of inadmissible information I have no definition of admissibility to offer, but must be content to suggest sufficient (or almost sufficient) conditions for admissibility I suggest that two different sorts of information are generally admissible

The first sort is historical information If a proposition is entirely about matters of particular fact at times no later than t, then as a rule that proposition is admissible at t Admissible information just before the toss of a coin, for example, includes the outcomes of all

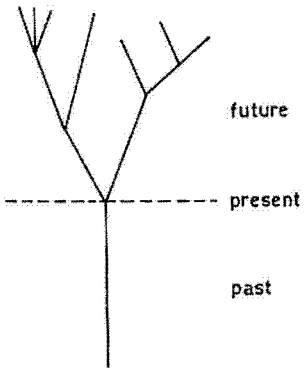
previous tosses of that coin and others like it It also includes every detail—no matter how hard it might be to discover—of the structure of the coin, the tosser, other parts of the set-up, and even anything nearby that might somehow intervene It also includes a great deal of other information that is completely irrelevant to the outcome of the toss

A proposition is *about* a subject matter—about history up to a certain time, for instance—if and only if that proposition holds at both or neither of any two worlds that match perfectly with respect to that subject matter (Or we can go the other way two worlds match perfectly with respect to a subject matter if and only if every proposition about that subject matter holds at both or neither) If our world and another are alike point for point, atom for atom, field for field, even spirit for spirit (if such there be) throughout the past and up until noon today, then any proposition that distinguishes the two cannot be entirely about the respects in which there is no difference It cannot be entirely about what goes on no later than noon today That is so even if its linguistic expression makes no overt mention of later times, we must beware lest information about the future is hidden in the predicates, as in "Fred was mortally wounded at 11 58" I doubt that any linguistic test of aboutness will work without circular restrictions on the language used Hence it seems best to take either "about" or "perfect match with respect to" as a primitive

Time-dependent chance and time-dependent admissibility go together Suppose the proposition A is about matters of particular fact at some moment or interval t_A, and suppose we are concerned with chance at time t If t is later than t_A, then A is admissible at t The Principal Principle applies with A for E If X is the proposition that the chance at t of A equals x, and if A and X are compatible, then

$$1 = C(A/XA) = x$$

Put contrapositively, this means that if the chance at t of A, according to X, is anything but one, then A and X are incompatible A implies that the chance at t of A, unless undefined, equals one What's past is no longer chancy The past, unlike the future, has no chance of being any other way than the way it actually is This temporal asymmetry of chance falls into place as part of our conception of the past as "fixed" and the future as "open"—whatever that may mean The asymmetry of fixity and of chance may be pictured by a tree The single trunk is



the one possible past that has any present chance of being actual The many branches are the many possible futures that have some present chance of being actual I shall not try to say here what features of the world justify our discriminatory attitude toward past and future possibilities, reflected for instance in the judgment that historical information is admissible and similar information about the future is not But I think they are contingent features, subject to exception and absent altogether from some possible worlds

That possibility calls into question my thesis that historical information is invariably admissible What if the commonplace *de facto* asymmetries between past and future break down? If the past lies far in the future, as we are far to the west of ourselves, then it cannot simply be that propositions about the past are admissible and propositions about the future are not And if the past contains seers with foreknowledge of what chance will bring, or time travelers who have witnessed the outcome of coin-tosses to come, then patches of the past are enough tainted with futurity so that historical information about them may well seem inadmissible That is why I qualified my claim that historical information is admissible, saying only that it is so "as a rule" Perhaps it is fair to ignore this problem in building a case that the Principal Principle captures our common opinions about chance, since those opinions may rest on a naive faith that past and future cannot possibly get mixed up Any serious physicist, if he remains at least open-minded both about the shape of the cosmos and about the existence of chance processes, ought to do better But I shall not, I shall carry on as if historical information is admissible without exception

Besides historical information, there is at least one other sort of admissible information hypothetical information about chance itself

Let us return briefly to our questionnaire and add one further supposition to each case Suppose you have various opinions about what the chance of heads would be under various hypotheses about the detailed nature and history of the chance set-up under consideration Suppose further that you have similar hypothetical opinions about other chance set-ups, past, present, and future (Assume that these opinions are consistent with your admissible historical information and your opinions about chance in the present case) It seems quite clear to me—and I hope it does to you also—that these added opinions do not change anything The correct answers to the questionnaire are just as before The added opinions do not bear in any overly direct way on the future outcomes of chance processes Therefore they are admissible

We must take care, though Some propositions about future chances do reveal inadmissible information about future history, and these are inadmissible Recall the case of the labyrinth you enter at 11 00, choosing your turns by chance, and hope to reach the center by noon Your subsequent chance of success depends on the point you have reached The proposition that at 11 30 your chance of success has fallen to 26% is not admissible information at 11 00, it is a giveaway about your bad luck in the first half hour What is admissible at 11 00 is a conditional version if you were to reach a certain point at 11 30, your chance of success would then be 26% But even some conditionals are tainted for instance, any conditional that could yield inadmissible information about future chances by *modus ponens* from admissible historical propositions Consider also the truth-functional conditional that if history up to 11 30 follows a certain course, then you will have a 98% chance of becoming a monkey's uncle before the year is out This conditional closely resembles the denial of its antecedent, and is inadmissible at 11 00 for the same reason

I suggest that conditionals of the following sort, however, are admissible, and indeed admissible at all times (1) The consequent is a proposition about chance at a certain time (2) The antecedent is a proposition about history up to that time, and further, it is a complete proposition about history up to that time, so that it either implies or else is incompatible with any other proposition about history up to that time It fully specifies a segment, up to the given time, of some possible course of history (3) The conditional is made from its consequent and antecedent not truth-functionally, but rather by means of a strong conditional operation of some sort This might well be the counterfactual conditional of Lewis (1973), but various rival versions would serve as well, since many differences do not matter for the case

at hand One feature of my treatment will be needed, however if the antecedent of one of our conditionals holds at a world, then both or neither of the conditional and its consequent hold there

These admissible conditionals are propositions about how chance depends (or fails to depend) on history They say nothing, however, about how history chances to go A set of them is a theory about the way chance works It may or may not be a complete theory, a consistent theory, a systematic theory, or a credible theory It might be a miscellany of unrelated propositions about what the chances would be after various fully specified particular courses of events Or it might be systematic, compressible into generalizations to the effect that after any course of history with property J there would follow a chance distribution with property K (For instance, it might say that any coin with a certain structure would be fair) These generalizations are universally quantified conditionals about single-case chance, if lawful, they are probabilistic laws in the sense of Railton (1978) (I shall not consider here what would make them lawful, but see Lewis (1973), §3 3, for a treatment that could cover laws about chance along with other laws) Systematic theories of chance are the ones we can express in language, think about, and believe to substantial degrees But a reasonable initial credence function does not reject any possibility out of hand It assigns some non-zero credence to any consistent theory of chance, no matter how unsystematic and incompressible it is

Historical propositions are admissible, so are propositions about the dependence of chance on history Combinations of the two, of course, are also admissible More generally, we may assume that any Boolean combination of propositions admissible at a time also is admissible at that time Admissibility consists in keeping out of a forbidden subject matter—how the chance processes turned out—and there is no way to break into a subject matter by making Boolean combinations of propositions that lie outside it

There may be sorts of admissible propositions besides those I have considered If so, we shall have no need of them in what follows

This completes an exposition of the Principal Principle We turn next to an examination of its consequences I maintain that they include all that we take ourselves to know about chance

THE PRINCIPLE REFORMULATED

Given a time t and world w, let us write P_{tw} for the *chance distribution* that obtains at t and w For any proposition A, $P_{tw}(A)$ is the chance, at

time t and world w, of A's holding (The domain of P_{tw} comprises those propositions for which this chance is defined)

Let us also write H_{tw} for the *complete history* of world w up to time t the conjunction of all propositions that hold at w about matters of particular fact no later than t H_{tw} is the proposition that holds at exactly those worlds that perfectly match w, in matters of particular fact, up to time t

Let us also write T_w for the *complete theory of chance* for world w the conjunction of all the conditionals from history to chance, of the sort just considered, that hold at w Thus T_w is a full specification, for world w, of the way chances at any time depend on history up to that time

Taking the conjunction $H_{tw}T_w$, we have a proposition that tells us a great deal about the world w It is nevertheless admissible at time t, being simply a giant conjunction of historical propositions that are admissible at t and conditionals from history to chance that are admissible at any time Hence the Principal Principle applies

$$C(A/XH_{tw}T_w) = x$$

when C is a reasonable initial credence function, X is the proposition that the chance at t of A is x, and $H_{tw}T_w$ is compatible with X

Suppose X holds at w That is so if and only if x equals $P_{tw}(A)$ Hence we can choose such an X whenever A is in the domain of P_{tw} $H_{tw}T_w$ and X both hold at w, therefore they are compatible But further, $H_{tw}T_w$ implies X The theory T_w and the history H_{tw} together are enough to imply all that is true (and contradict all that is false) at world w about chances at time t For consider the strong conditional with antecedent H_{tw} and consequent X This conditional holds at w, since by hypothesis its antecedent and consequent hold there Hence it is implied by T_w, which is the conjunction of all conditionals of its sort that hold at w, and this conditional and H_{tw} yield X by *modus ponens* Consequently, the conjunction $XH_{tw}T_w$ simplifies to $H_{tw}T_w$ Provided that A is in the domain of P_{tw} so that we can make a suitable choice of X, we can substitute $P_{tw}(A)$ for x, and $H_{tw}T_w$ for $XH_{tw}T_w$, in our instance of the Principal Principle Therefore we have

THE PRINCIPAL PRINCIPLE REFORMULATED Let C be any reasonable initial credence function Then for any time t, world w, and proposition A in the domain of P_{tw}

$$P_{tw}(A) = C(A/H_{tw}T_w)$$

In words the chance distribution at a time and a world comes from any reasonable initial credence function by conditionalizing on the complete history of the world up to the time, together with the complete theory of chance for the world

This reformulation enjoys less direct intuitive support than the original formulation, but it will prove easier to use It will serve as our point of departure in examining further consequences of the Principal Principle

CHANCE AND THE PROBABILITY CALCULUS

A reasonable initial credence function is, among other things, a probability distribution a non-negative, normalized, finitely additive measure It obeys the laws of mathematical probability theory There are well-known reasons why that must be so if credence is to rationalize courses of action that would not seem blatantly unreasonable in some circumstances

Whatever comes by conditionalizing from a probability distribution is itself a probability distribution Therefore a chance distribution is a probability distribution For any time t and world w, P_{tw} obeys the laws of mathematical probability theory These laws carry over from credence to chance via the Principal Principle We have no need of any independent assumption that chance is a kind of probability

Observe that although the Principal Principle concerns the relationship between chance and credence, some of its consequences concern chance alone We have seen two such consequences (1) The thesis that the past has no present chance of being otherwise than it actually is (2) The thesis that chance obeys the laws of probability More such consequences will appear later

CHANCE AS OBJECTIFIED CREDENCE

Chance is an objectified subjective probability in the sense of Jeffrey (1965), §12 7 Jeffrey's construction (omitting his use of sequences of partitions, which is unnecessary if we allow infinitesimal credences) works as follows Suppose given a partition of logical space a set of mutually exclusive and jointly exhaustive propositions Then we can define the *objectification* of a credence function, with respect to this

partition, at a certain world, as the probability distribution that comes from the given credence function by conditionalizing on the member of the given partition that holds at the given world. Objectified credence is credence conditional on the truth—not the whole truth, however, but exactly as much of it as can be captured by a member of the partition without further subdivision of logical space. The member of the partition that holds depends on matters of contingent fact, varying from one world to another; it does not depend on what we think (except insofar as our thoughts are relevant matters of fact) and we may well be ignorant or mistaken about it. The same goes for objectified credence.

Now consider one particular way of partitioning. For any time t, consider the partition consisting of the propositions $H_{tw}T_w$ for all worlds w. Call this the *history-theory partition* for time t. A member of this partition is an equivalence class of worlds with respect to the relation of being exactly alike both in respect of matters of particular fact up to time t and in respect of the dependence of chance on history. The Principal Principle tells us that the chance distribution, at any time t and world w, is the objectification of any reasonable credence function, with respect to the history-theory partition for time t, at world w. Chance is credence conditional on the truth—*if* the truth is subject to censorship along the lines of the history-theory partition, and *if* the credence is reasonable.

Any historical proposition admissible at time t, or any admissible conditional from history to chance, or any admissible Boolean combination of propositions of these two kinds—in short, any sort of admissible proposition we have considered—is a disjunction of members of the history-theory partition for t. Its borders follow the lines of the partition, never cutting between two worlds that the partition does not distinguish. Likewise for any proposition about chances at t. Let X be the proposition that the chance at t of A is x, let Y be any member of the history-theory partition for t, and let C be any reasonable initial credence function. Then, according to our reformulation of the Principal Principle, X holds at all worlds in Y if $C(A/Y)$ equals x, and at no worlds in Y otherwise. Therefore X is the disjunction of all members Y of the partition such that $C(A/Y)$ equals x.

We may picture the situation as follows. The partition divides logical space into countless tiny squares. In each square there is a black region where A holds and a white region where it does not. Now blur the focus, so that divisions within the squares disappear from view. Each square becomes a grey patch in a broad expanse covered with varying

shades of grey Any maximal region of uniform shade is a proposition specifying the chance of A The darker the shade, the higher is the uniform chance of A at the worlds in the region The worlds themselves are not grey—they are black or white, worlds where A holds or where it doesn't—but we cannot focus on single worlds, so they all seem to be the shade of grey that covers their region Admissible propositions, of the sorts we have considered, are regions that may cut across the contours of the shades of grey The conjunction of one of these admissible propositions and a proposition about the chance of A is a region of uniform shade, but not in general a maximal uniform region It consists of some, but perhaps not all, the members Y of the partition for which $C(A/Y)$ takes a certain value

We derived our reformulation of the Principal Principle from the original formulation, but have not given a reverse derivation to show the two formulations equivalent In fact the reformulation may be weaker, but not in any way that is likely to matter Let C be a reasonable initial credence function, let X be the proposition that the chance at t of A is x, let E be admissible at t (in one of the ways we have considered) and compatible with X According to the reformulation, as we have seen, XE is a disjunction of incompatible propositions Y, for each of which $C(A/Y)$ equals x If there were only finitely many Y's, it would follow that $C(A/XE)$ also equals x But the implication fails in certain cases with infinitely many Y's (and indeed we would expect the history-theory partition to be infinite) so we cannot quite recover the original formulation in this way The cases of failure are peculiar, however, so the extra strength of the original formulation in ruling them out seems unimportant

KINEMATICS OF CHANCE

Chance being a kind of probability, we may define conditional chance in the usual way as a quotient (leaving it undefined if the denominator is zero)

$$P_{tw}(A/B) =_{\mathrm{df}} P_{tw}(AB)/P_{tw}(B)$$

To simplify notation, let us fix on a particular world—ours, as it might be—and omit the subscript "w", let us fix on some particular reasonable initial credence function C, it doesn't matter which, and let us fix on a sequence of times, in order from earlier to later, to be called 1, 2, 3, (I do not assume they are equally spaced) For any time t in our

sequence, let the proposition I_t be the complete history of our chosen world in the interval from time t to time $t + 1$ (including $t + 1$ but not t) Thus I_t is the set of worlds that match the chosen world perfectly in matters of particular fact throughout the given interval

A complete history up to some time may be extended by conjoining complete histories of subsequent intervals H_2 is H_1I_1, H_3 is $H_1I_1I_2$, and so on Then by the Principal Principal we have

$$P_1(A) = C(A/H_1T),$$

$$P_2(A) = C(A/H_2T) = C(A/H_1I_1T) = P_1(A/I_1),$$

$$P_3(A) = C(A/H_3T) = C(A/H_1I_1I_2T) = P_2(A/I_2) = P_1(A/I_1I_2),$$

and in general

$$P_{t+n+1}(A) = P_t(A/I_t \quad I_{t+n})$$

In words a later chance distribution comes from an earlier one by conditionalizing on the complete history of the interval in between

The evolution of chance is parallel to the evolution of credence for an agent who learns from experience, as he reasonably might, by conditionalizing In that case a later credence function comes from an earlier one by conditionalizing on the total increment of evidence gained in the interval in between For the evolution of chance we simply put the world's chance distribution in place of the agent's credence function, and the totality of particular fact about a time in place of the totality of evidence gained at that time

In the interval from t to $t + 1$ there is a certain way that the world will in fact develop namely, the way given by I_t And at t, the last moment before the interval begins, there is a certain chance that the world will develop in that way $P_t(I_t)$, the endpoint chance of I_t Likewise for a longer interval, say from time 1 to time 18 The world will in fact develop in the way given by $I_1 \quad I_{17}$, and the endpoint chance of its doing so is $P_1(I_1 \quad I_{17})$ By definition of conditional chance

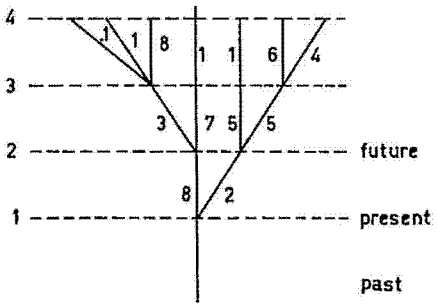
$$P_1(I_1 \quad I_{17}) = P_1(I_1)\,P_1(I_2/I_1)\,P_1(I_3/I_1I_2) \quad P_1(I_{17}/I_1 \quad I_{16}),$$

and by the Principal Principle, applied as above,

$$P_1(I_1 \quad I_{17}) = P_1(I_1)\,P_2(I_2)\,P_3(I_3) \quad P_{17}(I_{17})$$

In general, if an interval is divided into subintervals, then the endpoint chance of the complete history of the interval is the product of the endpoint chances of the complete histories of the subintervals

Earlier we drew a tree to represent the temporal asymmetry of chance Now we can embellish the tree with numbers to represent the kinematics of chance Take time 1 as the present Worlds—those of them that are compatible with a certain common past and a certain common theory of chance—lie along paths through the tree The numbers on each segment give the endpoint chance of the course of history represented by that segment, for any world that passes through that segment Likewise, for any path consisting of several segments, the product of numbers along the path gives the endpoint chance of the course of history represented by the entire path



CHANCE OF FREQUENCY

Suppose that there is to be a long sequence of coin tosses under more or less standardized conditions The first will be in the interval between time 1 and time 2, the second in the interval between 2 and 3, and so on Our chosen world is such that at time 1 there is no chance, or negligible chance, that the planned sequence of tosses will not take place And indeed it does take place The outcomes are given by a sequence of propositions A_1, A_2, Each A_t states truly whether the toss between t and $t + 1$ fell heads or tails A conjunction A_1 A_n then gives the history of outcomes for an initial segment of the sequence

The endpoint chance $P_1(A_1$ $A_n)$ of such a sequence of outcomes is given by a product of conditional chances By definition of conditional chance,

$$P_1(A_1 \quad A_n) = P_1(A_1) \quad P_1(A_2/A_1) \quad P_1(A_3/A_1A_2)$$
$$P_1(A_n/A_1 \quad A_{n-1})$$

Since we are dealing with propositions that give only incomplete histories of intervals, there is no general guarantee that these factors equal the endpoint chances of the A's The endpoint chance of A_2, $P_2(A_2)$, is given by $P_1(A_2/I_1)$, this may differ from $P_1(A_2/A_1)$ because the complete history I_1 includes some relevant information that the incomplete history A_1 omits about chance occurrences in the first interval Likewise for the conditional and endpoint chances pertaining to later intervals

Even though there is no general guarantee that the endpoint chance of a sequence of outcomes equals the product of the endpoint chances of the individual outcomes, yet it may be so if the world is right It may be, for instance, that the endpoint chance of A_2 does not depend on those aspects of the history of the first interval that are omitted from A_1—it would be the same regardless Consider the class of all possible complete histories up to time 2 that are compatible both with the previous history H_1 and with the outcome A_1 of the first toss These give all the ways the omitted aspects of the first interval might be For each of these histories, some strong conditional holds at our chosen world that tells what the chance at 2 of A_2 would be if that history were to come about Suppose all these conditionals have the same consequent whichever one of the alternative histories were to come about, it would be that X, where X is the proposition that the chance at 2 of A_2 equals x Then the conditionals taken together tell us that the endpoint chance of A_2 is independent of all aspects of the history of the first interval except the outcome of the first toss

In that case we can equate the conditional chance $P_1(A_2/A_1)$ and the endpoint chance $P_2(A_2)$ Note that our conditionals are of the sort implied by T, the complete theory of chance for our chosen world Hence A_1, H_1, and T jointly imply X It follows that A_1H_1T and XA_1H_1T are the same proposition It also follows that X holds at our chosen world, and hence that x equals $P_2(A_2)$ Note also that A_1H_1T is admissible at time 2 Now, using the Principal Principle first as reformulated and then in the original formulation, we have

$$P_1(A_2/A_1) = C(A_2/A_1H_1T) = C(A_2/XA_1H_1T) = x = P_2(A_2)$$

If we also have another such battery of conditionals to the effect that the endpoint chance of A_3 is independent of all aspects of the history of the first two intervals except the outcomes A_1 and A_2 of the first two

tosses, and another battery for A_4, and so on, then the multiplicative rule for endpoint chances follows

$$P_1(A_1 \quad A_n) = P_1(A_1)\ P_2(A_2)\ P_3(A_3) \quad P_n(A_n)$$

The conditionals that constitute the independence of endpoint chances mean that the incompleteness of the histories $A_1, A_2,$ doesn't matter The missing part wouldn't make any difference

We might have a stronger form of independence The endpoint chances might not depend on *any* aspects of history after time 1, not even the outcomes of previous tosses Then conditionals would hold at our chosen world to the effect that if any complete history up to time 2 which is compatible with H_1 were to come about, it would be that X (where X is again the proposition that the chance at 2 of A_2 equals x) We argue as before, leaving out A_1 T implies the conditionals, H_1 and T jointly imply X, H_1T and XH_1T are the same, X holds, x equals $P_2(A_2)$, H_1T is admissible at 2, so, using the Principal Principle in both formulations, we have

$$P_1(A_2) = C(A_2/H_1T) = C(A_2/XH_1T) = x = P_2(A_2)$$

Our strengthened independence assumption implies the weaker independence assumption of the previous case, wherefore

$$P_1(A_2/A_1) = P_2(A_2) = P_1(A_2)$$

If the later outcomes are likewise independent of history after time 1, then we have a multiplicative rule not only for endpoint chances but also for unconditional chances of outcomes at time 1

$$P_1(A_1 \quad A_n) = P_1(A_1)\,P_1(A_2)\,P_1(A_3) \quad P_1(A_n)$$

Two conceptions of independence are in play together One is the familiar probabilistic conception A_2 is independent of A_1, with respect to the chance distribution P_1, if the conditional chance $P_1(A_2/A_1)$ equals the unconditional chance $P_1(A_2)$, equivalently, if the chance $P_1(A_1A_2)$ of the conjunction equals the product $P_1(A_1)\,P_1(A_2)$ of the chances of the conjuncts The other conception involves batteries of strong conditionals with different antecedents and the same consequent (I consider this to be *causal* independence, but that's another story) The conditionals need not have anything to do with probability, for instance, my beard does not depend on my politics since I would have such a beard whether I were Republican, Democrat, Prohibitionist, Libertarian, Socialist Labor, or whatever But one sort of consequent that can be independent of a range of alternatives, as we

have seen, is a consequent about single-case chance What I have done is to use the Principal Principle to parlay battery-of-conditionals independence into ordinary probabilistic independence

If the world is right, the situation might be still simpler, and this is the case we hope to achieve in a well-conducted sequence of chance trials Suppose the history-to-chance conditionals and the previous history of our chosen world give us not only independence (of the stronger sort) but also uniformity of chances for any toss in our sequence, the endpoint chance of heads on that toss would be h (and the endpoint chance of tails would be $1 - h$) no matter which of the possible previous histories compatible with H_1 might have come to pass Then each of the A_t's has an endpoint chance of h if it specifies an outcome of heads, $1 - h$ if it specifies an outcome of tails By the multiplicative rule for endpoint chances,

$$P_1(A_1 \quad A_n) = h^{fn}(1-h)^{(n-fn)}$$

where f is the frequency of heads in the first n tosses according to $A_1 \quad A_n$

Now consider any other world that matches our chosen world in its history up to time 1 and in its complete theory of chance, but not in its sequence of outcomes By the Principal Principle, the chance distribution at time 1 is the same for both worlds Our assumptions of independence and uniformity apply to both worlds, being built into the shared history and theory So all goes through for this other world as it did for our chosen world Our calculation of the chance at time 1 of a sequence of outcomes, as a function of the uniform single-case chance of heads and the length and frequency of heads in the sequence, goes for any sequence, not only for the sequence $A_1, A_2,$ that comes about at our chosen world

Let F be the proposition that the frequency of heads in the first n tosses is f F is a disjunction of propositions each specifying a sequence of n outcomes with frequency f of heads, each disjunct has the same chance at time 1, under our assumptions of independence and uniformity, and the disjuncts are incompatible Multiplying the number of these propositions by the uniform chance of each, we get the chance of obtaining some or other sequence of outcomes with frequency f of heads

$$P_1(F) = \frac{n! \quad h^{fn} \quad (1-h)^{(n-fn)}}{(fn)! \quad (n-fn)!}$$

The rest is well known For fixed h and n, the right hand side of the

equation peaks for f close to h, the greater is n, the sharper is the peak If there are many tosses, then the chance is close to one that the frequency of heads is close to the uniform single-case chance of heads The more tosses, the more stringent we can be about what counts as "close" That much of frequentism is true, and that much is a consequence of the Principal Principle, which relates chance not only to credence but also to frequency

On the other hand, unless h is zero or one, the right hand side of the equation is non-zero So, as already noted, there is always some chance that the frequency and the single-case chance may differ as badly as you please That objection to frequentist analyses also turns out to be a consequence of the Principal Principle

EVIDENCE ABOUT CHANCES

To the subjectivist who believes in objective chance, particular or general propositions about chances are nothing special We believe them to varying degrees As new evidence arrives, our credence in them should wax and wane in accordance with Bayesian confirmation theory It is reasonable to believe such a proposition, like any other, to the degree given by a reasonable initial credence function conditionalized on one's present total evidence

If we look at the matter in closer detail, we find that the calculations of changing reasonable credence involve *likelihoods* credences of bits of evidence conditionally upon hypotheses Here the Principal Principle may act as a useful constraint Sometimes when the hypothesis concerns chance and the bit of evidence concerns the outcome, the reasonable likelihood is fixed, independently of the vagaries of initial credence and previous evidence What is more, the likelihoods are fixed in such a way that observed frequencies tend to confirm hypotheses according to which these frequencies differ not too much from uniform chances

To illustrate, let us return to our example of the sequence of coin tosses Think of it as an experiment, designed to provide evidence bearing on various hypotheses about the single-case chances of heads The sequence begins at time 1 and goes on for at least n tosses The evidence gained by the end of the experiment is a proposition F to the effect that the frequency of heads in the first n tosses was f (I assume that we use a mechanical counter that keeps no record of individual tosses The case in which there is a full record, however, is little different I also

assume, in an unrealistic simplification, that no other evidence whatever arrives during the experiment.) Suppose that at time 1 your credence function is $C(-/E)$, the function that comes from our chosen reasonable initial credence function C by conditionalizing on your total evidence E up to that time. Then if you learn from experience by conditionalizing, your credence function after the experiment is $C(-/FE)$. The impact of your experimental evidence F on your beliefs, about chances or anything else, is given by the difference between these two functions.

Suppose that before the experiment your credence is distributed over a range of alternative hypotheses about the endpoint chances of heads in the experimental tosses. (Your degree of belief that none of these hypotheses is correct may not be zero, but I am supposing it to be negligible and shall accordingly neglect it.) The hypotheses agree that these chances are uniform, and each independent of the previous course of history after time 1, but they disagree about what the uniform chance of heads is. Let us write G_h for the hypothesis that the endpoint chances of heads are uniformly h. Then the credences $C(G_h/E)$, for various h's, comprise the *prior distribution* of credence over the hypotheses, the credences $C(G_h/FE)$ comprise the *posterior distribution*, and the credences $C(F/G_hE)$ are the likelihoods. Bayes' Theorem gives the posterior distribution in terms of the prior distribution and the likelihoods:

$$C(G_h/FE) = \frac{C(G_h/E)\ \ C(F/G_hE)}{\Sigma_h\ [C(G_h/E)\ \ C(F/G_hE)]}$$

(Note that "h" is a bound variable of summation in the denominator of the right hand side, but a free variable elsewhere.) In words: to get the posterior distribution, multiply the prior distribution by the likelihood function and renormalize.

In talking only about a single experiment, there is little to say about the prior distribution. That does indeed depend on the vagaries of initial credence and previous evidence.

Not so for the likelihoods. As we saw in the last section, each G_h implies a proposition X_h to the effect that the chance at 1 of F equals x_h, where x_h is given by a certain function of h, n, and f. Hence G_hE and X_hG_hE are the same proposition. Further, G_hE and X are compatible (unless G_hE is itself impossible, in which case G_h might as well be omitted from the range of hypotheses). E is admissible at 1, being about matters of particular fact—your evidence—at times no later than 1. G_h also is admissible at 1. Recall from the last section that what

makes such a proposition hold at a world is a certain relationship between that world's complete history up to time 1 and that world's history-to-chance conditionals about the chances that would follow various complete extensions of that history Hence any member of the history-theory partition for time 1 either implies or contradicts G_b, G_b is therefore a disjunction of conjunctions of admissible historical propositions and admissible history-to-chance conditionals Finally, we supposed that C is reasonable So the Principal Principle applies

$$C(F/G_bE) = C(F/X_bG_bE) = x_b$$

The likelihoods are the endpoint chances, according to the various hypotheses, of obtaining the frequency of heads that was in fact obtained

When we carry the calculation through, putting these implied chances for the likelihoods in Bayes' theorem, the results are as we would expect An observed frequency of f raises the credences of the hypotheses G_b with b close to f at the expense of the others, the more sharply so, the greater is the number of tosses Unless the prior distribution is irremediably biased, the result after enough tosses is that the lion's share of the posterior credence will go to hypotheses putting the single-case chance of heads close to the observed frequency

CHANCE AS A GUIDE TO LIFE

It is reasonable to let one's choices be guided in part by one's firm opinions about objective chances or, when firm opinions are lacking, by one's degrees of belief about chances *Ceteris paribus*, the greater chance you think a lottery ticket has of winning, the more that ticket should be worth to you and the more you should be disposed to choose it over other desirable things Why so?

There is no great puzzle about why credence should be a guide to life Roughly speaking, what makes it be so that a certain credence function is *your* credence function is the very fact that you are disposed to act in more or less the ways that it rationalizes (Better what makes it be so that a certain reasonable initial credence function and a certain reasonable system of basic intrinsic values are both yours is that you are disposed to act in more or less the ways that are rationalized by the pair of them together, taking into account the modification of credence by conditionalizing on total evidence, and further, you would have been likewise disposed if your life history of experience, and conse-

quent modification of credence, had been different, and further, no other such pair would fit your dispositions more closely) No wonder your credence function tends to guide your life If its doing so did not accord to some considerable extent with your dispositions to act, then it would not be your credence function You would have some other credence function, or none

If your present degrees of belief are reasonable—or at least if they come from some reasonable initial credence function by conditionalizing on your total evidence—then the Principal Principle applies Your credences about outcomes conform to your firm beliefs and your partial beliefs about chances Then the latter guide your life because the former do The greater chance you think the ticket has of winning, the greater should be your degree of belief that it will win, and the greater is your degree of belief that it will win, the more, *ceteris paribus*, it should be worth to you and the more you should be disposed to choose it over other desirable things

PROSPECTS FOR AN ANALYSIS OF CHANCE

Consider once more the Principal Principle as reformulated

$$P_{tw}(A) = C(A/H_{tw}T_w)$$

Or in words the chance distribution at a time and a world comes from any reasonable initial credence function by conditionalizing on the complete history of the world up to the time, together with the complete theory of chance for the world

Doubtless it has crossed your mind that this has at least the form of an analysis of chance But you may well doubt that it is informative as an analysis, that depends on the distance between the analysandum and the concepts employed in the analysans

Not that it has to be informative *as an analysis* to be informative I hope I have convinced you that the Principal Principle is indeed informative, being rich in consequences that are central to our ordinary ways of thinking about chance

There are two different reasons to doubt that the Principal Principle qualifies as an analysis The first concerns the allusion in the analysans to reasonable initial credence functions The second concerns the allusion to complete theories of chance In both cases the challenge is the same could we possibly get any independent grasp on this concept, otherwise than by way of the concept of chance itself? In both

cases my provisional answer is most likely not, but it would be worth trying Let us consider the two problems in turn

It would be natural to think that the Principal Principle tells us nothing at all about chance, but rather tells us something about what makes an initial credence function be a reasonable one To be reasonable is to conform to objective chances in the way described Put this strongly, the response is wrong the Principle has consequences, as we noted, that are about chance and not at all about its relationship to credence (They would be acceptable, I trust, to a believer in objective single-case chance who rejects the very idea of degree of belief) It tells us more than nothing about chance But perhaps it is divisible into two parts one part that tells us something about chance, another that takes the concept of chance for granted and goes on to lay down a criterion of reasonableness for initial credence

Is there any hope that we might leave the Principal Principle in abeyance, lay down other criteria of reasonableness that do not mention chance, and get a good enough grip on the concept that way? It's a lot to ask For note that just as the Principal Principle yields some consequences that are entirely about chance, so also it yields some that are entirely about reasonable initial credence One such consequence is as follows There is a large class of propositions such that if Y is any one of these, and C_1 and C_2 are any two reasonable initial credence functions, then the functions that come from C_1 and C_2 by conditionalizing on Y are exactly the same (The large class is, of course, the class of members of history-theory partitions for all times) That severely limits the ways that reasonable initial credence functions may differ, and so shows that criteria adequate to pick them out must be quite strong What might we try? A reasonable initial credence function ought to (1) obey the laws of mathematical probability theory, (2) avoid dogmatism, at least by never assigning zero credence to possible propositions and pehaps also by never assigning infinitesimal credence to certain kinds of possible propositions, (3) make it possible to learn from experience by having a built-in bias in favor of worlds where the future in some sense resembles the past, and perhaps (4) obey certain carefully restricted principles of indifference, thereby respecting certain symmetries Of these, critera (1)–(3) are all very well, but surely not yet strong enough Given C_1 satisfying (1)–(3), and given any proposition Y that holds at more than one world, it will be possible to distort C_1 very slightly to produce C_2, such that $C_1(—/Y)$ and $C_2(—/Y)$ differ but C_2 also satisfies (1)–(3) It is less clear what (4) might be able to do for us Mostly that is because (4) is less clear *sim-*

plicater, in view of the fact that it is not possible to obey too many different restricted principles of indifference at once and it is hard to give good reasons to prefer some over their competitors It also remains possible, of course, that some criterion of reasonableness along different lines than any I have mentioned would do the trick

I turn now to our second problem the concept of a complete theory of chance In saying what makes a certain proposition be the complete theory of chance for a world (and for any world where it holds), I gave an explanation in terms of chance Could these same propositions possibly be picked out in some other way, without mentioning chance?

The question turns on an underlying metaphysical issue A broadly Humean doctrine (something I would very much like to believe if at all possible) holds that all the facts there are about the world are particular facts, or combinations thereof This need not be taken as a doctrine of analyzability, since some combinations of particular facts cannot be captured in any finite way It might be better taken as a doctrine of supervenience if two worlds match perfectly in all matters of particular fact, they match perfectly in all other ways too—in modal properties, laws, causal connections, chances, It seems that if this broadly Humean doctrine is false, then chances are a likely candidate to be the fatal counter-instance And if chances are not supervenient on particular fact, then neither are complete theories of chance For the chances at a world are jointly determined by its complete theory of chance together with propositions about its history, which latter plainly are supervenient on particular fact

If chances are not supervenient on particular fact, then neither chance itself nor the concept of a complete theory of chance could possibly be analyzed in terms of particular fact, or of anything supervenient thereon The only hope for an analysis would be to use something in the analysans which is itself not supervenient on particular fact I cannot say what that something might be

How might chance, and complete theories of chance, be supervenient on particular fact? Could something like this be right the complete theory of chance for a world is that one of all possible complete theories of chance that somehow best fits the global pattern of outcomes and frequencies of outcomes? It could not For consider any such global pattern, and consider a time long before the pattern is complete At that time, the pattern surely has some chance of coming about and some chance of not coming about There is surely some chance of a very different global pattern coming about, one which, according to the proposal under consideration, would make true some different

complete theory of chance But a complete theory of chance is not something that could have some chance of coming about or not coming about By the Principal Principle,

$$P_{tw}(T_w) = C(T_w/H_{tw}T_w) = 1$$

If T_w is something that holds in virtue of some global pattern of particular fact that obtains at world w, this pattern must be one that has no chance at any time (at w) of not obtaining If w is a world where many matters of particular fact are the outcomes of chance processes, then I fail to see what kind of global pattern this could possibly be

But there is one more alternative I have spoken as if I took it for granted that different worlds have different history-to-chance conditionals, and hence different complete theories of chance Perhaps this is not so perhaps all worlds are exactly alike in the dependence of chance on history Then the complete theory of chance for every world, and all the conditionals that comprise it, are necessary They are supervenient on particular fact in the trivial way that what is noncontingent is supervenient on anything—no two worlds differ with respect to it Chances are still contingent, but only because they depend on contingent historical propositions (information about the details of the coin and tosser, as it might be) and not also because they depend on a contingent theory of chance Our theory is much simplified if this is true Admissible information is simply historical information, the history-theory partition at t is simply the partition of alternative complete histories up to t, for any reasonable initial credence function C

$$P_{tw}(A) = C(A/H_{tw}),$$

so that the chance distribution at t and w comes from C by conditionalizing on the complete history of w up to t Chance is reasonable credence conditional on the whole truth about history up to a time The broadly Humean doctrine is upheld, so far as chances are concerned what makes it true at a time and a world that something has a certain chance of happening is something about matters of particular fact at that time and (perhaps) before

What's the catch? For one thing, we are no longer safely exploring the consequences of the Principal Principle, but rather engaging in speculation For another, our broadly Humean speculation that history-to-chance conditionals are necessary solves our second problem by making the first one worse Reasonable initial credence functions are constrained more narrowly than ever Any two of them, C_1 and C_2, are now required to yield the same function by conditionalizing on the com-

plete history of any world up to any time Put it this way according to our broadly Humean speculation (and the Principal Principle) if I were perfectly reasonable and knew all about the course of history up to now (no matter what that course of history actually is, and no matter what time is now) then there would be only one credence function I could have Any other would be unreasonable

It is not very easy to believe that the requirements of reason leave so little leeway as that Neither is it very easy to believe in features of the world that are not supervenient on particular fact But if I am right, that seems to be the choice I shall not attempt to decide between the Humean and the anti-Humean variants of my approach to credence and chance The Principal Principle doesn't

REFERENCES

Bernstein, Allen R and Wattenberg, Frank, "Non-Standard Measure Theory", in *Applications of Model Theory of Algebra, Analysis, and Probability*, ed by W Luxemburg, Holt, Reinhart, and Winston, 1969

Carnap, Rudolf, "The Two Concepts of Probability", *Philosophy and Phenomenological Research* 5 (1945), 513–32

Jeffrey, Richard C , *The Logic of Decision*, McGraw-Hill, 1965

Jeffrey, Richard C , review of articles by David Miller *et al* , *Journal of Symbolic Logic* 35 (1970), 124–27

Jeffrey, Richard C , ' Mises Redux", in *Basic Problems in Methodology and Linguistics Proceedings of the Fifth International Congress of Logic, Methodology and Philosophy of Science*, Part III, ed by R Butts and J Hintikka, D Reidel, Dordrecht, Holland, 1977

Lewis, David, "Counterpart Theory and Quantified Modal Logic", *Journal of Philosophy* 65 (1968), 113–26

Lewis, David, *Counterfactuals*, Blackwell, 1973

Mellor, D H , *The Matter of Chance*, Cambridge University Press, 1971

Quine, W V , "Propositional Objects' , in *Ontological Relativity and Other Essays*, Columbia University Press, 1969

Railton, Peter, "A Deductive-Nomological Model of Probabilistic Explanation", *Philosophy of Science* 45 (1978), 206–26

Skyrms, Brian, "Resiliency, Propensities, and Causal Necessity", *Journal of Philosophy* 74 (1977), 704–13

Postscripts to
"A Subjectivist's Guide to Objective Chance"

A NO ASSISTANCE NEEDED[1]

Henry Kyburg doubts that the Principal Principle has as much scope as my praise of it would suggest He offers a continuation of my questionnaire, says that his added questions fall outside the scope of the Principal Principle, and suggests that we need some Assistant Principle to deal with them His first added question is as follows [2]

> *Question* You are sure that a certain coin is fair It was tossed this morning, but you have no information about the outcome of the toss To what degree should you believe the proposition that it landed heads?
>
> *Answer* 50 per cent, of course

That's the right answer (provided the question is suitably interpreted) But the Principal Principle, unassisted, does suffice to yield that answer What we must bear in mind is that the Principle relates time-dependent chance to time-dependent admissibility of evidence, and that it applies to any time, not only the present

Kyburg thinks the Principle falls silent "since there is *no* chance that the coin fell other than the way it did," and quotes me to the effect that "what's past is no longer chancy " Right We won't get anywhere if we apply the Principle to *present* chances But what's past *was* chancy, if indeed the coin was fair, so let's see what we get by applying the Principle to a past time, and working back to present credences Notation

[1] In writing this postscript I have benefited from a discussion by W N Reinhardt (personal communication 1982) Reinhardt s treatment and mine agree on most but not all points

[2] Henry E Kyburg Jr , Principle Investigation, *Journal of Philosophy* 78 (1981) 772–78

t a time just before the toss,
C a reasonable initial credence function that will yield my later credences by conditionalizing on total evidence,
C_0 my present credence function,
A the proposition that the coin fell heads,
X the proposition that the coin was fair, that is that its chance at t of falling heads was 50%,
E the part of my present total evidence that is admissible at t,
F the rest of my present total evidence

Since *ex hypothesi* I'm certain of X, we have

(1) $C_0 = C_0(-/X)$

By definition of C, we have

(2) $C_0 = C(-/EF)$

Assuming that F is irrelevant to the tosses, we have

(3) $C(A/XEF) = C(A/XE)$

By the Principal Principle, applied not to the present but to t, we have

(4) $C(A/XE) = 50\%$

Now, by routine calculation from (1)–(4) we have

(5) $C_0(A) = 50\%$

which answers Kyburg's question

Step (3) deserves further examination, lest you suspect it of concealing an Assistant Principle Recall that F is the part of my present total evidence that was not admissible already at time t Presumably it consists of historical information about the interval between t and the present For historical information about earlier times would be already admissible at t, and historical information about later times, or nonhistorical information, could scarcely be part of my present total evidence (Here, as in the paper, I set aside strange possibilities in which the normal asymmetries of time break down So far as I can tell, Kyburg is content to join me in so doing) Thus if I had watched the toss, or otherwise received information about its outcome, that information would be included in F

However, Kyburg stipulated in his question that "you have no information about the outcome of the toss" We might reasonably construe that to mean that no information received between t and the present is

evidentially relevant to whether the coin fell heads, with evidential relevance construed in the usual way in terms of credence Then (3) comes out as a stipulated condition of the problem, not some extra principle

There is a different, stricter way that Kyburg's stipulation might perhaps be construed It might only exclude information that settles the outcome decisively, leaving it open that I have information that bears evidentially on the outcome without settling it For instance, it might be that the tosser promised to phone me if the toss fell heads, I got no phone call, but that is far from decisive because my phone is not reliable On that construal, we are not entitled to assume (3) But on that construal Kyburg's answer is wrong, or anyway it isn't right as a matter of course on the basis of what he tells us, so we don't want any principle that delivers that answer

Kyburg has a second added question to challenge the Principal Principle

> *Question* As above, but you know that the coin was tossed 100 times, and landed heads 86 times To what degree should you believe the proposition that it landed heads on the first toss?
>
> *Answer* 86 per cent

The strategy for getting the Principal Principle to yield an answer is the same as before, but the calculation is more complicated Notation as before, except for

- A the proposition that the coin fell heads *on the first toss*,
- B the proposition that the coin fell heads 86 times out of 100,
- X the proposition that the coin was fair, that is that its chance at t of falling heads was 50% *on each toss*,
- F the rest of my present total evidence, besides the part that was admissible at t, *and also besides the part* B,
- x the fraction of heads-tails sequences of length 100 in which there are 86 heads

Our equations this time are as follows They are justified in much the same way as the like-numbered equations above But this time, to get the new (2) we split the present total evidence into three parts B, E, and F And to get the new (4), we use the Principal Principle repeatedly to multiply endpoint chances, as was explained in the section of the paper dealing with chance of frequency

(1) As before,

(2) $C_0 = C(-/BEF)$,

(3) $C(A/XBEF) = C(A/XBE)$,

(4) $C(AB/XE) = x$ 86%, $C(B/XE) = x$,

(5) $C_0(A) = 86\%$

Kyburg also thinks I need an extra "Principle of Integration" which I neglected to state But this principle, it turns out, has nothing especially to do with chance! It is just a special case of a principle of infinite additivity for credences Indeed it could be replaced, at the point where he claims I tacitly used it, by *finite* additivity of credences (And finite additivity goes without saying, though I nevertheless did say it) To be sure, if we want to treat credences in the setting of nonstandard analysis, we are going to want some kind of infinite additivity And some kind of infinite additivity comes automatically when we start with finite additivity and then treat some infinite sets as if they were finite It is an interesting question what kind of infinite addivity of credences we can reasonably assume in the nonstandard setting But this question belongs entirely to the theory of credence—not to the connection between chance and credence that was the subject of my paper

B CHANCE WITHOUT CHANCE?

Isaac Levi thinks that I have avoided confronting "the most important problem about chance", which problem, it seems, is the reconciliation of chances with determinism, or of chances with different chances [3] Consider a toss of coin Levi writes that

> in typical cases, the agent will and should be convinced that information exists (though inaccessible to him) which is highly relevant [to the outcome] Thus, the agent may well be convinced that a complete history through [the onset of the toss] will include a specification of the initial mechanical state of the coin upon being tossed and boundary conditions which, taken together, determine the outcome to be heads up or tails up according to physical laws
>
> given the available knowledge of physics, we cannot [deny that the mechanical state of the coin at the onset of the toss determines the out-

[3] Isaac Levi review of *Studies in Inductive Logic and Probability* ed by R C Jeffrey *Philosophical Review* 92 (1983) 120–21

> come] provided we can assume the motion of the coin to be sealed off from substantial external influences But even if we allow for fluctuations in the boundary conditions, we would not suppose them so dramatic as to permit large deviations from 0 or 1 to be values of the chances of heads

And yet

> Lewis, however, appears ready to assign 5 to the chance of [the] coin landing heads up

So how do I square the supposition that the chance of heads is 50% with the fact that it is zero or one, or anyway it does not deviate much from zero or one?

I don't If the chance is zero or one, or close to zero or one, then it cannot also be 50% To the question how chance can be reconciled with determinism, or to the question how disparate chances can be reconciled with one another, my answer is *it can't be done*

It was not I, but the hypothetical "you" in my example, who appeared ready to assign a 50% chance of heads If my example concerned the beliefs of an ignoramus, it is none the worse for that

I myself am in a more complicated position than the character in this example (That is why I made an example of him, not me) I would not give much credence to the proposition that the coin has a chance of heads of 50% exactly I would give a small share of credence to the proposition that it is zero exactly, and an equal small share to the proposition that it is one exactly I would divide most of the rest of my credence between the vicinity of 50%, the vicinity of zero, and the vicinity of one

The small credence I give to the extremes, zero and one exactly, reflects my slight uncertainty about whether the world is chancy at all Accepted theory says it is, of course, but accepted theory is not in the best of foundational health, and the sick spot (reduction of the wave function brought on by measurement) is the very spot where the theory goes indeterministic But most of my credence goes to the orthodox view that there are plenty of chance processes in microphysics And not just the microphysics of extraordinary goings-on in particle accelerators! No, for instance the making and breaking of chemical bonds is chancy, so is the coherence of solids that stick together by means of chemical bonding, so is the elasticity of collisions between things that might bond briefly before they rebound, So is any process whatever that could be disrupted by chance happenings nearby—and infallible "sealing off" is not to be found

In Levi's physics, a coin coming loose from fingers and tumbling in

air until it falls flat on a table is a classical system, an oasis of determinism in a chancy microworld I do not see how that can be The coin, and the fingers and the air and the table, are too much a part of that microworld There are also the external influences, which cannot be dismissed either by requiring them to be substantial or by invoking fictitious seals, but never mind, let us concentrate on the toss itself There is chance enough in the processes by which the coin leaves the fingers, in the processes whereby it bounces off air molecules and sends them recoiling off, perhaps to knock other molecules into its path, in the process whereby the coin does or doesn't stretch a bit as it spins, thereby affecting its moment of inertia, and in the processes whereby it settles down after first touching the table In ever so many minute ways, what happens to the coin is a matter of chance

But all those chance effects are *so* minute —But a tossed coin is *so* sensitive to minute differences Which dominates—minuteness or sensitivity? That is a question to be settled not by asking what a philosopher would find it reasonable to suppose, but by calculation The calculations would be difficult We may not make them easier by approximations in which expected values replace chance distributions I have not heard of anyone who has attempted these calculations, and of course they are far beyond my own power Maybe they are beyond the state of the art altogether Without them, I haven't a clue whether the minuteness of the chance effects dominates, in which case the chance of heads is indeed close to zero or one, or whether instead the sensitivity dominates, in which case the chance of heads is close to 50% Hence my own distribution of credence

The hypothetical "you" in my example has a different, simpler distribution Why? He might be someone who has done the calculations and found that the sensitivity dominates Or he might have been so foolish as to intuit that the sensitivity would dominate Or he might be altogether misinformed

Well-informed people often say that ordinary gambling devices are deterministic systems Why? Perhaps it is a hangover of instrumentalism If we spoke as instrumentalists, we would be right to say so—meaning thereby not that they really *are* deterministic, but rather that it is sometimes instrumentally useful to pretend that they are To the extent that it is feasible to predict gambling devices at all—we can't predict heads or tails, but we can predict, for instance, that the coin won't tumble in mid-air until next year, and won't end up sticking to the wall—deterministic theories are as good predictive instruments as can be had Perhaps when the instrumentalist expert says that tossed

coins are deterministic, the philosopher misunderstands him, and thinks he means that tossed coins are deterministic

Can it be that Levi himself was speaking as an instrumentalist in the passages I cited? If so, then the problem of reconciling chance and determinism is not very hard It is just the problem of reconciling truth *simpliciter* with truth in fiction In truth, nobody lived at 221B Baker Street, in fiction, Holmes lived there In truth, most likely, the coin is chancy, in fiction, it is deterministic No worries The character in my example, of course, was meant to be someone who believed that the chance of heads was 50% in truth—not in fiction, however instrumentally useful such fiction might be

There is no chance without chance If our world is deterministic there are no chances in it, save chances of zero and one Likewise if our world somehow contains deterministic enclaves, there are no chances in those enclaves If a determinist says that a tossed coin is fair, and has an equal chance of falling heads or tails, he does not mean what I mean when he speaks of chance Then what *does* he mean? This, I suppose, is the question Levi would like to see addressed It is, of course, a more urgent question for determinists than it is for me

That question has been sufficiently answered in the writings of Richard Jeffrey and Brian Skyrms on objectified and resilient credence [4] Without committing themselves one way or the other on the question of determinism, they have offered a kind of counterfeit chance to meet the needs of the determinist It is a relative affair, and apt to go indeterminate, hence quite unlike genuine chance But what better could a determinist expect?

According to my second formulation of the Principal Principle, we have the history-theory partition (for any given time), and the chance distribution (for any given time and world) comes from any reasonable initial credence function by conditionalizing on the true cell of this partition That is, it is objectified in the sense of Jeffrey Let us note three things about the history-theory partition

(1) It seems to be a natural partition, not gerrymandered It is what we get by dividing possibilities as finely as possible in certain straightforward respects

[4] Richard C Jeffrey, *The Logic of Decision* (New York McGraw-Hill 1965 second edition, Chicago University of Chicago Press 1983) Section 12 7, Brian Skyrms, Resiliency Propensities and Causal Necessity, *Journal of Philosophy* 74 (1977) 704–13, Brian Skyrms, *Causal Necessity* (New Haven Yale University Press, 1980)

(2) It is to some extent feasible to investigate (before the time in question) which cell of this partition is the true cell, but
(3) it is unfeasible (before the time in question, and without peculiarities of time whereby we could get news from the future) to investigate the truth of propositions that divide the cells

Hence if we start with a reasonable initial credence function and do enough feasible investigation, we may expect our credences to converge to the chances, and no amount more feasible investigation (before the time) will undo that convergence That is, after enough investigation, our credences become resilient in the sense of Skyrms And our credences conditional on cells of the partition are resilient from the outset

Conditions (1)–(3) characterize the history-theory partition, but not uniquely Doubtless there are other, coarser partitions, that also satisfy the conditions How feasible is feasible? Some investigations are more feasible than others, depending on the resources and techniques available, and there must be plenty of boundaries to be drawn between the feasible and the unfeasible before we get to the ultimate boundary whereby investigations that divide the history-theory cells are the most unfeasible of all Any coarser partition, if it satisfies conditions (1)–(3) according to some appropriate standards of feasible investigation and of natural partitioning, gives us a kind of counterfeit chance suitable for use by determinists namely, reasonable credence conditional on the true cell of that partition Counterfeit chances will be relative to partitions, and relative, therefore, to standards of feasibility and naturalness, and therefore indeterminate unless the standards are somehow settled, or at least settled well enough that all remaining candidates for the partition will yield the same answers Counterfeit chances are therefore not the sort of thing we would want to find in our fundamental physical theories, or even in our theories of radioactive decay and the like But they will do to serve the conversational needs of determinist gamblers

C LAWS OF CHANCE

Despite the foundational problems of quantum mechanics, it remains a good guess that many processes are governed by probabilistic laws of nature These laws of chance, like other laws of nature, have the form of universal generalizations Just as some laws concern forces, which are magnitudes pertaining to particulars, so some laws concern single-case chances, which likewise are magnitudes pertaining to particulars

For instance, a law of chance might say that for any tritium atom and any time when it exists, there is such-and-such chance of that atom decaying within one second after that time [5] What makes it at least a regularity—a true generalization—is that for each tritium atom and time, the chance of decay is as the law says it is What makes it a law, I suggest, is the same thing that gives some others regularities the status of laws it fits into some integrated system of truths that combines simplicity with strength in the best way possible [6]

This is a kind of regularity theory of lawhood, but it is a collective and selective regularity theory Collective, since regularities earn their lawhood not by themselves, but by the joint efforts of a system in which they figure either as axioms or as theorems Selective, because not just any regularity qualifies as a law If it would complicate the otherwise best system to include it as an axiom, or to include premises that would imply it, and if it would not add sufficient strength to pay its way, then it is left as a merely accidental regularity

Five remarks about the best-system theory of lawhood may be useful before we return to our topic of how this theory works in the presence of chance

[5] Peter Railton employs laws of chance of just this sort to bring probabilistic explanation under the deductive-nomological model The outcome itself cannot be deduced, of course, but the single case chance of it can be See Railton, A Deductive Nomological Model of Probabilistic Explanation, *Philosophy of Science* 45 (1978) 206–26, and the final section of my Causal Explanation in this volume

[6] I advocate a best system theory of lawhood in *Counterfactuals* (Oxford Blackwell, 1973), pp 73–75 Similar theories of lawhood were held by Mill and, briefly by Ramsey See John Stuart Mill, *A System of Logic* (London Parker, 1843), Book III, Chapter IV, Section 1, and F P Ramsey Universals of Law and of Fact, in his *Foundations* (London Routledge & Kegan Paul, 1978) For further discussion, see John Earman, Laws of Nature The Empiricist Challenge, in *D M Armstrong*, ed by Radu J Bogdan (Dordrecht Reidel, 1984)

Mill s version is not quite the same as mine He says that the question what are the laws of nature could be restated thus What are the fewest general propositions from which all the uniformities which exist in the universe might be deductively inferred? , so it seems that the ideal system is supposed to be complete as regards uniformities, that it may contain only general propositions as axioms, and that its theorems do not qualify as laws

It is not clear to me from his brief statement whether Ramsey's version was quite the same as mine His summary statement (after changing his mind) that he had taken laws to be consequences of those propositions we should take as axioms if we knew everything and organized it as simply as possible into a deductive system (*Foundations*, p 138) is puzzling Besides Ramsey s needless mention of knowledge, his it with antecedent everything suggests that the ideal system is supposed to imply everything true Unless Ramsey made a stupid mistake, which is impossible, that can not have been his intent, it would make all regularities come out as laws

(1) The standards of simplicity, of strength, and of balance between them are to be those that guide us in assessing the credibility of rival hypotheses as to what the laws are In a way, that makes lawhood depend on us—a feature of the approach that I do not at all welcome! But at least it does not follow that lawhood depends on us in the most straightforward way namely, that if our standards were suitably different, then the laws would be different For we can take our actual standards as fixed, and apply them in asking what the laws would be in various counterfactual situations, including counterfactual situations in which people have different standards—or in which there are no people at all Likewise, it fortunately does not follow that the laws are different at other times and places where there live people with other standards

(2) On this approach, it is not to be said that certain generalizations are *lawlike* whether or not they are true, and the laws are exactly those of the lawlikes that are true There will normally be three possibilities for any given generalization that it be false, that it be true but accidental, and that it be true as a law Whether it is true accidentally or as a law depends on what else is true along with it, thus on what integrated systems of truths are available for it to enter into To illustrate the point it may be true accidentally that every gold sphere is less than one mile in diameter, but if gold were unstable in such a way that there was no chance whatever that a large amount of gold could last long enough to be formed into a one-mile sphere, then this same generalization would be true as a law

(3) I do not say that the competing integrated systems of truths are to consist entirely of regularities, however, only the regularities in the best system are to be laws It is open that the best system might include truths about particular places or things, in which case there might be laws about these particulars As an empirical matter, I do not suppose there are laws that essentially mention Smith's garden, the center of the earth or of the universe, or even the Big Bang But such laws ought not to be excluded *a priori* [7]

(4) It will trivialize our comparisons of simplicity if we allow our competing systems to be formulated with just any hoked-up primi-

[7] In defense of the possibility that there might be a special law about the fruit in Smith s garden, see Michael Tooley, The Nature of Laws, *Canadian Journal of Philosophy* 7 (1977) 667–98 especially p 687, and D M Armstrong, *What is a Law of Nature?* (Cambridge Cambridge University Press 1983), Sections 3 I, 3 II, and 6 VII In The Universality of Laws, *Philosophy of Science* 45 (1978) 173–81, John Earman observes that the best system theory of lawhood avoids any *a priori* guarantee that the laws will satisfy strong requirements of universality

tives So I take it that this kind of regularity theory of lawhood requires some sort of inegalitarian theory of properties simple systems are those that come out formally simple when formulated in terms of perfectly natural properties Then, sad to say, it's useless (though true) to say that the natural properties are the ones that figure in laws [8]

(5) If two or more systems are tied for best, then certainly any regularity that appears in all the tied systems should count as a law But what of a regularity that appears in some but not all of the tied systems? We have three choices it is not a law (take the intersection of the tied systems), it is a law (take the union), it is indeterminate whether it is law (apply a general treatment for failed presuppositions of uniqueness) If required to choose, I suppose I would favor the first choice, but it seems a reasonable hope that nature might be kind to us, and put some one system so far out front that the problem will not arise Likewise, we may hope that some system will be so far out front that it will win no matter what the standards of simplicity, strength, and balance are, within reason If so, it will also not matter if these standards themselves are unsettled To simplify, let me ignore the possibility of ties, or of systems so close to tied that indeterminacy of the standards matters, if need be, the reader may restore the needed complications

To return to laws of chance if indeed there are chances, they can be part of the subject matter of a system of truths, then regularities about them can appear as axioms or theorems of the best system, then such regularities are laws Other regularities about chances might fail to earn a place in the best system, those ones are accidental All this is just as it would be for laws about other magnitudes So far, so good

But there is a problem nearby, not especially a problem about laws of chance, but about laws generally in a chancy world We have said that a regularity is accidental if it cannot earn a place in the best system if it is too weak to enter as an axiom, and also cannot be made to follow as a theorem unless by overloading the system with particular information That is one way to be accidental, but it seems that a regularity might be accidental also for a different and simpler reason It might hold merely by chance It might be simple and powerful and well deserve a place in the ideal system and yet be no law For it might have, or it might once have had, some chance of failing to hold, whereas it seems very clear, *contra* the best-system theory as so far stated, that no genuine law ever could have had any chance of not holding A world of

[8] See my New Work for a Theory of Universals, *Australasian Journal of Philosophy* 61 (1983) 343–77 especially pp 366–68

lawful chance might have both sorts of accidental regularities, some disqualified by their inadequate contribution to simplicity and strength and others by their chanciness

Suppose that radioactive decay is chancy in the way we mostly believe it to be Then for each unstable nucleus there is an expected lifetime, given by the constant chance of decay for a nucleus of that species It might happen—there is some chance of it, infinitesimal but not zero—that each nucleus lasted for precisely its expected lifetime, no more and no less Suppose that were so The regularity governing lifetimes might well qualify to join the best system, just as the corresponding regularity governing *expected* lifetimes does Still, it is not a law For if it were a law, it would be a law with some chance—in fact, an overwhelming chance—of being broken That cannot be so [9]

(Admittedly, we do speak of defeasible laws, laws with exceptions, and so forth But these, I take it, are rough-and-ready approximations to the real laws There real laws have no exceptions, and never had any chance of having any)

Understand that I am not supposing that the constant chances of decay are *replaced* by a law of constant lifetimes That is of course possible What is not possible, unfortunately for the best-system theory, is for the constant chances to remain and to coexist with a law of constant lifetimes

If the lifetimes chanced to be constant, and if the matter were well investigated, doubtless the investigators would come to believe in a law of constant lifetimes But they would be mistaken, fooled by a deceptive coincidence It is one thing for a regularity to be a law, another thing for it to be so regarded, however reasonably Indeed, there are philosphers who seem oblivious to the distinction, but I think these philosophers misrepresent their own view They are sceptics, they do not believe in laws of nature at all, they resort to regarded-as-law regularities as a substitute, and they call their substitute by the name of the real thing

[9] At this point I am indebted to correspondence and discussion with Frank Jackson arising out of his discussion of Hume worlds in A Causal Theory of Counterfactuals, *Australasian Journal of Philosophy* 55 (1977) 3–21, especially pp 5–6 A Hume world, as Jackson describes it, is a possible world where every particular fact is as it is in our world, but there are no causes or effects at all Every regular conjunction is an accidental one, not a causal one I am not sure whether Jackson s Hume world is one with chances — lawless chances, of course — or without In the former case the bogus laws of the Hume world would be like our bogus law of constant lifetimes, but on a grander scale

So the best-system theory of lawhood, as it stands, is in trouble I propose this correction Previously, we held a competition between all true systems Instead, let us admit to the competition only those systems that are true not by chance, that is, those that not only are true, but also have never had any chance of being false The field of eligible competitors is thus cut down But then the competition works as before The best system is the one that achieves as much simplicity as is possible without excessive loss of strength, and as much strength as is possible without excessive loss of simplicity A law is a regularity that is included, as an axiom or as a theorem, in the best system

Then a chance regularity, such as our regularity of constant lifetimes, cannot even be included in any of the competing systems *A fortiori*, it cannot be included in the best of them Then it cannot count as a law It will be an accidental regularity, and for the right reason because it had a chance of being false Other regularities may still be accidental for our original reason These would be regularities that never had any chance of being false, but that don't earn their way into the best system because they don't contribute enough to simplicity and strength For instance suppose that (according to regularities that do earn a place in the best system) a certain quantity is strictly conserved, and suppose that the universe is finite in extent Then we have a regularity to the effect that the total of this quantity, over the entire universe, always equals a certain fixed value This regularity never had any chance of being false But it is not likely to earn a place in the best system and qualify as a law

In the paper, I made much use of the history-to-chance conditionals giving hypothetical information about the chance distribution that would follow a given (fully specified) initial segment of history Indeed, my reformulation of the Principal Principle involves a "complete theory of chance" which is the conjunction of all such history-to-chance conditionals that hold at a given world, and which therefore fully specifies the way chances at any time depend on history up to that time

It is to be hoped that the history-to-chance conditionals will follow, entirely or for the most part, from the laws of nature, and, in particular, from the laws of chance We might indeed impose a requirement to that effect on our competing systems I have chosen not to While the thesis that chances might be entirely governed by law has some plausibility, I am not sure whether it deserves to be built into the analysis of lawhood Perhaps rather it is an empirical thesis a virtue that we may hope distinguishes our world from more chaotic worlds

At any rate, we can be sure that the history-to-chance conditionals

will not conflict with the system of laws of chance Not, at any rate, in what they say about the outcomes and chances that would follow any initial segment of history that ever had any chance of coming about Let H be a proposition fully specifying such a segment Let t be a time at which there was some chance that H would come about Let L be the conjunction of the laws There was no chance, at t, of L being false Suppose for *reductio* first that we have a history-to-chance conditional "if H, then A" (where A might, for instance, specify chances at the end-time of the segment), and second that H and L jointly imply not-A, so that the conditional conflicts with the laws The conditional had no chance at t of being false—this is an immediate consequence of the reformulated Principal Principle Since we had some chance at t of H, we had some chance of H holding along with the conditional, hence some chance of H and A And since there was no chance that L would be false, there was some chance that all of H, A, and L would hold together, so some chance at t of a contradiction Which is impossible there never can be any chance of a contradiction

A more subtle sort of conflict also is ruled out Let t, L, and H be as before Suppose for *reductio* first that we have a history-to-chance conditional "if H, then there would be a certain positive chance of A", and second that H and L jointly imply not-A This is not the same supposition as before after all, it would be no contradiction if something had a positive chance and still did not happen But it is still a kind of conflict the definiteness of the law disagrees with the chanciness of the conditional To rule it out, recall that we had at t some chance of H, but no chance of the conditional being false, so at t there was a chance of H holding along with the conditional, so at t there was a chance that, later, there would be a chance of A following the history H, but chanciness does not increase with time (assuming, as always, the normal asymmetries), an earlier chance of a later chance of something implies an earlier chance of it, so already at t there was some chance of H and A holding together Now we can go on as before we have that at t there was no chance that L would be false, so some chance that all of H, A, and L would hold together, so some chance at t of a contradiction, which is impossible

The best-system theory of lawhood in its original form served the cause of Humean supervenience History, the pattern of particular fact throughout the universe, chooses the candidate systems, and the standards of selection do the rest So no two worlds could differ in laws without differing also in their history But our correction spoils that The laws—laws of chance, and other laws besides—supervene now on

the pattern of particular chances If the chances in turn somehow supervene on history, then we have Humean supervenience of the laws as well, if not, not The corrected theory of lawhood starts with the chances It does nothing to explain them

Once, *circa* 1975, I hoped to do better to extend the best-system approach in such a way that it would provide for the Humean supervenience of chances and laws together, in one package deal This was my plan We hold a competition of deductive systems, as before, but we impose less stringent requirements of eligibility to enter the competition, and we change the terms on which candidate systems compete We no longer require a candidate system to be entirely true, still less do we require that it never had any chance of being false Instead, we only require that a candidate system be true in what it says about history, we leave it open, for now, whether it also is true in what it says about chances We also impose a requirement of coherence each candidate system must imply that the chances are such as to give that very system no chance at any time of being false Once we have our competing systems, they vary in simplicity and in strength, as before But also they vary in what I shall call *fit* a system fits a world to the extent that the history of that world is a comparatively probable history according to that system (No history will be very probable, in fact, any history for a world like ours will be very improbable according to any system that deserves in the end to be accepted as correct, but still, some are more probable than others) If the histories permitted by a system formed a tree with finitely many branch points and finitely many alternatives at each point, and the system specified chances for each alternative at each branch point, then the fit between the system and a branch would be the product of these chances along that branch, and likewise, somehow, for the general, infinite case (Never mind the details if, as I think, the plan won't work anyway) The best system will be the winner, now, in a three-way balance between simplicity, strength, and fit As before, the laws are the generalizations that appear as axioms or theorems in the best system, further, the true chances are the chances as they are according to the best system So it turns out that the best system is true in its entirety— true in what it says about chances, as well as in what it says about history So the laws of chance, as well as other laws, turn out to be true, and further, to have had no chance at any time of being false We have our Humean supervenience of chances and of laws, because history selects the candidate systems, history determines how well each one fits, and our standards of selection do the rest We will tend, *ceteris paribus*, to get the proper agreement

between frequencies and uniform chances, because that agreement is conducive to fit But we leave it open that frequencies may chance to differ from the uniform chances, since *ceteris* may not be *paribus* and the chances are under pressure not only to fit the frequencies but also to fit into a simple and strong system All this seems very nice

But it doesn't work Along with simpler analyses of chance in terms of actual frequency, it falls victim to the main argument in the last section of the paper Present chances are determined by history up to now, together with history-to-chance conditionals These conditionals are supposed to supervene, via the laws of chance of the best system, on a global pattern of particular fact This global pattern includes future history But there are various different futures which have some present chance of coming about, and which would make the best system different, and thus make the conditionals different, and thus make the present chances different We have the actual present chance distribution over alternative futures, determined by the one future which will actually come about Using it, we have the expected values of the present chances the average of the present chances that would be made true by the various futures, weighted by the chances of those futures But these presently expected values of present chances may differ from the actual present chances A peculiar situation, to say the least

And worse than peculiar Enter the Principal Principle it says first that if we knew the present chances, we should conform our credences about the future to them But it says also that we should conform our credences to the expected values of the present chances [10] If the two

[10] Let A be any proposition, let P_1, P_2, be a partition of propositions to the effect that the present chance of A is x_1, x_2 , respectively, let these propositions have positive present chances of y_1, y_2, , respectively, let C be a reasonable initial credence function, let E be someone's present total evidence, which we may suppose to be presently admissible Suppose that $C(-/E)$ assigns probability 1 to the propositions that the present chance of P_1 is y_1, the present chance of P_2 is y_2 By additivity

(1) $C(A/E) = C(A/P_1E)C(P_1/E) + C(A/P_2E)C(P_2/E) +$

By the Principal Principle,

(2) $C(P_1/E) = y_1$

$C(P_2/E) = y_2$,

and

differ, we cannot do both So if the Principle is right (and if it is possible to conform our credences as we ought to), the two cannot differ So a theory that says they can is wrong

That was the strategy behind my argument in the paper But I streamlined the argument by considering one credence in particular Let T be a full specification of history up to the present and of present chances, and suppose for *reductio* that F is a nonactual future, with some positive present chance of coming about, that would give a different present distribution of chances What is a reasonable credence for F conditionally on T? Zero, because F contradicts T But not zero, by the Principal Principle, because it should equal the positive chance of F according to T This completes the *reductio*

This streamlining might hide the way the argument exploits a predicament that arises already when we consider chance alone Even one who rejects the very idea of credence, and with it the Principal Principle, ought to be suspicious of a theory that permits discrepancies between the chances and their expected values

If anyone wants to defend the best-system theory of laws and chances both (as opposed to the best-system theory of laws, given chances), I suppose the right move would be to cripple the Principal Principle by declaring that information about the chances at a time is *not*, in general, admissible at that time, and hence that hypothetical information about chances, which can join with admissible historical information to imply chances at a time, is likewise inadmissible The reason would be that, under the proposed analysis of chances, information about present chances is a disguised form of inadmissible information about future history—to some extent, it reveals the outcomes of matters that are presently chancy That crippling stops all versions of our *reductio* against positive present chances of futures that would

(3) $C(A/P_1E) = x_1,$

$C(A/P_2E) = x_2$

(Since the $C(P_i/E)$ s are positive, the $C(A/P\,E)$ s are well defined) So we have the prescription

(4) $C(A/E) = y_1x_1 + y_2x_2 +$

that the credence is to be equal to the expected value of chance

yield different present chances [11] I think the cost is excessive, in ordinary calculations with chances, it seems intuitively right to reply on this hypothetical information So, much as I would like to use the best-system approach in defense of Humean supervenience, I cannot support this way out of our difficulty

I stand by my view, in the paper, that if there is any hope for Humean supervenience of chances, it lies in a different direction the history-to-chance conditionals must supervene trivially, by not being contingent at all As noted, that would impose remarkably stringent standards on reasonable belief To illustrate on this hypothesis, enough purely historical information would suffice to tell a reasonable believer whether the half-life of radon is 3 825 days or 3 852 What is more enough purely historical information *about any initial segment of the universe*, however short, would settle the half-life! (It might even be a segment before the time when radon first appeared) For presumably the half-life of radon is settled by the laws of chance, any initial segment of history, aided by enough noncontingent history-to-chance conditionals, suffices to settle any feature of the world that never had a chance to be otherwise, and the laws are such a feature But just how is the believer, however reasonable, supposed to figure out the half-life given his scrap of ancient history? We can hope, I suppose, that some appropriate symmetries in the space of possibilities would do the trick But it seems hard to connect these hoped-for symmetries with anything we now know about the workings of radioactive decay!

D RESTRICTED DOMAINS

In reformulating the Principal Principle, I took care not to presuppose that the domain of a chance distribution would include all propositions Elsewhere I was less cautious I am grateful to Zeno Swijtink for

[11] As to the version in the paper declaring hypothetical information about chances inadmissible blocks my reformulation of the Principal Principle and it was this reformulation that I used in the *reductio*

As to the version in the previous footnote if information about present chances is inadmissible then it becomes very questionable whether the total evidence E can indeed be admissible given that $C(-/E)$ assigns probability 1 to propositions about present chance

As to the streamlined version in this postscript T includes information about pres ent chances, and its partial inadmissibility would block the use of the Principal Principle to prescribe positive credence for F conditionally on T

pointing out (personal communication, 1984) that if I am to be uniformly noncommital on this point, two passages in my final section need correction

I say that if C_1 and C_2 are any two reasonable initial credence functions, and Y is any member of the history-theory partition for any time, then $C_1(-/Y)$ and $C_2(-/Y)$ are "exactly the same" Not so The most I can say is that they agree exactly in the values they assign to the propositions in a certain (presumably large) set, namely, the domain of the chance distribution implied by Y My point stands I have a consequence of the Principal Principle that is entirely about credence, and that limits the ways in which reasonable initial credence functions can differ

Later I say that these differences are—implausibly—even more limited on the hypothesis that the complete theory of chance is the same for all worlds The same correction is required, this time with complete histories in place of history-theory conjunctions Again my point stands The limitation of difference is less than I said, but still implausibly stringent Unless, of course, there are very few propositions which fall in the domains of chance distributions, but that hypothesis also is very implausible, and so would not save the day for a noncontingent theory of chance and for Humean supervenience

My reason for caution was not that I had in mind some interesting class of special propositions—as it might be, about free choices—that would somehow fail to have well-defined chances Rather, I thought it might lead to mathematical difficulties to assume that a probability measure is defined on all propositions without exception In the usual setting for probability theory—values in the standard reals, sigma-additivity—that assumption is indeed unsafe by no means just any measure on a restricted domain of subsets of a given set can be extended to a measure on all the subsets I did not know whether there would be any parallel difficulty in the nonstandard setting, it probably depends on what sort of infinite additivity we wish to assume, just as the difficulty in the standard setting arises only when we require more than finite additivity

Plainly this reason for caution is no reason at all to think that the domains of chance distributions will be notably sparser than the domains of idealized credence functions

TWENTY

Probabilities of Conditionals and Conditional Probabilities

The truthful speaker wants not to assert falsehoods, wherefore he is willing to assert only what he takes to be very probably true He deems it permissible to assert that A only if $P(A)$ is sufficiently close to 1, where P is the probability function that represents his system of degrees of belief at the time Assertability goes by subjective probability

At least, it does in most cases But Ernest Adams has pointed out an apparent exception [1] In the case of ordinary indicative conditionals, it seems that assertability goes instead by the conditional subjective probability of the consequent, given the antecedent We define the conditional probability function $P(-/-)$ by a quotient of absolute probabilities, as usual

(1) $P(C/A) = \text{df } P(CA)/P(A)$, if $P(A)$ is positive

(If the denominator $P(A)$ is zero, we let $P(C/A)$ remain undefined) The truthful speaker evidently deems it permissible to assert the indicative conditional that if A, then C (for short, $A \rightarrow C$) only if $P(C/A)$ is

[1] Ernest Adams, The Logic of Conditionals *Inquiry* 8 (1965) 166–197, and Probability and the Logic of Conditionals *Aspects of Inductive Logic*, ed by Jaakko Hintikka and Patrick Suppes Dordrecht 1966 I shall not here consider Adams s subsequent work which differs at least in emphasis

sufficiently close to 1. Equivalently: only if $P(CA)$ is sufficiently much greater than $P(\bar{C}A)$.

Adams offers two sorts of evidence. There is direct evidence, obtained by contrasting cases in which we would be willing or unwilling to assert various indicative conditionals. There also is indirect evidence, obtained by considering various inferences with indicative conditional premises or conclusions. The ones that seem valid turn out to be just the ones that preserve assertability, if assertability goes by conditional probabilities for conditionals and by absolute probabilities otherwise.[2] Our judgements of validity are not so neatly explained by various rival hypotheses. In particular, they do not fit the hypothesis that the inferences that seem valid are just the ones that preserve truth if we take the conditionals as truth-functional.

Adams has convinced me. I shall take it as established that the assertability of an ordinary indicative conditional $A \rightarrow C$ does indeed go by the conditional subjective probability $P(C/A)$. But why? Why not rather by the absolute probability $P(A \rightarrow C)$?

The most pleasing explanation would be as follows. The assertability of $A \rightarrow C$ does go by $P(A \rightarrow C)$ after all, indicative conditionals are not exceptional. But also it goes by $P(C/A)$, as Adams says, for the meaning of $\rightarrow$ is such as to guarantee that $P(A \rightarrow C)$ and $P(C/A)$ are always equal (if the latter is defined). For short: *probabilities of conditionals are conditional probabilities*. This thesis has been proposed by various authors.[3]

If this is so, then of course the ordinary indicative conditional $A \rightarrow C$ cannot be the truth-functional conditional $A \supset C$. $P(A \supset C)$ and $P(C/A)$ are equal only in certain extreme cases. The indicative conditional must be something else; call it a *probability conditional*. We may or may not be able to give truth conditions for probability conditionals, but at least we may discover a good deal about their meaning and their logic just by using what we know about conditional probabilities.

Alas, this most pleasing explanation cannot be right. We shall see

[2] More precisely, just the ones that satisfy this condition: for any positive ε there is a positive δ such that if any probability function gives each premise an assertability within δ of 1 then it also gives the conclusion an assertability within ε of 1.

[3] Richard Jeffrey, If (abstract), *Journal of Philosophy* **61** (1964), 702–703; Brian Ellis, An Epistemological Concept of Truth, *Contemporary Philosophy in Australia*, ed. by Robert Brown and C. D. Rollins, London, 1969; Robert Stalnaker, Probability and Conditionals, *Philosophy of Science* **37** (1970), 64–80. We shall consider later whether to count Adams as another adherent of the thesis.

that there is no way to interpret a conditional connective so that, with sufficient generality, the probabilities of conditionals will equal the appropriate conditional probabilities If there were, probabilities of conditionals could serve as links to establish relationships between the probabilities of non-conditionals, but the relationships thus established turn out to be incorrect The quest for a probability conditional is futile, and we must admit that assertability does not go by absolute probability in the case of indicative conditionals

PRELIMINARIES

Suppose we are given an interpreted formal language equipped at least with the usual truth-functional connectives and with the further connective $\rightarrow$ These connectives may be used to compound any sentences in the language We think of the interpretation as giving the truth value of every sentence at every possible world Two sentences are *equivalent* iff they are true at exactly the same worlds, and *incompatible* iff there is no world where both are true One sentence *implies* another iff the second is true at every world where the first is true A sentence is *necessary*, *possible* or *impossible* iff it is true at all worlds, at some, or at none We may think of a probability function P as an assignment of numerical values to all sentences of this language, obeying these standard laws of probability

(2) $1 \geqslant P(A) \geqslant 0$,
(3) if A and B are equivalent, then $P(A) = P(B)$,
(4) if A and B are incompatible, then $P(A \vee B) = P(A) + P(B)$,
(5) if A is necessary, then $P(A) = 1$

The definition (1) gives us the multiplication law for conjunctions

Whenever $P(B)$ is positive, there is a probability function P' such that $P'(A)$ always equals $P(A/B)$, we say that P' *comes from* P *by conditionalizing on* B A class of probability functions is *closed under conditionalizing* iff any probability function that comes by conditionalizing from one in the class is itself in the class

Suppose that $\rightarrow$ is interpreted in such a way that, for some particular probability function P, and for any sentences A and C,

(6) $P(A \rightarrow C) = P(C/A)$, if $P(A)$ is positive,

iff so, let us call $\rightarrow$ a *probability conditional for* P Iff $\rightarrow$ is a probability conditional for every probability function in some class of probability

functions, then let us call $\rightarrow$ a *probability conditional for* the class And iff $\rightarrow$ is a probability conditional for all probability functions, so that (6) holds for any P, A, and C, then let us call $\rightarrow$ a *universal probability conditional*, or simply a *probability conditional*

Observe that if $\rightarrow$ is a universal probability conditional, so that (6) holds always, then (7) also holds always

(7) $P(A \rightarrow C/B) = P(C/AB)$, if $P(AB)$ is positive

To derive (7), apply (6) to the probability function P' that comes from P by conditionalizing on B, such a P' exists if $P(AB)$ and hence also $P(B)$ are positive Then (7) follows by several applications of (1) and the equality between $P'(\text{—})$ and $P(\text{—}/B)$ In the same way, if $\rightarrow$ is a probability conditional for a class of probability functions, and if that class is closed under conditionalizing, then (7) holds for any probability function P in the class, and for any A and C (It does not follow, however, that if (6) holds for a particular probability function P, then (7) holds for the same P)

FIRST TRIVIALITY RESULT

Suppose by way of *reductio* that $\rightarrow$ is a universal probability conditional Take any probability function P and any sentences A and C such that $P(AC)$ and $P(A\bar{C})$ both are positive Then $P(A)$, $P(C)$, and $P(\bar{C})$ also are positive By (6) we have

(8) $P(A \rightarrow C) = P(C/A)$

By (7), taking B as C or as $\bar{C}$ and simplifying the right-hand side, we have

(9) $P(A \rightarrow C/C) = P(C/AC) = 1$,
(10) $P(A \rightarrow C/\bar{C}) = P(C/A\bar{C}) = 0$

For any sentence D, we have the familiar expansion by cases

(11) $P(D) = P(D/C)\ \ P(C) + P(D/\bar{C})\ \ P(\bar{C})$

In particular, take D as $A \rightarrow C$ Then we may substitute (8), (9), and (10) into (11) to obtain

(12) $P(C/A) = 1\ \ P(C) + 0\ \ P(\bar{C}) = P(C)$

With the aid of the supposed probability conditional, we have reached the conclusion that if only $P(AC)$ and $P(A\bar{C})$ both are positive, then A

and C are probabilistically independent under P That is absurd For instance, let P be the subjective probability function of someone about to throw what he takes to be a fair die, let A mean that an even number comes up, and let C mean that the six comes up $P(AC)$ and $P(A\bar{C})$ are positive But, *contra* (12), $P(C/A)$ is $\frac{1}{3}$ and $P(C)$ is $\frac{1}{6}$, A and C are not independent More generally, let C, D, and E be possible but pairwise incompatible There are probability functions that assign positive probability to all three let P be any such Let A be the disjunction $C \vee D$ Then $P(AC)$ and $P(A\bar{C})$ are positive but $P(C/A)$ and $P(C)$ are unequal

Our supposition that $\rightarrow$ is a universal probability conditional has led to absurdity, but not quite to contradiction If the given language were sufficiently weak in expressive power, then our conclusion might be unobjectionable There might not exist any three possible but pairwise incompatible sentences to provide a counterexample to it For all I have said, such a weak language might be equipped with a universal probability conditional Indeed, consider the extreme case of a language in which there are none but necessary sentences and impossible ones For this very trivial language, the truth-functional conditional itself is a universal probability conditional

If an interpreted language cannot provide three possible but pairwise incompatible sentences, then we may justly call it a *trivial language* We have proved this theorem *any language having a universal probability conditional is a trivial language*

SECOND TRIVIALITY RESULT

Since our language is not a trivial one, our indicative conditional must not be a universal probability conditional But all is not yet lost for the thesis that probabilities of conditionals are conditional probabilities A much less than universal probability conditional might be good enough Our task, after all, concerns subjective probability probability functions used to represent people's systems of beliefs We need not assume, and indeed it seems rather implausible, that any probability function whatever represents a system of beliefs that it is possible for someone to have We might set aside those probability functions that do not If our indicative conditional were a probability conditional for a limited class of probability functions, and if that class were inclusive enough to contain any probability function that might ever represent a speaker's system of beliefs, that would suffice to

explain why assertability of indicative conditionals goes by conditional subjective probability

Once we give up on universality, it may be encouraging to find that probability conditionals for particular probability functions, at least, commonly do exist Given a probability function P, we may be able to tailor the interpretation of $\rightarrow$ to fit [4] Suppose that for any A and C there is some B such that $P(B/\bar{A})$ and $P(C/A)$ are equal if both defined, this should be a safe assumption when P is a probability function rich enough to represent someone's system of beliefs If for any A and C we arbitrarily choose such a B and let $A \rightarrow C$ be interpreted as equivalent to $AC \vee \bar{A}B$, then $\rightarrow$ is a probability conditional for P But such piecemeal tailoring does not yet provide all that we want Even if there is a probability conditional for each probability function in a class, it does not follow that there is one probability conditional for the entire class Different members of the class might require different interpretations of $\rightarrow$ to make the probabilities of conditionals and the conditional probabilities come out equal But presumably our indicative conditional has a fixed interpretation, the same for speakers with different beliefs, and for one speaker before and after a change in his beliefs Else how are disagreements about a conditional possible, or changes of mind? Our question, therefore, is whether the indicative conditional might have one fixed interpretation that makes it a probability conditional for the entire class of all those probability functions that represent possible systems of beliefs

This class, we may reasonably assume, is closed under conditionalizing Rational change of belief never can take anyone to a subjective probability function outside the class, and there are good reasons why the change of belief that results from coming to know an item of new evidence should take place by conditionalizing on what was learned [5]

Suppose by way of *reductio* that $\rightarrow$ is a probability conditional for a class of probability functions, and that the class is closed under conditionalizing The argument proceeds much as before Take any probability function P in the class and any sentences A and C such that

[4] I am indebted to Bas van Fraassen for this observation He has also shown that by judicious selection of the B s we can give $\rightarrow$ some further properties that might seem appropriate to a conditional connective See Bas van Fraassen, Probabilities of Conditionals , in *Foundations of Probability Theory Statistical Inference and Statistical Theories of Science*, Volume I, ed by W Harper and C A Hooker, D Reidel Dordrecht Holland, 1976, p 261

[5] These reasons may be found in Paul Teller, Conditionalization and Observation , *Synthese* 26 (1973) 218–258

$P(AC)$ and $P(A\bar{C})$ are positive Again we have (6) and hence (8), (7) and hence (9) and (10), (11) and hence by substitution (12) $P(C/A)$ and $P(C)$ must be equal But if we take three pairwise incompatible sentences C, D, and E such that $P(C)$, $P(D)$, and $P(E)$ are all positive and if we take A as the disjunction $C \vee D$, then $P(AC)$ and $P(A\bar{C})$ are positive but $P(C/A)$ and $P(C)$ are unequal So there are no such three sentences Further, P has at most four different values Else there would be two different values of P, x and y, strictly intermediate between 0 and 1 and such that $x + y \neq 1$ But then if $P(F) = x$ and $P(G) = y$ it follows that at least three of $P(FG)$, $P(\bar{F}G)$, $P(F\bar{G})$, and $P(\bar{F}\bar{G})$ are positive, which we have seen to be impossible

If a probability function never assigns positive probability to more than two incompatible alternatives, and hence is at most four-valued, then we may call it a *trivial probability function* We have proved this theorem *if a class of probability functions is closed under conditionalizing, then there can be no probability conditional for that class unless the class consists entirely of trivial probability functions* Since some probability functions that represent possible systems of belief are not trivial, our indicative conditional is not a probability conditional for the class of all such probability functions Whatever it may mean, it cannot possibly have a meaning such as to guarantee, for all possible subjective probability functions at once, that the probabilities of conditionals equal the corresponding conditional probabilities There is no such meaning to be had We shall have to grant that the assertability of indicative conditionals does not go by absolute probability, and seek elsewhere for an explanation of the fact that it goes by conditional probability instead

THE INDICATIVE CONDITIONAL AS NON-TRUTH-VALUED

Assertability goes in general by probability because probability is probability of truth and the speaker wants to be truthful If this is not so for indicative conditionals, perhaps the reason is that they have no truth values, no truth conditions, and no probabilities of truth Perhaps they are governed not by a semantic rule of truth but by a rule of assertability

We might reasonably take it as the goal of semantics to specify our prevailing rules of assertability Most of the time, to be sure, that can best be done by giving truth conditions plus the general rule that

speakers should try to be truthful, or in other words that assertability goes by probability of truth But sometimes the job might better be done another way for instance, by giving truth conditions for antecedents and for consequents, but not for whole conditionals, plus the special rule that the assertability of an indicative conditional goes by the conditional subjective probability of the consequent given the antecedent Why not? We are surely free to institute a new sentence form, without truth conditions, to be used for making it known that certain of one's conditional subjective probabilities are close to 1 But then it should be no surprise if we turn out to have such a device already

Adams himself seems to favor this hypothesis about the semantics of indicative conditionals [6] He advises us, at any rate, to set aside questions about their truth and to concentrate instead on their assertability There is one complication Adams does *say* that conditional probabilities are probabilities of conditionals Nevertheless he does not mean by this that the indicative conditional is what I have here called a probability conditional, for he does not claim that the so-called "probabilities" of conditionals are probabilities of truth, and neither does he claim that they obey the standard laws of probability They are probabilities only in name Adams's position is therefore invulnerable to my triviality results, which were proved by applying standard laws of probability to the probabilities of conditionals

Would it make sense to suppose that indicative conditionals *do not* have truth values, truth conditions, or probabilities of truth, but that they *do* have probabilities that obey the standard laws? Yes, but only if we first restate those laws to get rid of all mention of truth We must continue to permit unrestricted compounding of sentences by means of the usual connectives, so that the domain of our probability functions will be a Boolean algebra (as is standardly required), but we can no longer assume that these connectives always have their usual truth-functional interpretations, since truth-functional compounding of non-truth-valued sentences makes no sense Instead we must choose some deductive system—any standard formalization of sentential logic will do—and characterize the usual connectives by their deductive role in this system We must replace mention of equivalence, incompatibility, and necessity in laws (3) through (5) by mention of their syntactic substitutes in the chosen system inter-deducibility, deductive inconsistency, and deducibility In this way we could describe the

[6] The Logic of Conditionals

probability functions for our language without assuming that all probabilities of sentences, or even any of them, are probabilities of truth We could still hold that assertability goes in most cases by probability, though we could no longer restate this as a rule that speakers should try to tell the truth

Merely to deny that probabilities of conditionals are probabilities of truth, while retaining all the standard laws of probability in suitably adapted form, would not yet make it safe to revive the thesis that probabilities of conditionals are conditional probabilities It was not the connection between truth and probability that led to my triviality results, but only the application of standard probability theory to the probabilities of conditionals The proofs could just as well have used versions of the laws that mentioned deducibility instead of truth Whoever still wants to say that probabilities of conditionals are conditional probabilities had better also employ a non-standard calculus of "probabilities" He might drop the requirement that the domain of a probability function is a Boolean algebra, in order to exclude conjunctions with conditional conjuncts from the language Or he might instead limit (4), the law of additivity, refusing to apply it when the disjuncts A and B contain conditional conjuncts Either maneuver would block my proofs But if it be granted that the "probabilities" of conditionals do not obey the standard laws, I do not see what is to be gained by insisting on calling them "probabilities" It seems to me that a position like Adams's might best be expressed by saying that indicative conditionals have neither truth values nor probabilities, and by introducing some neutral term such as "assertability" or "value" which denotes the probability of truth in the case of nonconditionals and the appropriate conditional probability in the case of indicative conditionals

I have no conclusive objection to the hypothesis that indicative conditionals are non-truth-valued sentences, governed by a special rule of assertability that does not involve their nonexistent probabilities of truth I have an inconclusive objection, however the hypothesis requires too much of a fresh start It burdens us with too much work still to be done, and wastes too much that has been done already So far, we have nothing but a rule of assertability for conditionals with truth-valued antecedents and consequents But what about compound sentences that have such conditionals as constituents? We think we know how the truth conditions for compound sentences of various kinds are determined by the truth conditions of constituent subsentences, but this knowledge would be useless if any of those subsen-

tences lacked truth conditions Either we need new semantic rules for many familiar connectives and operators when applied to indicative conditionals—perhaps rules of truth, perhaps special rules of assertability like the rule for conditionals themselves—or else we need to explain away all seeming examples of compound sentences with conditional constituents

THE INDICATIVE CONDITIONAL AS TRUTH-FUNCTIONAL

Fortunately a more conservative hypothesis is at hand H P Grice has given an elegant explanation of some qualitative rules governing the assertability of indicative conditionals [7] It turns out that a quantitative hypothesis based on Grice's ideas gives us just what we want the rule that assertability goes by conditional subjective probability

According to Grice, indicative conditionals *do* have truth values, truth conditions, and probabilities of truth In fact, the indicative conditional $A \rightarrow C$ is simply the truth-functional conditional $A \supset C$ But the assertability of this truth-functional conditional does not go just by $P(A \supset C)$, its subjective probability of truth It goes by the resultant of that and something else

It may happen that a speaker believes a truth-functional conditional to be true, yet he ought not to assert it Its assertability might be diminished for various reasons, but let us consider one in particular The speaker ought not to assert the conditional if he believes it to be true predominantly because he believes its antecedent to be false, so that its probability of truth consists mostly of its probability of vacuous truth In this situation, why assert the conditional instead of denying the antecedent? It is pointless to do so And if it is pointless, then also it is worse than pointless it is misleading The hearer, trusting the speaker not to assert pointlessly, will assume that he has not done so The hearer may then wrongly infer that the speaker has additional reason to believe that the conditional is true, over and above his disbelief in the antecedent

This consideration detracts from the assertability of $A \supset C$ to the extent that both of two conditions hold first, that the probability $P(\bar{A})$

[7] H P Grice, Logic and Conversation The William James Lectures, given at Harvard University in 1967

of vacuity is high, and second, that the probability $P(\overline{C}A)$ of falsity is a large fraction of the total probability $P(A)$ of nonvacuity The product

(13) $P(\overline{A})\ \ (P(\overline{C}A)/P(A))$

of the degrees to which the two conditions are met is therefore a suitable measure of diminution of assertability Taking the probability $P(A \supset C)$ of truth, and subtracting the diminution of assertability as measured by (13), we obtain a suitable measure of resultant assertability

(14) $P(A \supset C) - P(\overline{A})\ \ (P(\overline{C}A)/P(A))$

But (14) may be simplified, using standard probability theory, and so we find that the resultant assertability, probability of truth minus the diminution given by (13), is equal to the conditional probability $P(C/A)$ That is why assertability goes by conditional probability

Diminished assertability for such reasons is by no means special to conditionals It appears also with uncontroversially truth-functional constructions such as negated conjunction We are gathering mushrooms, I say to you "You won't eat that one and live " A dirty trick I thought that one was safe and especially delicious, I wanted it myself, so I hoped to dissuade you from taking it without actually lying I thought it highly probable that my trick would work, that you would not eat the mushroom, and therefore that I would turn out to have told the truth But though what I said had a high subjective probability of truth, it had a low assertability and it was a misdeed to assert it Its assertability goes not just by probability but by the resultant of that and a correction term to take account of the pointlessness and misleadingness of denying a conjunction when one believes it false predominantly because of disbelieving one conjunct Surely few would care to explain the low assertability of what I said by rejecting the usual truth-functional semantics for negation and conjunction, and positing instead a special probabilistic rule of assertability

There are many considerations that might detract from assertability Why stop at (14)? Why not add more terms to take account of the diminished assertability of insults, of irrelevancies, of long-winded pomposities, of breaches of confidence, and so forth? Perhaps part of the reason is that, unlike the diminution of assertability when the probability of a conditional is predominantly due to the improbability of the antecedent, these other diminutions depend heavily on miscellaneous features of the conversational context In logic we are accustomed to consider sentences and inferences in abstraction from

context Therefore it is understandable if, when we philosophize, our judgements of assertability or of assertability-preserving inference are governed by a measure of assertability such as (14), that is $P(C/A)$, in which the more context-dependent dimensions of assertability are left out

There is a more serious problem, however What of conditionals that have a high probability predominantly because of the probability of the consequent? If we are on the right track, it seems that there should be a diminution of assertability in this case also, and one that should still show up if we abstract from context we could argue that in such a case it is pointless, and hence also misleading, to assert the conditional rather than the consequent This supposed diminution is left out, and I think rightly so, if we measure the assertability of a conditional $A \supset C$ (in abstraction from context) by $P(C/A)$ If A and C are probabilistically independent and each has probability 9, then the probability of the conditional (91) is predominantly due to the probability of the consequent (9), yet the conditional probability $P(C/A)$ is high (9) so we count the conditional as assertable And it does seem so, at least in some cases "I'll probably flunk, and it doesn't matter whether I study, I'll flunk if I do and I'll flunk if I don't "

The best I can do to account for the absence of a marked diminution in the case of the probable consequent is to concede that considerations of conversational pointlessness are not decisive They create only tendencies toward diminished assertability, tendencies that may or may not be conventionally reinforced In the case of the improbable antecedent, they are strongly reinforced In the case of the probable consequent, apparently they are not

In conceding this, I reduce the distance between my present hypothesis that indicative conditionals are truth-functional and the rival hypothesis that they are non-truth-valued and governed by a special rule of assertability Truth conditions plus general conversational considerations are not quite the whole story They go much of the way toward determining the assertability of conditionals, but a separate convention is needed to finish the job The point of ascribing truth conditions to indicative conditionals is not that we can thereby get rid entirely of special rules of assertability

Rather, the point of ascribing truth conditions is that we thereby gain at least a prima facie theory of the truth conditions and assertability of compound sentences with conditional constituents We need not waste whatever general knowledge we have about the way the truth conditions of compounds depend on the truth conditions of their con-

stituents Admittedly we might go wrong by proceeding in this way We have found one explicable discrepancy between assertability and probability in the case of conditionals themselves, and there might be more such discrepancies in the case of various compounds of conditionals (For instance, the assertability of a negated conditional seems not to go by its probability of truth, but rather to vary inversely with the assertability of the conditional) It is beyond the scope of this paper to survey the evidence, but I think it reasonable to hope that the discrepancies are not so many, or so difficult to explain, that they destroy the explanatory power of the hypothesis that the indicative conditional is truth-functional

PROBABILITIES OF STALNAKER CONDITIONALS

It is in some of the writings of Robert Stalnaker that we find the fullest elaboration of the thesis that conditional probabilities are probabilities of conditionals [8] Stalnaker's conditional connective $>$ has truth conditions roughly as follows a conditional $A > C$ is true iff the least drastic revision of the facts that would make A true would make C true as well Stalnaker conjectures that this interpretation will make $P(A > C)$ and $P(C/A)$ equal whenever $P(A)$ is positive He also lays down certain constraints on $P(A > C)$ for the case that $P(A)$ is zero, explaining this by means of an extended concept of conditional probability that need not concern us here

Stalnaker supports his conjecture by exhibiting a coincidence between two sorts of validity The sentences that are true no matter what, under Stalnaker's truth conditions, turn out to be exactly those that have positive probability no matter what, under his hypothesis about probabilities of conditionals Certainly this is weighty evidence, but it is not decisive Cases are known in modal logic, for instance, in which very different interpretations of a language happen to validate the very same sentences And indeed our triviality results show that

[8] Probabilities and Conditionals The Stalnaker conditional had been introduced in Robert Stalnaker, A Theory of Conditionals , *Studies in Logical Theory*, ed by Nicholas Rescher, Oxford, 1968 I have discussed the Stalnaker conditional in *Counterfactuals*, Oxford, 1973, pp 77–83, arguing there that an interpretation quite similar to Stalnaker s is right for counterfactuals but wrong for indicative conditionals

Stalnaker's conjecture cannot be right, unless we confine our attention to trivial probability functions [9]

But it is almost right, as we shall see Probabilities of Stalnaker conditionals do not, in general, equal the corresponding conditional probabilities [10] But they do have some of the characteristic properties of conditional probabilities

A possible totality of facts corresponds to a possible world, so a revision of facts corresponds to a transition from one world to another For any given world W and (possible) antecedent A, let W_A be the world we reach by the least drastic revision of the facts of W that makes A true There is to be no gratuitous revision W_A may differ from W as much as it must to permit A to hold, but no more Balancing off respects of similarity and difference against each other according to the importance we attach to them, W_A is to be the closest in overall similarity to W among the worlds where A is true Then the Stalnaker conditional $A > C$ is true at the world W iff C is true at W_A, the closest A-world to W (In case the antecedent A is impossible, so that there is no possible A-world to serve as W_A, we take $A > C$ to be vacuously true at all worlds For simplicity I speak here only of absolute impossibility, Stalnaker works with impossibility relative to worlds) Let us introduce this notation

$$(15) \qquad W(A) = \text{df} \begin{cases} 1 \text{ if } A \text{ is true at the world } W \\ 0 \text{ if } A \text{ is false at } W \end{cases}$$

Then we may give the truth conditions for nonvacuous Stalnaker conditionals as follows

$$(16) \quad W(A > C) = W_A(C), \text{ if } A \text{ is possible}$$

[9] Once it is recognized that the Stalnaker conditional is not a probability conditional, the coincidence of logics has a new significance The hypothesis that assertability of indicative conditionals goes by conditional probabilities, though still sufficiently well supported by direct evidence, is no longer unrivalled as an explanation of our judgements of validity for inferences with indicative conditional premises or conclusions The same judgements could be explained instead by the hypothesis that the indicative conditional is the Stalnaker conditional and we judge valid those inferences that preserve truth

[10] Although the probabilities of Stalnaker conditionals and the corresponding conditional probabilities cannot always be equal, they often are They are equal whenever the conditional (and perhaps some non conditional state of affairs on which it depends) is probabilistically independent of the antecedent For example, my present subjective probabilities are such that the conditional probability of finding a penny in my pocket, given that I look for one equals the probability of the conditional I look for a penny > I find one The reason is that both are equal to the absolute probability that there is a penny in my pocket now

It will be convenient to pretend, from this point on, that there are only finitely many possible worlds That will trivialize the mathematics but not distort our conclusions Then we can think of a probability function P as a distribution of probability over the worlds Each world W has a probability $P(W)$, and these probabilities of worlds sum to 1 We return from probabilities of worlds to probabilities of sentences by summing the probabilities of the worlds where a sentence is true

(17) $P(A) = \Sigma_W P(W) \quad W(A)$

I shall also assume that the worlds are distinguishable for any two, some sentence of our language is true at one but not the other Thus we disregard phenomena that might result if our language were sufficiently lacking in expressive power

Given any probability function P and any possible A, there is a probability function P' such that, for any world W',

(18) $P'(W') = \Sigma_W P(W) \quad \begin{cases} 1 \text{ if } W_A \text{ is } W' \\ 0 \text{ otherwise} \end{cases}$

Let us say that P' *comes from* P *by imaging on* A, and call P' the *image of* P *on* A Intuitively, the image on A of a probability function is formed by shifting the original probability of each world W over to W_A, the closest A-world to W Probability is moved around but not created or destroyed, so the probabilities of worlds still sum to 1 Each A-world keeps whatever probability it had originally, since if W is an A-world then W_A is W itself, and it may also gain additional shares of probability that have been shifted away from $\bar{A}$-worlds The $\bar{A}$-worlds retain none of their original probability, and gain none All the probability has been concentrated on the A-worlds And this has been accomplished with no gratuitous movement of probability Every share stays as close as it can to the world where it was originally located

Suppose that P' comes from P by imaging on A, and consider any sentence C

(19) $P'(C) = \Sigma_W P'(W') \quad W'(C)$, by (17) applied to P',

$= \Sigma_W \left(\Sigma_W P(W) \quad \begin{cases} 1 \text{ if } W_A \text{ is } W' \\ 0 \text{ otherwise} \end{cases} \right) \quad W'(C)$, by (18),

$= \Sigma_W P(W) \quad \left(\Sigma_W \begin{cases} 1 \text{ if } W_A \text{ is } W' \\ 0 \text{ otherwise} \end{cases} W'(C) \right)$, by algebra,

$= \Sigma_W P(W) \quad W_A(C)$, simplifying the inner sum,

$= \Sigma_W P(W) \quad W(A > C)$, by (16),

$= P(A > C)$, by (17)

We have proved this theorem *the probability of a Stalnaker conditional with a possible antecedent is the probability of the consequent after imaging on the antecedent*

Conditionalizing is one way of revising a given probability function so as to confer certainty—probability of 1—on a given sentence Imaging is another way to do the same thing The two methods do not in general agree (Example let $P(W)$, $P(W')$, and $P(W'')$ each equal $\frac{1}{3}$, let A hold at W and W' but not W'', and let W' be the closest A-world to W'' Then the probability function that comes from P by conditionalizing on A assigns probability $\frac{1}{2}$ to both W and W', whereas the probability function that comes from P by imaging on A assigns probability $\frac{1}{3}$ to W and $\frac{2}{3}$ to W') But though the methods differ, either one can plausibly be held to give minimal revisions to revise the given probability function as much as must be done to make the given sentence certain, but no more Imaging P on A gives a minimal revision in this sense unlike all other revisions of P to make A certain, it involves no gratuitous movement of probability from worlds to dissimilar worlds Conditionalizing P on A gives a minimal revision in this different sense unlike all other revisions of P to make A certain, it does not distort the profile of probability ratios, equalities, and inequalities among sentences that imply A [11]

Stalnaker's conjecture divides into two parts This part is true the probability of a nonvacuous Stalnaker conditional is the probability of the consequent, after minimal revision of the original probability function to make the antecedent certain But it is not true that this minimal revision works by conditionalizing Rather it must work by imaging Only when the two methods give the same result does the probability of a Stalnaker conditional equal the corresponding conditional probability

Stalnaker gives the following instructions for deciding whether or not you believe a conditional [12]

> First, add the antecedent (hypothetically) to your stock of beliefs, second, make whatever adjustments are required to maintain consistency (without modifying the hypothetical belief in the antecedent), finally, consider whether or not the consequent is true

That is right, for a Stalnaker conditional, if the feigned revision of beliefs works by imaging However the passage suggests that the thing

[11] Teller, Conditionalization and Observation

[12] A Theory of Conditionals , p 102

to do is to feign the sort of revision that would take place if the antecedent really were added to your stock of beliefs That is wrong If the antecedent really were added, you should (if possible) revise by conditionalizing The reasons in favor of responding to new evidence by conditionalizing are equally reasons against responding by imaging instead

PROBABILITY-REVISION CONDITIONALS

Suppose that the connective $\rightarrow$ is interpreted in such a way that for any probability function P, and for any sentences A and C,

(20) $P(A \rightarrow C) = P_A(C)$, if A is possible,

where P_A is (in some sense) the minimal revision of P that raises the probability of A to 1 Iff so, let us call $\rightarrow$ a *probability-revision conditional* Is there such a thing? We have seen that it depends on the method of revision Conditionalizing yields revisions that are minimal in one sense, and if P_A is obtained (when possible) by conditionalizing, then no probability-revision conditional exists (unless the language is trivial) Imaging yields revisions that are minimal in another sense, and if P_A is obtained by imaging then the Stalnaker conditional is a probability-revision conditional Doubtless there are still other methods of revision, yielding revisions that are minimal in still other senses than we have yet considered Are there any other methods which, like imaging and unlike conditionalizing, can give us a probability-revision conditional? There are not, as we shall see The only way to have a probability-revision conditional is to interpret the conditional in Stalnaker's way and revise by imaging

Since we have not fixed on a particular method of revising probability functions, our definition of a probability-revision conditional should be understood as tacitly relative to a method To make this relativity explicit, let us call $\rightarrow$ a *probability-revision conditional for* a given method iff (20) holds in general when P_A is taken to be the revision obtained by that method

Our definition of a Stalnaker conditional should likewise be understood as tacitly relative to a method of revising worlds Stalnaker's truth conditions were deliberately left vague at the point where they mention the minimal revision of a given world to make a given antecedent true With worlds, as with probability functions, different

methods of revision will yield revisions that are minimal in different senses We can indeed describe any method as selecting the antecedent-world closest in overall similarity to the original world, but different methods will fit this description under different resolutions of the vagueness of similarity, resolutions that stress different respects of comparison To be explicit, let us call $\rightarrow$ a *Stalnaker conditional for* a given method of revising worlds iff (16) holds in general when W_A is taken to be the revision obtained by that method (and $A \rightarrow C$ is true at all worlds if A is impossible) I spoke loosely of "the" Stalnaker conditional, but henceforth it will be better to speak in the plural of the Stalnaker conditionals for various methods of revising worlds

We are interested only in those methods of revision, for worlds and for probability functions, that can be regarded as giving revisions that are in some reasonable sense minimal We have no hope of saying in any precise way just which methods those are, but at least we can list some formal requirements that such a method must satisfy The requirements were given by Stalnaker for revision of worlds, but they carry over *mutatis mutandis* to revision of probability functions also First, a minimal revision to reach some goal must be one that does reach it For worlds, W_A must be a world where A is true, for probability functions, P_A must assign to A a probability of 1 Second, there must be no revision when none is needed For worlds, if A is already true at W then W_A must be W itself, for probability functions, if $P(A)$ is already 1, then P_A must be P Third, the method must be consistent in its comparisons For worlds, if B is true at W_A and A is true at W_B then W_A and W_B must be the same, else W_A would be treated as both less and more of a revision of W than is W_B Likewise for probability functions, if $P_A(B)$ and $P_B(A)$ both are 1, then P_A and P_B must be the same

Let us call any method of revision of worlds or of probability functions *eligible* iff it satisfies these three requirements We note that the methods of revising probability functions that we have considered are indeed eligible Conditionalizing is an eligible method, or, more precisely, conditionalizing can be extended to an eligible method applicable to any probability function P and any possible A (Choose some fixed arbitrary well-ordering of all probability functions In case P_A cannot be obtained by conditionalizing because $P(A)$ is zero, let it be the first, according to the arbitrary ordering, of the probability functions that assign to A a probability of 1) Imaging is also an eligible method More precisely, imaging on the basis of any eligible method of revising worlds is an eligible method of revising probability functions

Our theorem of the previous section may be restated as follows *If*

→ is a Stalnaker conditional for any eligible method of revising worlds, then → is also a probability-revision conditional for an eligible method of revising probability functions, namely, for the method that works by imaging on the basis of the given method of revising worlds Now we shall prove the converse *if → is a probability-revision conditional for an eligible method of revising probability functions, then → is also a Stalnaker conditional for an eligible method of revising worlds* In short, the *probability-revision conditionals are exactly the Stalnaker conditionals*

Suppose that we have some eligible method of revising probability functions, and suppose that → is a probability-revision conditional for this method

We shall need to find a method of revising worlds, therefore let us consider the revision of certain special probability functions that stand in one-to-one correspondence with the worlds For each world W, there is a probability function P that gives all the probability to W and none to any other world Accordingly, by (17),

$$(21) \qquad P(A) = \begin{Bmatrix} 1 \text{ if } A \text{ is true at } W \\ 0 \text{ if } A \text{ is false at } W \end{Bmatrix} = W(A)$$

for any sentence A Call such a probability function *opinionated*, since it would represent the beliefs of someone who was absolutely certain that the world W was actual and who therefore held a firm opinion about every question, and call the world W where P concentrates all the probability the *belief world of P*

Our given method of revising probability functions preserves opinionation Suppose P were opinionated and P_A were not, for some possible A That is to say that P_A gives positive probability to two or more worlds We have assumed that our language has the means to distinguish the worlds, so there is some sentence C such that $P_A(C)$ is neither 0 nor 1 But since P is opinionated, $P(A \rightarrow C)$ is either 0 or 1, contradicting the hypothesis that → is a probability-revision conditional so that $P_A(C)$ and $P(A \rightarrow C)$ are equal

Then we have the following method of revising worlds Given a world W and possible sentence A, let P be the opinionated probability function with belief world W, revise P according to our given method of revising probability functions, and let W_A be the belief world of the resulting opinionated probability function P_A Since the given method of revising probability functions is eligible, so is this derived method of revising worlds

Consider any world W and sentences A and C Let P be the opinion-

ated probability function with belief world W, and let W_A be as above Then if A is possible,

$$(22) \qquad W(A \rightarrow C) = P(A \rightarrow C), \text{ by (21)},$$
$$= P_A(C), \text{ by (20)},$$
$$= W_A(C), \text{ by (21) applied to } W_A$$

So $\rightarrow$ is a Stalnaker conditional for the derived method of revising worlds *Quod erat demonstrandum* [13]

Postscript to

"Probabilities of Conditionals and Conditional Probabilities"

INDICATIVE CONDITIONALS BETTER EXPLAINED

I retract the positive theory of indicative conditionals that I proposed in the paper I now prefer the alternative theory advanced by Frank Jackson [1]

The two theories have much in common Both agree (1) that the indicative conditional has the truth conditions of the truth-functional conditional $A \supset C$, yet (2) its assertability goes by the conditional subjective probability $P(C/A)$, provided that we abstract from special considerations—of etiquette, say—that apply in special cases Both theories further agree, therefore, (3) that there is a discrepancy between truth- and assertability-preserving inference involving indicative conditionals, and (4) that our intuitions about valid reasoning with con-

[13] An earlier version of this paper was presented at a Canadian Philosophical Association colloquium on probability semantics for conditional logic at Montreal in June 1972 I am grateful to many friends and colleagues, and especially to Ernest Adams and Robert Stalnaker, for valuable comments

[1] On Assertion and Indicative Conditionals, *Philosophical Review* 88 (1979) 565–89 Conditionals and Possibilia, *Proceedings of the Aristotelian Society* 81 (1981) 125–37

ditionals are apt to concern the latter, and so to be poor evidence about the former (As to whether "validity" should be the word for truth- or for assertability-preservation, that seems a non-issue if ever there was one) Further, the theories agree (5) that the discrepancy between the assertability $P(C/A)$ and the probability of truth $P(A \supset C)$ is due to some sort of Gricean implicature, and (6) that an adequate account of this implicature must use the premise that the conditional has the truth conditions of $A \supset C$ I still hold these six theses

But what sort of implicature is involved? Formerly, I thought it was predominantly a conversational implicature, akin to the implicature from "Here, you have a good point" to "Elsewhere, you mostly don't " According to Jackson, it is a conventional implicature, akin to the implicature from "She votes Liberal but she's no fool" to "Liberal voters mostly are fools "

I said, following Grice if $P(A \supset C)$ is high mostly because $P(A)$ is low, what's the sense of saying $A \supset C$? Why not say the stronger thing that's almost as probable, not-A? If you say the weaker thing, you will be needlessly uninformative Besides, you will mislead those who rely on you not to be needlessly uninformative, and who will infer that you were not in a position to say the stronger thing

To which Jackson replies that we often do say weaker things than we believe true, and for a very good reason I speak to you (or to my future self, via memory) in the expectation that our belief systems will be much alike, but not exactly alike If there were too little in common, my attempts to convey information would fail, if there were too much in common, they would serve no purpose I do not know quite what other information you (or I in future) may possess from other sources Maybe you (or I in future) know something that now seems to me improbable I would like to say something that will be useful even so So let me not say the strongest thing I believe Let me say something a bit weaker, if I can thereby say something that will not need to be given up, that will remain useful, even if a certain hypothesis that I now take to be improbable should turn out to be the case If I say something that I would continue to believe even if I should learn that the improbable hypothesis is true, then that will be something that I think you can take my word for even if you already believe the hypothesis

Let us say that A is *robust* with respect to B (according to someone's subjective probabilities at a certain time) iff the unconditional probability of A and the probability of A conditionally on B are close together, and both are high, so that even if one were to learn that B,

one would continue to find A probable Then Jackson's point is that one might say the weaker thing rather than the stronger for the sake of robustness The weaker might be more robust with respect to some case that one judges to be improbable, but that one nevertheless does not wish to ignore

If it is pointless to say the weaker instead of the stronger, how much more pointless to say the weaker and the stronger both! And yet we do I might say "Bruce is asleep in the rag box, or anyway somewhere downstairs " Jackson can explain that There's point in saying the stronger, and there's point in saying the more robust, and they're different, so I say them both

It could be useful to point out that one is saying something robust One might say "I am saying A not because I do not believe anything stronger, but because I want to say something which is robust with respect to B—something you may rely on even if, unlike me, you believe that B " But that's clumsy It would be a good idea if we had conventional devices to signal robustness more concisely So it would be no surprise to find out that we do Jackson suggests that we have various such devices, and that the indicative conditional construction is one of them

An indicative conditional is a truth-functional conditional that conventionally implicates robustness with respect to the antecedent Therefore, an indicative conditional with antecedent A and consequent C is assertable iff (or to the extent that) the probabilities $P(A \supset C)$ and $P(A \supset C/A)$ both are high If the second is high, the first will be too, and the second is high iff $P(C/A)$ is high, and that is the reason why the assertability of indicative conditionals goes by the corresponding conditional probability

Jackson lists several advantages of his implicature-of-robustness theory over my assert-the-stronger theory I will mention only one (which is not to suggest that I find the rest unpersuasive) I can say "Fred will not study, and if he does he still won't pass " If the conditional is assertable only when the denial of its antecedent is not, as the assert-the-stronger theory predicts, then how can it happen that the conditional and the denial of its antecedent *both* are assertable? As already noted, Jackson can explain such things The conditional was added for the sake of robustness, so that even if you happen to think I'm wrong about Fred not studying, you can still take my word for it that if he studies he still won't pass

So far, I have just been retailing Jackson But I think that one complication ought to be added (Jackson tells me that he agrees) Above, I

introduced robustness by what was in effect a double definition first in terms of probability, then in terms of what would happen if something were learned Let us distinguish more carefully

> *A* is *robust*$_1$ with respect to *B* iff *P*(*A*) and *P*(*A*/*B*) are close, and both are high
>
> *A* is *robust*$_2$ with respect to *B* iff *P*(*A*) is high, and would remain high even if one were to learn that *B*

Robustness$_1$ is robustness as Jackson officially defines it, and it is the implicature of robustness$_1$ that explains why assertability of conditionals goes by conditional probability But our reasons for wanting to say what's robust, and for needing signals of robustness, seem to apply to robustness$_2$ Most of the time, fortunately, the distinction doesn't matter Suppose that if one were to learn that *B*, one would learn only that *B*, and nothing else (or nothing else relevant) And suppose that one would then revise one's beliefs by conditionalizing Then we have robustness$_2$ of *A* with respect to *B* iff we have robustness$_1$ In this normal case, the distinction makes no difference

However, there may be abnormal cases cases in which *B* could not be learned all by itself, but would have to be accompanied by some extra information *E* Suppose *A* is robust$_1$ with respect to *B* alone, but not with respect to *B* and *E* in conjunction Then *A* will not be robust$_2$ with respect to *B* Example *A* is "I'll never believe that Reagan works for the KGB", *B* is "Reagan works for the KGB", and *E* is not-*A* My thought is that if the KGB were successful enough to install their man as president, surely they'd also be successful enough to control the news completely So *P*(*A*) and *P*(*A*/*B*) are both high, but of course *P*(*A*/*BE*) = 0 Yet if I did learn that Reagan worked for the KGB, I'd *ipso facto* learn that I believed it—despite my prior expectation that the KGB would be able to keep me from suspecting So *A* is not at all robust$_2$ with respect to *B* [2]

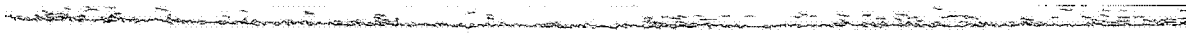
When the two senses of robustness come apart in special cases, which one does the indicative conditional signal? What really matters is robustness$_2$, so it would be more useful to signal that On the other

[2] A closely related point appears in Bas van Fraassen s review of Brian Ellis *Rational Belief Systems*, *Canadian Journal of Philosophy* 10 (1980) 497–511, with an illustrative example due to Richmond Thomason a man accepts If my wife were deceiving me, I would believe that she was not (because she is so clever), but that doesn t mean that if he were to come to believe the antecedent he would then believe the consequent

hand, it would be much easier to signal robustness$_1$. Robustness$_2$ with respect to B amounts roughly to robustness$_1$ with respect to the whole of what would be learned if B were learned. (The two are equivalent under the assumption that the learner would conditionalize.) But it might be no easy thing to judge what would be learned if B were learned, in view of the variety of ways that something might be learned. For the most part, robustness$_1$ is a reasonable guide to the robustness$_2$ that really matters—a fallible guide, as we've seen, but pretty good most of the time. So it's unsurprising if what we have the means to signal is the former rather than the latter. And if this gets conventionalized, it should be unsurprising to find that we signal robustness$_1$ even when that clearly diverges from robustness$_2$. That is exactly what happens. Example: I can perfectly well say "If Reagan works for the KGB, I'll never believe it."

PART SIX

Causation



TWENTY-ONE

Causation*

Hume defined causation twice over He wrote "we may define a cause to be *an object followed by another, and where all the objects, similar to the first, are followed by objects similar to the second* Or, in other words, *where, if the first object had not been, the second never had existed* "[1]

Descendants of Hume's first definition still dominate the philosophy of causation a casual succession is supposed to be a succession that instantiates a regularity To be sure, there have been improvements Nowadays we try to distinguish the regularities that count—the "causal laws"—from mere accidental regularities of succession We subsume causes and effects under regularities by means of descriptions they satisfy, not by over-all similarity And we allow a cause to be only one indispensable part, not the whole, of the total situation that is followed by the effect in accordance with a law In present-day regularity analyses, a cause is defined (roughly) as any member of any minimal set of actual conditions that are jointly sufficient, given the laws, for the existence of the effect

More precisely, let C be the proposition that *c* exists (or occurs) and

* I thank the American Council of Learned Societies, Princeton University, and the National Science Foundation for research support

[1] *An Enquiry concerning Human Understanding*, Section VII

let E be the proposition that e exists Then c causes e, according to a typical regularity analysis,[2] iff (1) C and E are true, and (2) for some nonempty set $\mathfrak{L}$ of true law-propositions and some set $\mathfrak{F}$ of true propositions of particular fact, $\mathfrak{L}$ and $\mathfrak{F}$ jointly imply $C \supset E$, although $\mathfrak{L}$ and $\mathfrak{F}$ jointly do not imply E and $\mathfrak{F}$ alone does not imply $C \supset E$ [3]

Much needs doing, and much has been done, to turn definitions like this one into defensible analyses Many problems have been overcome Others remain in particular, regularity analyses tend to confuse causation itself with various other causal relations If c belongs to a minimal set of conditions jointly sufficient for e, given the laws, then c may well be a genuine cause of e But c might rather be an effect of e one which could not, given the laws and some of the actual circumstances, have occurred otherwise than by being caused by e Or c might be an epiphenomenon of the causal history of e a more or less inefficacious effect of some genuine cause of e Or c might be a preempted potential cause of e something that did not cause e, but that would have done so in the absence of whatever really did cause e

It remains to be seen whether any regularity analysis can succeed in distinguishing genuine causes from effects, epiphenomena, and preempted potential causes—and whether it can succeed without falling victim to worse problems, without piling on the epicycles, and without departing from the fundamental idea that causation is instantiation of regularities I have no proof that regularity analyses are beyond repair, nor any space to review the repairs that have been tried Suffice it to say that the prospects look dark I think it is time to give up and try something else

A promising alternative is not far to seek Hume's "other words"—that if the cause had not been, the effect never had existed—are no mere restatement of his first definition They propose something altogether different a counterfactual analysis of causation

The proposal has not been well received True, we do know that causation has something or other to do with counterfactuals We think

[2] Not one that has been proposed by any actual author in just this form, so far as I know

[3] I identify a *proposition*, as is becoming usual, with the set of possible worlds where it is true It is not a linguistic entity Truth functional operations on propositions are the appropriate Boolean operations on sets of worlds, logical relations among propositions are relations of inclusion, overlap, etc among sets A sentence of a language *expresses* a proposition iff the sentence and the proposition are true at exactly the same worlds No ordinary language will provide sentences to express all propositions, there will not be enough sentences to go around

of a cause as something that makes a difference, and the difference it makes must be a difference from what would have happened without it Had it been absent, its effects—some of them, at least, and usually all—would have been absent as well Yet it is one thing to mention these platitudes now and again, and another thing to rest an analysis on them That has not seemed worth while [4] We have learned all too well that counterfactuals are ill understood, wherefore it did not seem that much understanding could be gained by using them to analyze causation or anything else Pending a better understanding of counterfactuals, moreover, we had no way to fight seeming counterexamples to a counterfactual analysis

But counterfactuals need not remain ill understood, I claim, unless we cling to false preconceptions about what it would be like to understand them Must an adequate understanding make no reference to unactualized possibilities? Must it assign sharply determinate truth conditions? Must it connect counterfactuals rigidly to covering laws? Then none will be forthcoming So much the worse for those standards of adequacy Why not take counterfactuals at face value as statements about possible alternatives to the actual situation, somewhat vaguely specified, in which the actual laws may or may not remain intact? There are now several such treatments of counterfactuals, differing only in details [5] If they are right, then sound foundations have been laid for analyses that use counterfactuals

In this paper, I shall state a counterfactual analysis, not very different from Hume's second definition, of some sorts of causation Then I shall try to show how this analysis works to distinguish genuine causes from effects, epiphenomena, and preempted potential causes

My discussion will be incomplete in at least four ways Explicit preliminary settings-aside may prevent confusion

1 I shall confine myself to causation among *events*, in the everyday sense of the word flashes, battles, conversations, impacts, strolls, deaths, touchdowns, falls, kisses, and the like Not that events are the only things that can cause or be caused, but I have no full list of the others, and no good umbrella-term to cover them all

2 My analysis is meant to apply to causation in particular cases It

[4] One exception Ardon Lyon, Causality, *British Journal for the Philosophy of Science*, XVIII, 1 (May 1967) 1–20

[5] See, for instance, Robert Stalnaker A Theory of Conditionals, in Nicholas Rescher, ed , *Studies in Logical Theory* (Oxford Blackwell, 1968), and my *Counterfactuals* (Oxford Blackwell, 1973)

is not an analysis of causal generalizations Presumably those are quantified statements involving causation among particular events (or non-events), but it turns out not to be easy to match up the causal generalizations of natural language with the available quantified forms A sentence of the form "C-events cause E-events," for instance, can mean any of

(a) For some *c* in C and some *e* in E, *c* causes *e*
(b) For every *e* in E, there is some *c* in C such that *c* causes *e*
(c) For every *c* in C, there is some *e* in E such that *c* causes *e*

not to mention further ambiguities Worse still, "Only C-events cause E-events" ought to mean

(d) For every *c*, if there is some *e* in E such that *c* causes *e*, then *c* is in C

if "only" has its usual meaning But no, it unambiguously means (b) instead! These problems are not about causation, but about our idioms of quantification

3 We sometimes single out one among all the causes of some event and call it "the" cause, as if there were no others Or we single out a few as the "causes," calling the rest mere "causal factors" or "causal conditions " Or we speak of the "decisive" or "real" or "principal" cause We may select the abnormal or extraordinary causes, or those under human control, or those we deem good or bad, or just those we want to talk about I have nothing to say about these principles of invidious discrimination [6] I am concerned with the prior question of what it is to be one of the causes (unselectively speaking) My analysis is meant to capture a broad and nondiscriminatory concept of causation

4 I shall be content, for now, if I can give an analysis of causation that works properly under determinism By determinism I do not mean any thesis of universal causation, or universal predictability-in-principle, but rather this the prevailing laws of nature are such that there do not exist any two possible worlds which are exactly alike up to some time, which differ thereafter, and in which those laws are never violated Perhaps by ignoring indeterminism I squander the most striking advantage of a counterfactual analysis over a regularity analy-

[6] Except that Morton G White s discussion of causal selection, in *Foundations of Historical Knowledge* (New York Harper & Row 1965), pp 105–181, would meet my needs despite the fact that it is based on a regularity analysis

sis that it allows undetermined events to be caused [7] I fear, however, that my present analysis cannot yet cope with all varieties of causation under indeterminism The needed repair would take us too far into disputed questions about the foundations of probability

COMPARATIVE SIMILARITY

To begin, I take as primitive a relation of *comparative over-all similarity* among possible worlds We may say that one world is *closer to actuality* than another if the first resembles our actual world more than the second does, taking account of all the respects of similarity and difference and balancing them off one against another

(More generally, an arbitrary world w can play the role of our actual world In speaking of our actual world without knowing just which world is ours, I am in effect generalizing over all worlds We really need a three-place relation world w_1 is closer to world w than world w_2 is I shall henceforth leave this generality tacit)

I have not said just how to balance the respects of comparison against each other, so I have not said just what our relation of comparative similarity is to be Not for nothing did I call it primitive But I have said what *sort* of relation it is, and we are familiar with relations of that sort We do make judgments of comparative overall similarity—of people, for instance—by balancing off many respects of similarity and difference Often our mutual expectations about the weighting factors are definite and accurate enough to permit communication I shall have more to say later about the way the balance must go in particular cases to make my analysis work But the vagueness of over-all similarity will not be entirely resolved Nor should it be The vagueness of similarity does infect causation, and no correct analysis can deny it

The respects of similarity and difference that enter into the over-all similarity of worlds are many and varied In particular, similarities in matters of particular fact trade off against similarities of law The prevailing laws of nature are important to the character of a world, so similarities of law are weighty Weighty, but not sacred We should not take it for granted that a world that conforms perfectly to our actual

[7] That this ought to be allowed is argued in G E M Anscombe *Causality and Determination An Inaugural Lecture* (Cambridge University Press, 1971), and in Fred Dretske and Aaron Snyder Causal Irregularity, *Philosophy of Science*, XXXIX, 1 (March 1972) 69–71

laws is *ipso facto* closer to actuality than any world where those laws are violated in any way at all It depends on the nature and extent of the violation, on the place of the violated laws in the total system of laws of nature, and on the countervailing similarities and differences in other respects Likewise, similarities or differences of particular fact may be more or less weighty, depending on their nature and extent Comprehensive and exact similarities of particular fact throughout large spatiotemporal regions seem to have special weight It may be worth a small miracle to prolong or expand a region of perfect match

Our relation of comparative similarity should meet two formal constraints (1) It should be a weak ordering of the worlds an ordering in which ties are permitted, but any two worlds are comparable (2) Our actual world should be closest to actuality, resembling itself more than any other world resembles it We do *not* impose the further constraint that for any set A of worlds there is a unique closest A-world, or even a set of A-worlds tied for closest Why not an infinite sequence of closer and closer A-worlds, but no closest?

COUNTERFACTUALS AND COUNTERFACTUAL DEPENDENCE

Given any two propositions A and C, we have their *counterfactual* $A \mathbin{\Box\!\!\rightarrow} C$ the proposition that if A were true, then C would also be true The operation $\mathbin{\Box\!\!\rightarrow}$ is defined by a rule of truth, as follows $A \mathbin{\Box\!\!\rightarrow} C$ is true (at a world w) iff either (1) there are no possible A-worlds (in which case $A \mathbin{\Box\!\!\rightarrow} C$ is *vacuous*), or (2) some A-world where C holds is closer (to w) than is any A-world where C does not hold In other words, a counterfactual is nonvacuously true iff it takes less of a departure from actuality to make the consequent true along with the antecedent than it does to make the antecedent true without the consequent

We did not assume that there must always be one or more closest A-worlds But if there are, we can simplify $A \mathbin{\Box\!\!\rightarrow} C$ is nonvacuously true iff C holds at all the closest A-worlds

We have not presupposed that A is false If A is true, then our actual world is the closest A-world, so $A \mathbin{\Box\!\!\rightarrow} C$ is true iff C is Hence $A \mathbin{\Box\!\!\rightarrow} C$ implies the material conditional $A \supset C$, and A and C jointly imply $A \mathbin{\Box\!\!\rightarrow} C$

Let A_1, A_2, be a family of possible propositions, no two of which are compossible, let C_1, C_2, be another such family (of

equal size) Then if all the counterfactuals $A_1 \boxrightarrow C_1$, $A_2 \boxrightarrow C_2$, between corresponding propositions in the two families are true, we shall say that the C's *depend counterfactually* on the A's We can say it like this in ordinary language whether C_1 or C_2 or depends (counterfactually) on whether A_1 or A_2 or

Counterfactual dependence between large families of alternatives is characteristic of processes of measurement, perception, or control Let R_1, R_2, be propositions specifying the alternative readings of a certain barometer at a certain time Let P_1, P_2, specify the corresponding pressures of the surrounding air Then, if the barometer is working properly to measure the pressure, the R's must depend counterfactually on the P's As we say it the reading depends on the pressure Likewise, if I am seeing at a certain time, then my visual impressions must depend counterfactually, over a wide range of alternative possibilities, on the scene before my eyes And if I am in control over what happens in some respect, then there must be a double counterfactual dependence, again over some fairly wide range of alternatives The outcome depends on what I do, and that in turn depends on which outcome I want [8]

CAUSAL DEPENDENCE AMONG EVENTS

If a family C_1, C_2, depends counterfactually on a family A_1, A_2, in the sense just explained, we will ordinarily be willing to speak also of causal dependence We say, for instance, that the barometer reading depends causally on the pressure, that my visual impressions depend causally on the scene before my eyes, or that the outcome of something under my control depends causally on what I do But there are exceptions Let G_1, G_2, be alternative possible laws of gravitation, differing in the value of some numerical constant Let M_1, M_2, be suitable alternative laws of planetary motion Then the M's may depend counterfactually on the G's, but we would not call this dependence causal Such exceptions as this, however, do not involve any sort of dependence among distinct particular events The hope remains that causal dependence among events, at least, may be analyzed simply as counterfactual dependence

[8] Analyses in terms of counterfactual dependence are found in two papers of Alvin I Goldman Toward a Theory of Social Power, *Philosophical Studies*, XXIII (1972) 221–268, and Discrimination and Perceptual Knowledge presented at the 1972 Chapel Hill Colloquium

We have spoken thus far of counterfactual dependence among propositions, not among events Whatever particular events may be, presumably they are not propositions But that is no problem, since they can at least be paired with propositions To any possible event e, there corresponds the proposition $O(e)$ that holds at all and only those worlds where e occurs This $O(e)$ is the proposition that e occurs [9] (If no two events occur at exactly the same worlds—if, that is, there are no absolutely necessary connections between distinct events—we may add that this correspondence of events and propositions is one to one) Counterfactual dependence among events is simply counterfactual dependence among the corresponding propositions

Let c_1, c_2, and e_1, e_2, be distinct possible events such that no two of the c's and no two of the e's are compossible Then I say that the family e_1, e_2, of events *depends causally* on the family c_1, c_2, iff the family $O(e_1)$, $O(e_2)$, of propositions depends counterfactually on the family $O(c_1)$, $O(c_2)$, As we say it whether e_1 or e_2 or occurs depends on whether c_1 or c_2 or occurs

We can also define a relation of dependence among single events rather than families Let c and e be two distinct possible particular events Then e *depends causally* on c iff the family $O(e)$, $\sim O(e)$ depends counterfactually on the family $O(c)$, $\sim O(c)$ As we say it whether e occurs or not depends on whether c occurs or not The dependence consists in the truth of two counterfactuals

[9] Beware if we refer to a particular event e by means of some description that e satisfies, then we must take care not to confuse $O(e)$, the proposition that e itself occurs with the different proposition that some event or other occurs which satisfies the description It is a contingent matter in general, what events satisfy what descriptions Let e be the death of Socrates—the death he actually died, to be distinguished from all the different deaths he might have died instead Suppose that Socrates had fled, only to be eaten by a lion Then e would not have occurred, and $O(e)$ would have been false, but a different event would have satisfied the description the death of Socrates that I used to refer to e Or suppose that Socrates had lived and died just as he actually did and afterwards was resurrected and killed again and resurrected again, and finally became immortal Then no event would have satisfied the description (Even if the temporary deaths are real deaths, neither of the two can be *the* death) But e would have occurred and $O(e)$ would have been true Call a description of an event e *rigid* iff (1) nothing but e could possibly satisfy it, and (2) e could not possibly occur without satisfying it I have claimed that even such commonplace descriptions as the death of Socrates are nonrigid, and in fact I think that rigid descriptions of events are hard to find That would be a problem for anyone who needed to associate with every possible event e a sentence $\Phi(e)$ true at all and only those worlds where e occurs But we need no such sentences—only propositions, which may or may not have expressions in our language

$O(c) \mathbin{\Box\!\!\rightarrow} O(e)$ and $\sim O(c) \mathbin{\Box\!\!\rightarrow} \sim O(e)$ There are two cases If c and e do not actually occur, then the second counterfactual is automatically true because its antecedent and consequent are true so e depends causally on c iff the first counterfactual holds That is, iff e would have occurred if c had occurred But if c and e are actual events, then it is the first counterfactual that is automatically true Then e depends causally on c iff, if c had not been, e never had existed I take Hume's second definition as my definition not of causation itself, but of causal dependence among actual events

CAUSATION

Causal dependence among actual events implies causation If c and e are two actual events such that e would not have occurred without c, then c is a cause of e But I reject the converse Causation must always be transitive, causal dependence may not be, so there can be causation without causal dependence Let c, d, and e be three actual events such that d would not have occurred without c and e would not have occurred without d Then c is a cause of e even if e would still have occurred (otherwise caused) without c

We extend causal dependence to a transitive relation in the usual way Let c, d, e, be a finite sequence of actual particular events such that d depends causally on c, e on d, and so on throughout Then this sequence is a *causal chain* Finally, one event is a *cause* of another iff there exists a causal chain leading from the first to the second This completes my counterfactual analysis of causation

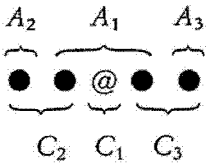
COUNTERFACTUAL VERSUS NOMIC DEPENDENCE

It is essential to distinguish counterfactual and causal dependence from what I shall call *nomic dependence* The family C_1, C_2, of propositions depends nomically on the family A_1, A_2, iff there are a nonempty set $\mathfrak{L}$ of true law-propositions and a set $\mathfrak{F}$ of true propositions of particular fact such that $\mathfrak{L}$ and $\mathfrak{F}$ jointly imply (but $\mathfrak{F}$ alone does not imply) all the material conditionals $A_1 \supset C_1$, $A_2 \supset C_2$, between the corresponding propositions in the two families (Recall that these same material conditionals are implied by the counterfactuals that would comprise a counterfactual dependence) We shall say

also that the nomic dependence holds *in virtue of* the premise sets $\mathfrak{L}$ and $\mathfrak{F}$

Nomic and counterfactual dependence are related as follows Say that a proposition B is *counterfactually independent* of the family A_1, A_2, of alternatives iff B would hold no matter which of the A's were true—that is, iff the counterfactuals $A_1 \boxrightarrow B$, $A_2 \boxrightarrow B$, all hold If the C's depend nomically on the A's in virtue of the premise sets $\mathfrak{L}$ and $\mathfrak{F}$, and if in addition (all members of) $\mathfrak{L}$ and $\mathfrak{F}$ are counterfactually independent of the A's, then it follows that the C's depend counterfactually on the A's In that case, we may regard the nomic dependence in virtue of $\mathfrak{L}$ and $\mathfrak{F}$ as explaining the counterfactual dependence Often, perhaps always, counterfactual dependences may be thus explained But the requirement of counterfactual independence is indispensable Unless $\mathfrak{L}$ and $\mathfrak{F}$ meet that requirement, nomic dependence in virtue of $\mathfrak{L}$ and $\mathfrak{F}$ does not imply counterfactual dependence, and, if there is counterfactual dependence anyway, does not explain it

Nomic dependence is reversible, in the following sense If the family C_1, C_2, depends nomically on the family A_1, A_2, in virtue of $\mathfrak{L}$ and $\mathfrak{F}$, then also A_1, A_2, depends nomically on the family AC_1, AC_2, , in virtue of $\mathfrak{L}$ and $\mathfrak{F}$, where A is the disjunction $A_1 \vee A_2 \vee$ Is counterfactual dependence likewise reversible? That does not follow For, even if $\mathfrak{L}$ and $\mathfrak{F}$ are independent of A_1, A_2, and hence establish the counterfactual dependence of the C's on the A's, still they may fail to be independent of AC_1, AC_2, , and hence may fail to establish the reverse counterfactual dependence of the A's on the AC's Irreversible counterfactual dependence is shown below @ is our actual world, the dots are the other worlds, and distance on the page represents similarity "distance"



The counterfactuals $A_1 \boxrightarrow C_1$, $A_2 \boxrightarrow C_2$, and $A_3 \boxrightarrow C_3$ hold at the actual world, wherefore the C's depend on the A's But we do not have the reverse dependence of the A's on the AC's, since instead of the needed $AC_2 \boxrightarrow A_2$ and $AC_3 \boxrightarrow A_3$ we have $AC_2 \boxrightarrow A_1$ and $AC_3 \boxrightarrow A_1$

Just such irreversibility is commonplace The barometer reading

depends counterfactually on the pressure—that is as clear-cut as counterfactuals ever get—but does the pressure depend counterfactually on the reading? If the reading had been higher, would the pressure have been higher? Or would the barometer have been malfunctioning? The second sounds better a higher reading would have been an incorrect reading To be sure, there are actual laws and circumstances that imply and explain the actual accuracy of the barometer, but these are no more sacred than the actual laws and circumstances that imply and explain the actual pressure Less sacred, in fact When something must give way to permit a higher reading, we find it less of a departure from actuality to hold the pressure fixed and sacrifice the accuracy, rather than vice versa It is not hard to see why The barometer, being more localized and more delicate than the weather, is more vulnerable to slight departures from actuality [10]

We can now explain why regularity analyses of causation (among events, under determinism) work as well as they do Suppose that event c causes event e according to the sample regularity analysis that I gave at the beginning of this paper, in virtue of premise sets $\mathfrak{L}$ and $\mathfrak{F}$ It follows that $\mathfrak{L}$, $\mathfrak{F}$ and $\sim O(c)$ jointly do not imply $O(e)$ Strengthen this suppose further that they do imply $\sim O(e)$ If so, the family $O(e)$, $\sim O(e)$ depends nomically on the family $O(c)$, $\sim O(c)$ in virtue of $\mathfrak{L}$ and $\mathfrak{F}$ Add one more supposition that $\mathfrak{L}$ and $\mathfrak{F}$ are counterfactually independent of $O(c)$, $\sim O(c)$ Then it follows according to my counterfactual analysis that e depends counterfactually and causally on c, and hence that c causes e If I am right, the regularity analysis gives conditions that are almost but not quite sufficient for explicable causal dependence That is not quite the same thing as causation, but causation without causal dependence is scarce, and if there is inexplicable causal dependence we are (understandably!) unaware of it [11]

[10] Granted, there are contexts or changes of wording that would incline us the other way For some reason, If the reading had been higher, that would have been because the pressure was higher invites my assent more than If the reading had been higher, the pressure would have been higher The counterfactuals from readings to pressures are much less clear-cut than those from pressures to readings But it is enough that some legitimate resolutions of vagueness give an irreversible dependence of readings on pressures Those are the resolutions we want at present, even if they are not favored in all contexts

[11] I am not here proposing a repaired regularity analysis The repaired analysis would gratuitously rule out inexplicable causal dependence, which seems bad Nor would it be squarely in the tradition of regularity analyses any more Too much else would have been added

EFFECTS AND EPIPHENOMENA

I return now to the problems I raised against regularity analyses, hoping to show that my counterfactual analysis can overcome them

The *problem of effects*, as it confronts a counterfactual analysis, is as follows Suppose that *c* causes a subsequent event *e*, and that *e* does not also cause *c* (I do not rule out closed causal loops a priori, but this case is not to be one) Suppose further that, given the laws and some of the actual circumstances, *c* could not have failed to cause *e* It seems to follow that if the effect *e* had not occurred, then its cause *c* would not have occurred We have a spurious reverse causal dependence of *c* on *e*, contradicting our supposition that *e* did not cause *c*

The *problem of epiphenomena*, for a counterfactual analysis, is similar Suppose that *e* is an epiphenomenal effect of a genuine cause *c* of an effect *f* That is, *c* causes first *e* and then *f*, but *e* does not cause *f* Suppose further that, given the laws and some of the actual circumstances, *c* could not have failed to cause *e*, and that, given the laws and others of the circumstances, *f* could not have been caused otherwise than by *c* It seems to follow that if the epiphenomenon *e* had not occurred, then its cause *c* would not have occurred and the further effect *f* of that same cause would not have occurred either We have a spurious causal dependence of *f* on *e*, contradicting our supposition that *e* did not cause *f*

One might be tempted to solve the problem of effects by brute force insert into the analysis a stipulation that a cause must always precede its effect (and perhaps a parallel stipulation for causal dependence) I reject this solution (1) It is worthless against the closely related problem of epiphenomena, since the epiphenomenon *e* does precede its spurious effect *f* (2) It rejects a priori certain legitimate physical hypotheses that posit backward or simultaneous causation (3) It trivializes any theory that seeks to define the forward direction of time as the predominant direction of causation

The proper solution to both problems, I think, is flatly to deny the counterfactuals that cause the trouble If *e* had been absent, it is not that *c* would have been absent (and with it *f*, in the second case) Rather, *c* would have occurred just as it did but would have failed to cause *e* It is less of a departure from actuality to get rid of *e* by holding *c* fixed and giving up some or other of the laws and circumstances in virtue of which *c* could not have failed to cause *e*, rather than to hold those laws and circumstances fixed and get rid of *e* by going back and abolishing its cause *c* (In the second case, it would of course be point-

less not to hold *f* fixed along with *c*.) The causal dependence of *e* on *c* is the same sort of irreversible counterfactual dependence that we have considered already.

To get rid of an actual event *e* with the least over-all departure from actuality, it will normally be best not to diverge at all from the actual course of events until just before the time of *e*. The longer we wait, the more we prolong the spatiotemporal region of perfect match between our actual world and the selected alternative. Why diverge sooner rather than later? Not to avoid violations of laws of nature. Under determinism *any* divergence, soon or late, requires some violation of the actual laws. If the laws were held sacred, there would be no way to get rid of *e* without changing all of the past, and nothing guarantees that the change could be kept negligible except in the recent past. That would mean that if the present were ever so slightly different, then all of the past would have been different—which is absurd. So the laws are not sacred. Violation of laws is a matter of degree. Until we get up to the time immediately before *e* is to occur, there is no general reason why a later divergence to avert *e* should need a more severe violation than an earlier one. Perhaps there are special reasons in special cases—but then these may be cases of backward causal dependence.

PREEMPTION

Suppose that c_1 occurs and causes *e*, and that c_2 also occurs and does not cause *e*, but would have caused *e* if c_1 had been absent. Thus c_2 is a potential alternate cause of *e*, but is preempted by the actual cause c_1. We may say that c_1 and c_2 overdetermine *e*, but they do so asymmetrically.[12] In virtue of what difference does c_1 but not c_2 cause *e*?

As far as causal dependence goes, there is no difference. *e* depends neither on c_1 nor on c_2. If either one had not occurred, the other would have sufficed to cause *e*. So the difference must be that, thanks to c_1, there is no causal chain from c_2 to *e*, whereas there is a causal chain of two or more steps from c_1 to *e*. Assume for simplicity that two steps are enough. Then *e* depends causally on some intermediate event *d*,

[12] I shall not discuss symmetrical cases of overdetermination, in which two overdetermining factors have equal claim to count as causes. For me these are useless as test cases because I lack firm naive opinions about them.

and d in turn depends on c_1 Causal dependence is here intransitive c_1 causes e via d even though e would still have occurred without c_1

So far, so good It remains only to deal with the objection that e does *not* depend causally on d, because if d had been absent then c_1 would have been absent and c_2, no longer preempted, would have caused e We may reply by denying the claim that if d had been absent then c_1 would have been absent That is the very same sort of spurious reverse dependence of cause on effect that we have just rejected in simpler cases I rather claim that if d had been absent, c_1 would somehow have failed to cause d But c_1 would still have been there to interfere with c_2, so e would not have occurred

Postcripts to

"Causation"

A PIECEMEAL CAUSATION

Suppose that c and e are large, prolonged processes, each composed of many smaller events Suppose it is not true (or not clearly true) that e, taken as a whole, causally depends on c, taken as a whole, suppose even that they are not connected by a chain of causal dependence It may nevertheless be that c and e are divisible into parts in such a way that every part of e is causally dependent on (or connected by a chain of causal dependence to) some part of c In that case we might well simply speak of c as a cause of e, though it is not so under the analysis I gave

Self-sustaining processes exhibit piecemeal causation For instance, suppose a public address system is turned up until it howls from feedback The howling, from start to finish, is an event If it had not occurred, it would not have occurred, but this is certainly not counterfactual dependence between *distinct* events, therefore it does not qualify as causal dependence on my account Nor is there a closed causal loop, as in time travel stories, in which the howling causes itself because it depends causally on some distinct event which in turn depends causally on it So it is not true, on my account, that the howling *taken as a whole* causes itself What is true is that the howling

causes itself piecemeal It is divisible into parts in such a way that each part except the first is caused by an earlier part, and each part except the last causes a later part This causing of part by part is unproblematic cause and effect are distinct events, wherefore their counterfactual dependence qualifies as causal We might well say that the howling causes itself, this is to be accepted, but only in a derivative sense Similarly, if two prolonged events sustain one another, each causes the other piecemeal The example of the howling illustrates this case also the sound in the air sustains the signal in the wires, and *vice versa*

It may be that when we speak of causation in history we are often speaking of piecemeal causation [1] A depression causes a wave of bankruptcies what are we to make of this? If the depression had not occurred That is puzzling To suppose away an entire depression takes us a long way from actuality And the farther we depart from actuality, the more we lose control over our counterfactuals For the more different respects of similarity and difference we have to balance, the more of a problem it is that we have left it vague just how to do the balancing, so the less clearly we know what is and what isn't to be held fixed in our counterfactualizing (For instance, what if many of the firms that went broke came into existence during, and because of, the depression itself? Shall we hold their existence fixed in asking what would have happened without the depression?) But the depression is a big event that is divisible into many parts Although it is hard to say what would happened without the entire depression, it is comparatively easy to say that without this or that event which was part of the depression, this or that one of the bankruptcies would not have taken place Now, our counterfactuals are much more under control, because they stay much closer to home So even if it is unclear what the depression taken as a whole might have caused, it is at any rate clear that various parts of it caused the various bankruptcies That is to say that the depression was at least a piecemeal cause of the wave of bankruptcies

There is a well-known dilemma about actions Consider an action of raising my arm First something goes on within my brain, then signals go out my nerves, then my muscles contract, and as they do, my arm rises There seems to be a conflict between two things we want to say (1) The action of raising my arm is a prolonged event with diverse parts It is the whole causal process just described It may begin within

[1] Here I am indebted to a lecture given by Martin Putnam at Princeton in 1976

me, but it is not over until my arm rises Its earlier parts cause its later parts, and its final part is the bodily movement But (2) just as my action of raising a flag would be an event that causes a flag to rise, so my action of raising my arm is an event that causes my arm to rise The raising, whether of flag or of arm, is so-called because it causes the rising

Distinguish the inclusion of one event in another from mere involvement of one in the naming of the other So far as involvement in naming goes, the two cases are on a par For the flag and the arm alike, the raising deserves its name only if, and perhaps only after, a rising ensues But with respect to inclusion, the two cases seem to differ If I raise a flag by delayed action, I can be done raising it long before it rises My action is over when I have done my part, the process that ends when the flag is up consists of more than just my action (Beware ambiguity the phrase "my raising the flag" might denote just my action, or it might denote the whole affair) But if I raise my arm by delayed action—say that I have *very* sluggish nerves—then it takes me a long time to raise my arm In this case, my part of the process is the whole of the process So long as the signal is traveling through my sluggish nerves, so long as my muscles are contracting and my arm is rising, my part of the affair is still going on

I would like to assent to both (1) and (2), the apparent obstacle is that we have two events, the raising and the rising, and according to (1) they are not wholly distinct, yet according to (2) one causes the other But if this is a case of piecemeal causation, we have no problem If an early part of the raising causes the rising which is a late part of the raising, we may still say simply that the raising causes the rising, just as, when an early part of the depression causes a bankruptcy which is a later part of the depression, we may still say simply that the depression causes the bankruptcy

There is a second version of the problem The rising of my arm is not the only event which is caused by the initial inner part of my action and yet takes place before my arm has risen The same may be true of various side effects, events which definitely are not to be included as parts of the action Suppose, for instance, that the nerves leading into my arm are monitored so that whenever I raise my arm the nerve signal produces a trace on an oscillograph Because I can produce the trace by raising my arm, we ought to be free to say that my action causes the trace And yet the trace appears before the arm rises Shall we say that the effect precedes its cause? Or that the action which causes the effect is over sooner than we think? Neither it is a case of piecemeal

causation Like the rising of the arm, the trace on the oscillograph is caused by an initial part of the action, and thereby is caused by the action [2]

B CHANCY CAUSATION

In the paper, I confined my discussion to the deterministic case for the sake of brevity [3] But I certainly do not think that causation requires determinism (Hence I regard "causality" as a naughty word, since it is ambiguous between "causation" and "determinism") Events that happen by chance may nevertheless be caused Indeed, it seems likely that most actual causation is of just this sort Whether that is so or not, plenty of people do think that our world is chancy, and chancy enough so that most things that happen had some chance, immediately beforehand, of not happening These people are seldom observed to deny commonplace causal statements, except perhaps when they philosophize An analysis that imputes widespread error is *prima facie* implausible Moreover, it is dishonest to accept it, if you yourself persist in the "error" when you leave the philosophy room We had better provide for causation under indeterminism, causation of events for which prior conditions were not lawfully sufficient

One kind of chancy causation is already covered by my analysis, with no modification needed *c* occurs, *e* has some chance of occurring, as it happens *e* does occur, but if *c* had not occurred, then *e* would have had no chance at all of occurring, and so would not have occurred Then *e* depends causally on *c*, and *c* is a cause of *e*, according to my original analysis So far, so good

(Some would object to my step from "*e* would have had no chance of occurring" to "*e* would not have occurred" They say that things

[2] See Jennifer Hornsby, *Actions* (London Routledge & Kegan Paul, 1980), Chapter II and for the second version, see also G H von Wright, *Explanation and Understanding* (London Routledge and Kegan Paul, 1971) pp 76–81 I am indebted to Hornsby, and to Alison McIntyre, for discussion on this point

[3] The paper was shortened at the request of the Program Chairman of the American Philosophical Association (Eastern Division) The full-length version (May 1973) advocated the same treatment of probabilistic causation that is presented in this post script

with no chance at all of occurring, that is with probability zero, do nevertheless happen, for instance when a fair spinner stops at one angle instead of another, yet any precise angle has probability zero I think these people are making a rounding error they fail to distinguish zero chance from infinitesimal chance Zero chance is *no* chance, and nothing with zero chance ever happens The spinner's chance of stopping exactly where it did was not zero, it was infinitesimal, and infinitesimal chance is still *some* chance)

But there is a second case to be considered *c* occurs, *e* has some chance *x* of occurring, and as it happens *e* does occur, if *c* had not occurred, *e* would still have had some chance *y* of occurring, but only a very slight chance since *y* would have been very much less than *x* We cannot quite say that without the cause, the effect would not have occurred, but we can say that without the cause, the effect would have been very much less probable than it actually was In this case also, I think we should say that *e* depends causally on *c*, and that *c* is a cause of *e*

It does not matter whether *x* itself, the actual chance of the effect, is high or low Suppose you mischievously hook up a bomb to a randomizer—a genuinely chancy one, if need be one that works by counting clicks in a counter near a radioactive source If you set the randomizer to a high probability, that makes it likely that your act of setting up the bomb will cause an explosion If you set the randomizer to a low probability, that makes it less likely that your act will cause an explosion But no matter how you set the randomizer, if the bomb does chance to go off, then your act does cause the explosion For no matter how you set the randomizer, we can be sure that the explosion would have been very much less probable still if you hadn't set up the bomb at all

(You took it in stride when you read my words if you set the randomizer low, that makes it less likely that your act will cause an explosion That proves my point For suppose that improbable events cannot be caused the actual chance *x* has to be high, or at least has to exceed some lowish threshold, in order to have a case of causation Then if you set the randomizer low enough, that doesn't just make it *unlikely* that your act will cause an explosion—it makes it downright *impossible* But "unlikely" did *seem* the right word "Don't worry—set the randomizer below 0 17% and you can't *possibly* cause an explosion "—Not so!)

Several points of clarification may be helpful (1) Chances are time-dependent an event may have different chances at different times

before it occurs The actual chance x of e is to be its chance at the time immediately after c, and the counterfactual is to concern chance at that same time (2) I do not assume that there is some y that would definitely have been the chance of e in the absence of c Maybe so, maybe not Maybe in that case the chance of e might have had any of various values, all of them much less than the actual chance x In saying that without c, e would have had some chance y much less than x, "some chance y" is a quantifier whose scope is limited to the consequent of the conditional (3) "Much less" means less by a large factor—not by a large difference If x is already small, the difference of y and x could not be large It is x that sets the standard for how small the chance of e must be without c We could have one case in which the absence of a cause would lower the chance of an effect from 100% to 10%, another in which the lowering would be from 10% to 1%, yet another in which the lowering would be from 1% to 0 1%, and all would count equally as cases of chancy causal dependence So it will not do to simplify our counterfactual and say that without c, the chance of e would be low *simpliciter* (4) A chance event may be caused, but we should not say that it is caused to happen *rather than not* Contrastive causal statements differ from plain ones According to what I say about contrastive questions and statements in "Causal Explanation" (in this volume), there can be no contrastive causal explanation of why a chance event occurs rather than not

Many probabilistic theories of causation share the motivating idea that a cause increases the probability of the effect Mine differs from some of the others in two respects [4] First, it is meant to apply to causation in the single case causation by one particular event of another event, not conduciveness of one kind of event to another kind Hence its probabilities are single-case chances, as opposed to finite or

[4] For the other sort of probabilistic theories of causation see *inter alia* Patrick Suppes, *A Probabilistic Theory of Causality* (Amsterdam North-Holland, 1970), and Nancy Cartwright, Causal Laws and Effective Strategies, *Nous* 13 (1979) 419–37, reprinted with additions in her *How the Laws of Physics Lie* (Oxford Clarendon Press, 1983) Cartwright does not offer her theory as an analysis, as such it would be circular, but it might nevertheless succeed as a constraint relating causation to probabilities

An analysis much closer to mine, except that it does not provide for (what I would call) causation without causal dependence and it avoids reference to events, is that of D H Mellor, Fixed Past, Unfixed Future in Barry Taylor, ed , *Contributions to Philosophy Michael Dummett* (The Hague Nijhoff, 1986)

limiting frequencies You may not like single-case chances—I don't either—but I cannot see how to make sense of certain well-established scientific theories without them If we need them anyway, we may as well use them here (I discuss single-case chances, and the reason for disliking them, elsewhere in this volume, see "A Subjectivist's Guide to Objective Chance," especially the final section and Postcript C, also my discussion of Humean supervenience in the introduction)

Second, my analysis is in terms of counterfactual conditionals about probability, not in terms of conditional probabilities If we try to use an inequality of conditional probabilities to express that event *c* raises the probability of event *e*, we run into a well-known difficulty The inequality may well hold not because *c* causes *e*, but rather because *c* and *e* are two effects of a common cause One cure is to use fancier conditional probabilities conditionalize not just on the absence of *c*, but on that together with a specification of background Then the problem is to say, preferably without circular mention of causation, what information should be included in this background

But even if that problem can be solved, another remains Conditional probabilities, as standardly understood, are quotients They go undefined if the denominator is zero If we want to say, using conditional probabilities, that *c* raises the probability of *e*, we will need probabilities conditional on the non-occurrence of *c* (plus background, perhaps) But there is no guarantee that this conditional probability will be defined What if the probability that *c* occurs (given background) is one? What if *c* has been predetermined through all of past time—what if its probability has *always* been one, so that even by going back in time we cannot find a non-zero chance of *c*'s failing to occur? For that matter, what if we want to apply our probabilistic analysis of causation to a deterministic world in which all probabilities (at all times) are extreme one for all events that do occur, zero for all that don't? The requisite conditional probabilities will go undefined, and the theory will fall silent That is not acceptable Earlier, I said that it would not do to impute error to indeterminists who accept commonplace causal statements, therefore we cannot accept an analysis of causation that works only under determinism Likewise it would not do to impute error to determinists who accept commonplace causal statements, therefore we cannot accept an analysis that works only under indeterminism An adequate analysis must be neutral It must work in both cases And it must work in a uniform way, for it does not seem that our concept of causation is disjunctive A probabilistic analysis (of single-case causation) that uses conditional probabilities is

not neutral It is made for indeterminism My analysis, on the other hand, can serve alike under indeterminism or determinism [5]

My motivating idea is that causes make their effects more probable, but that is written into the analysis of causal dependence, not of causation itself As in my original analysis, we have causation when we have a causal chain one or more steps of causal dependence The effect need not depend on the cause directly When we have causation without direct causal dependence, as in some cases of preemption, it is not necessarily true that the cause at the beginning of the chain raises the probability of the effect at the end The cause might lower the probability of the effect, or might leave it unchanged At each step in the chain, we have a cause raising the probability of its immediate effect But since counterfactuals are not transitive, that does not settle whether there is raising over the entire chain

Suppose we have two redundant systems to produce the same effect One is much more reliable than the other—that is, much less subject to random failure part way along the causal chain The reliable system is already started, left to itself, it will very probably produce the effect But I do not leave it to itself There is a switch that both turns off the reliable system and turns on the unreliable system, and I throw this switch As luck would have it, the unreliable system works The effect ensues, just as it would probably have done without my act My act did not make the effect more probable, but rather less, since I put the unreliable system in place of the reliable one Nevertheless, I did cause the effect And the reason is plain if we consider some intermediate event in the causal chain that actually occurred, something that happened well after the reliable system was already turned off That event was part of the working of the unreliable system, so it would not have occurred, or at least it would have been improbable, if I had not thrown the switch But by the time of the intermediate event, the reliable system was already out of action So without that event, the effect would not have occurred, or at least it would have been very improbable (Here it is crucial that the counterfactual be governed by a similarity relation that does not conduce to backtracking, see "Coun-

[5] It would be possible to squander this advantage of the counterfactual analysis, of course One could interpret the counterfactuals themselves in such a way that they make non trivial sense only under indeterminism take as accessible counterfactual situations only those courses of events that once had some non-zero probability of coming to pass A probabilistic theory of counterfactuals along these lines would make it child s play to confute the determinist out of his own mouth—an advantage that might commend it to some philosophers, but to me seems a sufficient *reductio*

terfactual Dependence and Time's Arrow" in this volume) My act raised the probability of the intermediate event, and thereby caused it And the intermediate event raised the probability of the effect, and thereby caused it So my act caused a cause of the effect, and thereby caused it—despite lowering its probability [6]

I have said that if distinct events *c* and *e* both occur, and if the actual chance of *e* (at a time *t* immediately after *c*) is sufficiently greater than the counterfactual chance of *e* without *c*, that implies outright that *c* is a cause of *e* Some philosophers find this counterintuitive They would correct me thus

> No, if there would have been some residual chance of *e* even without *c*, then the raising of probability only makes it *probable* that in this case *c* is a cause of *e* Suppose, for instance, that the actual chance of *e*, with *c*, was 88%, but that without *c*, there would still have been a 3% probability of *e* Then most likely (probability 97%) this is a case in which *e* would not have happened without *c*, then *c* is indeed a cause of *e* But this just might be (probability 3%) a case in which *e* would have happened anyway, then *c* is not a cause of *e* We can't tell for sure which kind of case this is

It is granted, *ex hypothesi*, that it would have been a matter of chance whether *e* occurred Even so, the objection presupposes that the case must be of one kind or the other either *e* definitely *would* have occurred without *c*, or it definitely would *not* have occurred If that were so, then indeed it would be sensible to say that we have causation only in case *e* definitely would not have occurred without *c* My original analysis would serve, the amendment suggested in this postscript would be unwise, and instead of having a plain case of probabilistic causation we would have a probable case of plain causation

But I reject the presupposition that there are two different ways the world could be, giving us one definite counterfactual or the other That presupposition is a metaphysical burden quite out of proportion to its intuitive appeal, what is more, its intuitive appeal can be explained away

The presupposition is that there is some hidden feature which may or may not be present in our actual world, and which if present would

[6] Compare Wesley C Salmon s discussion of explanations that do not increase weight in Salmon *et al* , *Statistical Explanation and Statistical Relevance* (Pittsburgh University of Pittsburgh Press, 1971), pp 62–65 For a more wholehearted adherence to the thesis that causes make their effects more probable, see Nancy Cartwright, Causal Laws and Effective Strategies, and D H Mellor Fixed Past, Unfixed Future

make true the counterfactual that *e* would have occurred anyway without *c*. If this counterfactual works as others do, then the only way this hidden feature could make the counterfactual true is by carrying over to the counterfactual situation and there being part of a set of conditions jointly sufficient for *e*.

What sort of set of conditions? We think at once that the set might consist in part of laws of nature, and in part of matters of historical fact prior to the time *t*, which would together predetermine *e*. But *e cannot* be predetermined in the counterfactual situation. For it is supposed to be a matter of chance, in the counterfactual situation as in actuality, whether *e* occurs. That is stipulated as a hypothesis of the case. When an event is predetermined, there cannot be any genuine chance that it will not happen. Genuine chance gives us the residue of uncertainty that is left after *all* laws and prior conditions have been taken into account.

(Here I assume that we are not dealing with an extraordinary situation, involving time travel perhaps, in which the normal asymmetries of time break down, and the past contains news from the future. That is fair. The objection concerns what should be said about *ordinary* cases of probabilistic causation.)

So the hidden feature must be something else. But what else can it be? Not the historical facts prior to *t*, not the chances, not the laws of nature or the history-to-chance conditionals that say how those chances depend on the prior historical facts. For all those are already taken account of, and they suffice only for a chance and not a certainty of *e*.

There is the rest of history: everything that happens after *t*. These future historical facts are not relevant to the chances at *t*, *e* can still have a chance of not occurring even if there are facts of later history that suffice for its occurrence. As there will be if it does occur, that is itself a fact of later history. In the terminology of "A Subjectivist's Guide to Objective Chance" (in this volume) later history is "inadmissible." So perhaps that is where the hidden feature of the world is to be found.

But this also will not do. For we know very well that if we give weight to future similarities, so that facts of the later history of our world tend to carry over into counterfactual situations, then we will get into trouble. We will get counterfactuals that seem false in themselves, and that also yield false conclusions about causation. We must make sure, either by fiat or else by tailoring our standards of similarity to exploit the *de facto* asymmetries of time, that future similarities will

normally carry no weight (See "Counterfactual Dependence and Time's Arrow" in this volume.) Features of our actual future history may be well hidden, sure enough, and they might well enter into sets of conditions and laws sufficient to postdetermine *e*, but what they will not do is carry over into the counterfactual situation without *c*.

(Normally I am forced to admit exceptions of two kinds, for reasons discussed in Postcript D to "Counterfactual Dependence and Time's Arrow" in this volume. If a reconvergence to actual history could be accomplished without widespread miracles or quasi-miraculous coincidences, then I would admit that actual future history carries over into the counterfactual situation, and I would admit that the absence of such quasi-miracles carries over. But I think the first cannot apply to the truth of counterfactuals at a world like ours, and the second could apply only to the special case where *e* itself would be quasi-miraculous. So these exceptions are not relevant to our present discussion.)

So the hidden feature must be something else still: not a feature of the history of this world, and also not a feature of its chances, or of the laws or conditionals whereby its chances depend on its history. It fails to supervene on those features of the world on which, so far as we know, all else supervenes. To accept any such mysterious extra feature of the world is a serious matter. We need some reason much more weighty than the isolated intuition on which my opponent relies.[7] Without such a reason, it would be better to suppress the intuition.

[7] Some people *do* have more weighty reasons, though I do not think they are reasons that we ought to accept. Theological reasons, perhaps: if God is to be properly omniscient, and if He is to exercise divine providence without running risks, He had better know just what would happen if He made creatures whose choices were not predetermined. Then there have to be definite counterfactual facts for Him to know, even if they cannot supervene on any features of the world that we would otherwise believe in, and accordingly de Molina, Suarez, and (sometimes) Plantinga posited that there are these facts. See Robert M. Adams, "Middle Knowledge and the Problem of Evil," *American Philosophical Quarterly* 14 (1977): 109–17. Or physical reasons, perhaps: P. H. Eberhard, "Bell's Theorem without Hidden Variables," *Il Nuovo Cimento* 38 B (1977): 75–79, and likewise Nick Herbert and Jack Karush, "Generalization of Bell's Theorem," *Foundations of Physics* 8 (1978): 313–17, fulfill the promise of their titles by appeal to a principle of 'counterfactual definiteness'. This principle says that even if a measurement was not made, and its outcome would have been a matter of chance if it had been made, nevertheless there is some definite value that it would have given. These counterfactual measurement outcomes do not supervene on the wave function which is the usual complete quantum mechanical description of a physical system. It is considered nice that we can get Bell's Theorem using just the counterfactual outcomes, instead of trafficking in hidden variables as traditionally conceived: though for my own part, I cannot tell the difference.

Which is all the easier if it rests on a mistake in the first place, and I think it does I suspect that my opponent is someone who has not wholeheartedly accepted my stipulation of the case in question Stipulation or no, he remains at least somewhat inclined to think that the case involves not genuine chance, but a kind of counterfeit chance that is compatible with determinism (See Postcript B to "A Subjectivist's Guide to Objective Chance" in this volume) Perhaps he clear-headedly thinks that counterfeit chance is all the chance there could ever be, and so is all that could be meant by the word "chance " Or perhaps he thinks double, and thinks of the case half one way and half the other

If it *is* a case of counterfeit chance, then his objection is well taken For then *e* is after all predetermined one way or the other, both in actuality and in the counterfactual situation without *c*, but predetermined partly by details of prior historical fact that are far too minute to be discovered in advance So we do indeed have an unproblematic hidden feature of the actual world—namely, the relevant configuration of minute details—that carries over to the counterfactual situation and there joins in predetermining the outcome one way or the other

That is all very well, but then his objection is off target I was not speaking of a case of counterfeit chance, I insist, but of a different case probabilistic causation of a genuine chance event If my opponent believes that my case is impossible because counterfeit chance is all the chance there can be, let him say so, but let him not reinterpret my case to fit his own doctrines

When my opponent says that either *e* would have occurred without *c* or else *e* would not have occurred without *c*, he sounds like Robert Stalnaker [8] But his position is not the same, though he accepts the same disjunction of counterfactuals, and Stalnaker's defense of such disjunctions is no use to him My opponent thinks there are two relevant ways the world might be, one of them would make true one of the disjoined counterfactuals, the other would make true the other, so the disjunction is true either way Stalnaker, like me, thinks there is only one relevant way for the world to be, and it does not make either counterfactual determinately true But Stalnaker, unlike me, thinks the disjoined counterfactuals are true or false relative to alternative arbitrary resolutions of a semantic indeterminacy, what makes the coun-

[8] See Robert C Stalnaker, A Theory of Conditionals, in Nicholas Rescher, ed , *Studies in Logical Theory* (Oxford Blackwell 1968) and A Defense of Conditional Excluded Middle in *Ifs*, ed by William Harper *et al* (Dordrecht Reidel, 1980) I discuss Stalnaker s theory in Counterfactuals and Comparative Possibility and Causal Decision Theory, both in this volume

terfactuals lack determinate truth is that different resolutions go different ways, but every resolution makes one or the other true, so the disjunction is determinately true despite the complementary indeterminacies of its disjuncts A resolution of an alleged semantic indeterminacy is not a hidden fact about the world, and that is the difference between Stalnaker and my opponent Stalnaker disagrees with me on a small point of semantics, my opponent, on a large point of ontology A resolution of an indeterminacy might indeed be *mistaken* for a hidden fact about the world—Stalnaker suggets, plausibly, that such mistakes are common So if we accepted Stalnaker's view on the point of semantics, that would give us a second way to explain away my opponent's problematic intuition

C INSENSITIVE CAUSATION

Killing, so they say, is causing to die I am sure that I—and likewise you, and each of us—have caused ever so many people to die, most of them people yet unborn Acts of mine are connected to their deaths by long chains of causal dependence [9] But I have never killed anyone—I hope

For instance, suppose I write a strong recommendation that lands someone a job, so someone else misses out on that job and takes another, which displaces a third job-seeker, this third job-seeker goes elsewhere, and there meets and marries someone, their offspring and all their descendants forevermore would never have lived at all, and *a fortiori* would never have died, and so presumably their deaths would not have occurred, but for my act [10] Maybe there is a time after which *every* death that occurs is one that would not have occurred but for my act It would be strange to single out my act as *the* cause of all those deaths But it is *a* cause of them, under my analysis and also according to our common usage And still I deny that I have ever killed

For a still more striking case, consider the Big Bang This event, I

[9] Not acts of omission if such there be In the next postcript I shall consider causation by omission, but for the present I am discussing cases in which we have what is uncontroversially a genuine act—or more generally, a genuine event—to do the causing

[10] It has been observed in other connections that who will live in the future is a very sensitive matter depending very much on the great and small events of the present and past See Derek Parfit, Future Generations Further Problems, *Philosophy and Public Affairs* 11 (1982) 113–72, and Robert M Adams, Existence Self-Interest, and the Problem of Evil *Nous* 13 (1979) 53–65

take it, is a cause of *every* later event without exception Then it is a cause of every death But the Big Bang did not kill anyone

So killing must be a special kind of causing to die But what distinguishes this special kind of causation?

Not that there must be one single step of causal dependence, as opposed to an intransitive chain An act of killing can be a preempting cause It can be you who kills the victim, even though another killer was standing by who would have done the job for you—causing the victim to die the very same death—if he had not seen you lay the poison yourself

Not that the chance of the effect must be high If you hook up a bomb to a randomizer and hide it in a crowded place, and it happens to go off, you can kill no matter how low you set the chance

Not that the causal chain must be short You can kill by delayed action If you set a hidden time bomb with a thousand-year fuse, you may well kill someone yet unborn

Not that the chain must be simple You can kill someone by means of a lethal Rube Goldberg machine

Not that the chain must be foreseeable You can kill someone no matter how good your reasons were for thinking the gun was not loaded, or no matter how unfeasible it would have been for you to discover in advance his lethal allergy to what you fed him

Not that the chain must pass through no later human actions [11] If you kill by setting a baited mantrap, or by making a gift of poisoned chocolates, your unsuspecting victim's action is an intermediate step in the causal chain whereby you kill him In other cases, an action by a third party may be an intermediate step you make a gift of poisoned chocolates to the host, who offers them to the guest

Perhaps a cluster of these conditions, inadequate if taken one by one, would work to distinguish the kind of causing that can be killing I think not But the counterexamples get too contrived to be very persuasive imagine a lethal Rube Goldberg machine with a randomizer at one step, a thousand-year fuse at another, an alternative waiting in reserve at another, dependence on some action of the unsuspecting victim at another, and no way to discover how it works

[11] *Pace* Jennifer Hornsby *Actions*, pp 127–30 While disagreeing with Hornsby s general claim, I disagree less about the examples that motivate it, examples in which somebody causes a dinghy to sink by ordering someone else to sink it or causes a death by ordering someone else to kill See the final part of this postcript I am indebted on this point to discussion with Hornsby

I suggest a different way to distinguish the right kind of causing by its insensitivity to circumstances When an effect depends counterfactually on a cause, in general it will depend on much else as well If the cause had occurred but other circumstances had been different, the effect would not have occurred To the extent that this is so, the dependence is sensitive Likewise if a causal chain consists of several steps of causal dependence, we can say that the chain is sensitive to the extent that its steps are (On average? Or at worst?) Sensitivity is a matter of degree, however It may be that the causation depends on an exceptionally large and miscellaneous bundle of circumstances all being just right If any little thing had been different, that cause would not have caused that effect But sometimes causation is comparatively insensitive to small differences in the circumstances When my strong recommendation causes lives and then deaths, that is comparatively sensitive causation—there are many differences that would have deflected the chain of events But if you shoot at your victim point-blank, only some very remarkable difference in circumstances would prevent his death The same is true if you set a Rube Goldberg machine, or a delayed-action bomb, working inexorably toward its lethal outcome The case of the bomb with a randomizer also is comparatively insensitive the bomb might very well have chanced not to go off, but it isn't the fine details of the circumstances that would make the difference

Jonathan Bennett restates my suggestion this way killing requires "that the causal chain run through a stable and durable structure rather than depending on intervening coincidental events "[12] A lethal Rube Goldberg machine may work in many steps, it may be full of thousand-year fuses and randomizers and alternatives waiting in reserve, its working may require the responses of unsuspecting agents, there may be no way to discover how it is built or understand how it would work, and yet it may be no less "stable and durable" for all that, and the causal chain running through it may be far more independent of "intervening coincidental events" than are most of the causal chains in the wider world

So it seems that the reason why a lot of causing to die is not killing is, at least partly, that the causing to die in killing must be causation of a comparatively insensitive kind And if this is so for killing, perhaps it is so likewise for other causatives Consider the ways in which you can and can't make, break, wake, or bake things

[12] Killing and Letting Die, in Sterling W McMurrin, ed , *The Tanner Lectures on Human Values*, Volume II (Cambridge Cambridge University Press, 1981) p 71

Insensitivity is not the same thing as any of the unsatisfactory conditions that I considered above, but of course it is connected to several of them *Ceteris paribus*, shortness and simplicity of the chain will make for insensitivity, insensitivity, in turn, will make for foreseeability The more the chain depends on a lot of circumstances being just right, the harder it is for a would-be predictor to know all he needs to know about the circumstances The sensitivity of the chain is an obstacle to prediction Unforseeability does not imply sensitivity, since any of many other obstacles to prediction might be at work But unforseeability sets a minimum standard If a chain is insensitive enough that you can predict it, then it is insensitive enough that you can kill by it Perhaps our common knowledge of what can normally be predicted sets a common standard for everyone Or perhaps the standard varies What if you are much better than I am at predicting chains that are somewhat sensitive? I am inclined to say that if so, then indeed you can kill in ways that I cannot If your act and mine cause death by chains that are exactly alike, and if the duplicate chains are insensitive enough to fall within your powers of prediction but sensitive enough to frustrate mine, then you kill but I do not [13]

My suggestion faces a problem Recall that you can kill by a causal chain that has someone else's action as an intermediate step you give someone poisoned chocolates, he unsuspectingly serves them to his guest, and thereby you kill the guest (It is true as well that the host unwittingly kills the guest But that is beside the point, the question is whether you kill the guest, and I submit that you and the host both do) But if you tell the host that the chocolates are poisoned, and you order or hire or coerce or persuade him to serve them anyway, then it seems that you do not kill the guest You may be no less guilty, morally and in the eyes of the law, than if you had killed him, or no less praiseworthy, if the guest was Hitler Be that as it may, it seems that you don't kill by getting someone else to kill *knowingly* Why not, on the suggestion I have advanced? It seems that if someone else is ready to kill knowingly when ordered or hired or coerced or persuaded to, his readiness well might be a stable and durable structure, so that by depending on this readiness, the causal chain from your action to a death well might be fairly insensitive to fortuitous circumstances

I reply that indeed that might be so, and nevertheless we might speak as if it were not so That would be no surprise Part of our habitual

[13] At this point, I am indebted to Jonathan Bennett

respect for other people consists in thinking that they are sensitive to a great variety of considerations, and therefore not easy to predict or control It is all very well to take for granted that someone is ready to offer a guest what he takes to be harmless chocolates, to that extent, it is not disrespectful to regard his dispositions as a stable and durable structure Offering chocolates is no big deal It is another thing to take for granted that someone can be ordered or hired or coerced or persuaded to kill knowingly That is to take him altogether too much for granted The relevant disrespect lies not in thinking him willing to kill, whether that is disrespectful depends on the circumstances and the victim Rather, it lies in thinking of his readiness to kill as stable and durable, inexorable, insensitive to fortuitous circumstances of the case, so that he is disposed to make weighty choices with unseemly ease

Such disrespect might be well deserved We might know very well that this dull thug before us would never think twice about killing for a small fee Therefore, we might be sure that when you hire him, the causal chain from your action to the victim's death is as inexorable and insensitive as if it had passed instead through some strong and sturdy machine But we might know this, and yet be halfhearted in putting our mouths where our minds are Some vestige of our habitual respect might well influence how we speak If I am right, when you cause death by hiring this thug, you are in literal truth a killer, no less than the thug himself is If we deny it, I suggest that we are paying the thug a gesture of respect—insincere, undeserved, yet unsurprising

That was an uncompromising version of my reply I can offer an alternative version that runs as follows If you hire the thug just considered, you are *not* in literal truth a killer The truth conditions for "kill" are not just a matter of insensitive causation They make an exception for insensitive causal chains that run through someone else's action of knowingly killing However, insensitivity remains the underlying idea The extra twist in the truth conditions is not just a brute complication of the concept, it is there, understandably, thanks to our respectful presumption that a causal chain through someone else's weighty decision will not be insensitive The two versions agree about what we say, and why we say it, they differ only about what is literally true *Ceteris paribus* it is bad to claim that we say what we know is literally false, but *ceteris paribus* it is bad to build complicating exceptions into the conditions of literal truth Between the version that does one and the version that does the other, I think there is little to choose I am not even confident that there is a genuine issue between the two

D CAUSATION BY OMISSION[14]

An omission consists of the nonoccurrence of any event of a certain sort To suppose away the omission is, exactly, to suppose that some event of the given sort does occur We say that omissions may be caused, and may cause, and I have no wish to deny this I would like to be able to provide for causation by omission within the general framework given in this paper and in "Events" (in this volume) Unfortunately, I do not see how to make it fit with all that I say in general about events and about their causal dependence So, one way or another, a special case it must be

Omissions as effects are no special problem I must allow in any case that sometimes, by causing suitable events, causes can create a pattern of events, and that a fact can supervene on this pattern even when there is no genuine event that can be called the obtaining of that fact, in which case the causes of the events in the pattern can also be said to cause that fact to obtain [15] For instance, it is at least a fact that Xanthippe became a widow I think there is no genuine event that can be called Xanthippe's becoming a widow But the causes of her marriage together with the causes of Socrates's death may nevertheless be said to have caused her to become a widow they caused genuine events that comprised a pattern on which the fact that she became a widow supervened Certainly this fact is not beyond the reach of causal explanation Likewise I can say that various distractions caused Fred to omit the precautions he should have taken, and in saying this, I needn't grant that there was any such thing as an event of omission If there are events of omission, well and good But I don't need them as effects

Do I need them as causes? There are two opposite strategies that I might follow, and a third which is a compromise between those two One way or another, all of them treat causation by omission as a special case While I would guess that any of the three could be made to work, I am not in a position to prove it by presenting fully developed versions I am not sure how much the three really differ, certainly some of their difference is just terminological

[14] In this postcript, I am much indebted to discussion with Jonathan Bennett and with Alison McIntyre

[15] Here I do not rely on any fancy theory of facts, they are simply truths That is to say they are the true ones among whatever entities can be said to bear truth values On this view as opposed to some fancy theories, most facts are only accidentally facts They are contingent truths, and might have been falsehoods

The first strategy accepts that there are events of omission What is more, there are events essentially specifiable as omissions For instance, Fred's omission of precautions, essentially specifiable as such, is an event that would have occurred no matter how he omitted them, no matter what else he did instead, and that could not have occurred if he had taken the precautions For any event, there are necessary and sufficient conditions, normally hard to state, for that very event to occur Some descriptions of an event are built into its conditions of occurrence, others are not The first strategy says that the description of this event as an omission *is* built in Then to suppose counterfactually that this event of omission does not occur is equivalent to supposing that Fred does take the precautions So the counterfactual analysis of causation can apply to events of omission just as it does to all other events, and it is safe to say, as we ought to, that the effects which depend causally on Fred's omission are those which would not have occurred if he had taken the precautions This strategy requires no exception to what I say about causation in general

But it does require an exception to what I say about events in general For I say that a theory of events, if it is built to serve the needs of my analysis of causation, must reject overly disjunctive events An alleged event would be disjunctive if, or to the extent that, it could have occurred in various dissimilar ways (The point is not that its conditions of occurrence could be formulated as a disjunction—anything can be formulated as a disjunction—but that they could be formulated as a disjunction of overly varied disjuncts) An alleged event that is essentially specifiable as a talking-or-walking, and which could have occurred either as a talking or as a walking, is an example of what ought to be rejected The reason is that if it were accepted as an event, then it could qualify as a cause, but it is intuitively very wrong to say that the talking-or-walking causes anything But if we are to accept events of omission, in the way we are considering, then we may not reject disjunctive events without exception For an event of omission, essentially specifiable as such, is highly disjunctive Fred omits the precautions if he does something else during the period in which he was supposed to attend to them So there are as many different ways for the event of omission to occur as there are alternative ways for Fred to spend the time An event essentially specifiable as an omission amounts to an event essentially specifiable as a sleeping-or-loafing-or-chatting-or- with a disjunct for everything Fred might do other than attending to the precautions If omissions are accepted as genuine events and as causes, while other alleged disjunctive events are rejected,

that makes causation by omission a special case [16] The unfinished business for this strategy, of course, is to draw the line how do we distinguish the genuine omissions from other alleged events that we should still reject? For instance, I think we ought not to say that the laws of nature, or other regularities, cause things, yet regularities may be made out to be omissions on a cosmic scale—the universe omits to contain events that would violate them We must somehow deny that we have here a genuine event of omission

The second, opposite, strategy says that there are no events of omission Then there is no need to make a place for them within a theory of events, and no need to worry that they would be like other alleged events that are to be rejected So far, so good But in that case, I need to make an exception to what I say about causation itself For it is not to be denied that there is causation by omission, and I cannot analyze this in my usual way, in terms of counterfactual dependence between distinct events Instead I have to switch to a different kind of counterfactual for the special case The counterfactual is not if event *c* (the

[16] Compare Jonathan Bennett s account of the distinction between killing and letting die in Whatever the Consequences , *Analysis* 26 (1966) 83–102, especially pp 94–96 He presents the same distinction more fully in Killing and Letting Die , but there gives it a new name—positive versus negative instrumentality—because he observes that other considerations somewhat affect the ordinary usage of the ordinary terms I agree, but shall ignore those considerations here

There are ever so many ways you might move (or hold still—let us count this as one way of moving) during a period of time Suppose that if you were to move in any way that falls within the range *L*, someone would live, whereas if you were to move in any way that falls within the complementary range *D*, he would die, and you move in a way that falls within range D, so he dies Have you killed him? Or have you let him die, in other words omitted to save his life? (To avoid irrelevant issues, suppose (1) that the dependences are insensitive in the sense of the previous postcript, (2) that he would die the same death no matter how you moved within the range *D*, and (3) that this is not one of those special cases in which you could be said *both* to kill someone *and* to let him die, and by the very same conduct) Bennett suggests, I think rightly, that if the range *L* is wide and varied compared to *D*, then you have killed him, whereas if the range *L* is narrow and uniform compared to *D*, then you have let him die

I note that if the range *D* is wide, then an alleged event essentially specified as a moving some-way-in-*D* is disjunctive and therefore suspect and this suspect event would be essentially specified as a letting-die and thus as an event of omission Not so if the range *D* is narrow On the strategy presently under consideration there are such events of omission, on the strategy to be considered next there are not Of course I am not suggesting that these two strategies have different moral implications Whatever events there may or may not be what matters is that someone s life depended on how you moved

omission) had *not* occurred It is rather if some event of kind *K* (the omitted kind) *had* occurred

But if we use special counterfactuals for the special case, that opens several questions Again we need to draw a line I thought it necessary to block some counterexamples against a counterfactual analysis of causation by insisting that counterfactual dependence was to be between events If we give that up, what new line shall we retreat to? As before, alleged causation by the laws of nature, regarded as cosmic omissions, will illustrate the problem Also, I thought it necessary to insist on distinctness between events that stand in causal dependence, and by distinctness I meant more than nonidentity (See "Events ") But how does distinctness apply to causation by omission? Fred sleeps, thereby omitting precautions against fire and also omitting precautions against burglary Two distinct omissions?

The third, compromise strategy accepts events of omission as causes, but this time, the events of omission are not essentially specified as such Fred omits the precautions, sleeping through the time when he was supposed to attend to them His nap was a genuine event, it is not objectionably disjunctive There are many and varied ways in which he could have omitted the precautions, but there is just one way that he did omit them We could plausibly say, then, that his nap *was* his omission of precautions But accidentally so His nap could have occurred without being an omission of precautions if (1) that very nap had been taken somewhat later, with the precautions seen to beforehand, or conceivably (2) if he had taken the precautions somehow in his sleep, or (3) if that very nap could have been taken by someone else, or (4) if the precautions had not been his responsibility (I take it that (2)–(4) are problematic in various ways, so I rest my case mainly on (1)) And an omission of precautions might very well have occurred without being that nap he might have stayed awake and done any of many other things instead of attending to the precautions Still, *as it was*, the nap was what happened instead of the taking of precautions So we may call it an event of omission, though we do not thereby capture its essence We can have events of omission, so understood, and still reject disjunctive events without exception

But this third strategy, like the second, demands special counterfactuals for the special case Even if Fred's nap was his omission of precautions, it is one thing to suppose that this very event did not occur, and it is another thing to suppose that no event that occurred (this or any other) was an omission of the precautions It is one thing to suppose away the event *simpliciter*, another thing to suppose it away *qua*

omission It is the second counterfactual supposition, not the first, that is relevant to causation by omission For it is the second supposition that is equivalent to supposing that the precautions were taken But this is special In other cases the relevant counterfactuals are those that suppose away an event *simpliciter*, we do not in general need to suppose away events *qua* satisfying some or other accidental specification

As with the second strategy, resort to special counterfactuals for the special case threatens to undo our defenses against various counterexamples It remains to be seen how, if at all, those defenses could be rebuilt This strategy, like the others, leaves us with unfinished business

E REDUNDANT CAUSATION[17]

Suppose we have two events c_1 and c_2, and another event e distinct from both of them, and in actuality all three occur, and if either one of c_1 and c_2 had occurred without the other, then also e would have occurred, but if neither c_1 nor c_2 had occurred, then e would not have occurred Then I shall say that c_1 and c_2 are *redundant causes* of e

(There might be redundant causation with a set of more than two redundant causes There might be probabilistic redundant causation, in which e would have had some small chance of occurring even if neither c_1 nor c_2 had occurred There might be stepwise redundant causation without direct dependence, as described by Louis Loeb [18] I pass over these complications and consider redundant causation in its simplest form)

As in my definition of ordinary causation, the counterfactuals concern particular events, not event-kinds So it is not redundant causation if you shoot a terminal cancer patient—or, for that matter, a healthy young mortal—who would sooner or later have died anyway Without your act

[17] In this postscript I am much indebted to discussion with John Bigelow with John Etchemendy and with Louis Loeb

[18] Causal Theories and Causal Overdetermination *Journal of Philosphy* 71 (1974), 525–44 The simplest stepwise case is as follows, there could be more steps or more events at any step We have five actual events c_1, c_2 d_1, d_2, e with e distinct from the d s and the d s distinct from the c s If neither of the c s had occurred then neither of the d s would have occurred but if either of the c s had occurred alone, then one of the d s would have occurred If neither of the d s had occurred, then e would not have occurred but if either of the d s had occurred alone, then e would have occurred So the c s redundantly cause e by way of the d s But if neither of the c s had occurred, e would have occurred anyway, so we do not have direct redundant causation of e by the c s

he would have died a different death numerically different, because very different in time and manner The particular event which is the death he actually dies would not have occurred If you shoot a man who is being stalked by seven other gunmen, that *may* be redundant causation—the answer depends partly on details of the underdescribed case, partly on unsettled standards of how much difference it takes to make a different event If you shoot a man who is simultaneously being shot by seven other members of your firing squad, that doubtless is redundant causation The exact number of bullets through the heart matters little

If one event is a redundant cause of another, then is it a cause *simpliciter*? Sometimes yes, it seems, sometimes no, and sometimes it is not clear one way or the other When common sense delivers a firm and uncontroversial answer about a not-too-far-fetched case, theory had better agree If an analysis of causation does not deliver the common-sense answer, that is bad trouble But when common sense falls into indecision or controversy, or when it is reasonable to suspect that far-fetched cases are being judged by false analogy to commonplace ones, then theory may safely say what it likes Such cases can be left as spoils to the victor, in D M Armstrong's phrase We can reasonably accept as true whatever answer comes from the analysis that does best on the clearer cases It would be still better, however, if theory itself went indecisive about the hard cases If an analysis says that the answer for some hard case depends on underdescribed details, or on the resolution of some sort of vagueness,[19] that would explain nicely why common sense comes out indecisive

In my paper, I distinguished one kind of case—preemption with chains of dependence—in which common sense delivers clear positive and negative answers, and my counterfactual analysis succeeds in agreeing I left all other cases of redundant causation as spoils to the victor, doubting that common-sense opinions about them would be firm and uncontroversial enough to afford useful tests of the analysis

[19] Or on resolution of an ambiguity Loeb (*op cit*) has offered a counterfactual analysis of causation in a broad sense—he calls such causes C conditions —which would include redundant causes whether or not they are causes on my narrower analysis Likewise Ardon Lyon s counterfactual analysis in Causality' *British Journal for the Philosophy of Science* 18 (1967) 1–20 is modified so that it includes some redundant causes I fear that such analyses, though perhaps suited to Loeb s purpose in formulating causal theories of memory *et al* are too broad to correspond to any ordinary sense of the word cause Be that as it may, it remains possible that the hard cases are causes in one sense but not in another If so, then if the counterfactual approach is right it ought to afford analyses for all the senses

Now I would distinguish more varieties of redundant causation Sometimes my analysis, as it stands, agrees with clear common-sense answers, positive or negative Sometimes it reproduces common-sense indecision Sometimes I am still content to leave far-fetched cases as spoils to the victor But sometimes it seems that additions to my original analysis are needed

I consider first a class of cases distinguished by doubt as to whether they exhibit redundant causation at all I have already mentioned one example you shoot a man who is being stalked by seven other gunmen As it actually happens, the man dies on Tuesday morning, face down on the ground, his heart pierced by your bullet, with an entry wound in his back and an exit wound in his chest Without your act he would have died on Wednesday evening, slumped in a chair, his heart pierced by someone else's bullet, with an entry wound in his chest and an exit wound in his back

Is it that without your act he would have died a different death—numerically different because somewhat different in time and manner?[20] If so, there is no redundancy The particular death he actually died depends counterfactually on your act, without which that very

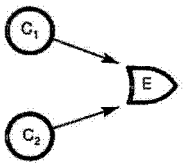
[20] Here and in what follows, I assume that difference in time or manner is what it takes to make a numerical difference between an event that actually occurs and one that would have occurred under some counterfactual supposition That is contrary to a view put forward by Peter van Inwagen in Ability and Responsibility, *Philosophical Review* 87 (1978) 201–24, especially pp 208–209, and in *An Essay on Free Will* (Oxford Clarendon Press 1983), pp 167–70 He suggests that an event which actually occurs as the product of certain causes could not have occurred without being the product of those causes, nor could those causes have had a different event as their product He finds this view plausible in part because of its analogy to the view that human beings, tables, etc should be individuated by their causal origins

I reject his view for two reasons First, because it would ruin my project of analyzing causation in terms of counterfactual dependence It would trivialize any counterfactual to the effect that without the cause the effect would not have occurred Second, because it is *prima facie* implausible I can legitimately entertain alternative hypotheses about how an event (or for that matter a human being, or a table) was caused, or I can entertain alternative plans about how some desired future event is to be caused But if I do, then I certainly seem to be presupposing that one and the same event might be produced by various different causes (Compare van Inwagen s own remark that we seem to presuppose that one and the same event might have had different effects) But van Inwagen s view implies that things are not as they seem either my hypotheses or plans (with at most one exception) are hidden impossibilities, or else they are not about a particular event at all, but rather they involve some highly specific kind of event These reconstruals seem artificial and not to be accepted without better reason than van Inwagen gives

event would not have occurred This is straightforward causation Or is it rather that without your act he would have died the very same death—numerically the same, despite slight differences in time and manner? If so, there is genuine redundancy In that case your act would be a redundant cause, whether it would be a cause *simpliciter* awaits our discussion of the varieties of genuine redundant causation

It is hard to say which is true It would remain hard, I think, no matter how fully we described the details of what actually happened, and of what would have happened under our counterfactual hypothesis

Here is another example Suppose three neurons are hooked up thus



Suppose that a neuron fires if stimulated by the firing of one or more other neurons connected to it by a stimulatory synapse (shown by a forward arrowhead) But suppose—fictitiously, I believe—that a neuron fires much more vigorously if it is doubly stimulated than if it is singly stimulated Neurons C_1 and C_2 fire simultaneously, thereby doubly stimulating E, which fires vigorously Is this vigorous firing of E a different event from the feeble firing that would have occurred if either one of C_1 and C_2 had fired alone? Then we have joint causation, in which the effect depends counterfactually on each of the causes, and there is no redundancy Or is it that numerically the same firing would have occurred, despite a difference in manner, with single stimulation? Then we have redundant causation Again it is hard to say, and again the difficulty cannot be blamed on underdescription of the details

Call an event *fragile* if, or to the extent that, it could not have occurred at a different time, or in a different manner A fragile event has a rich essence, it has stringent conditions of occurrence In both our examples we have redundant causation if the effect is not too fragile, ordinary causal dependence on joint causes otherwise

Don't say here we have *the events*—how fragile are *they*? Instead it should be here we have various candidates, some more fragile and some less—which ones do we call the events? (For instance under my proposal in "Events," in this volume, the candidates will be smaller and larger classes of possible spatiotemporal regions, more and less

tightly unified by similarity) Properly posed, the question need not have a fully determinate answer, settled once and for all Our standards of fragility might be both vague and shifty

As of course they are You can say the performance should have been postponed until the singer was over his laryngitis, then *it* would have been better You can just as well say, and mean nothing different the performance should have been cancelled, and another, which would have been better, scheduled later to replace it There's no right answer to the question how fragile the performance is Not because there is something—the performance—with an indeterminate size in logical space[1] But because there are various things, with various sizes, and we haven't troubled to decide which one is "the performance " Likewise every region of the earth has exact boundaries and a determinate size Silicon Valley, whatever exactly that is, is no exception However we haven't decided exactly how big a region is called "Silicon Valley " That's why there's no right answer to the question whether these words (written on the Stanford campus) were written in Silicon Valley

So there may be no right answer to the question whether we have a case of joint causation without redundancy, or whether instead we have a case of redundant causation, which might or might not count as causation according to considerations to be discussed later The answer depends on the resolution of vague standards of fragility If common sense falls into indecision and controversy over such cases, that is only to be expected

It is a common suggestion to adopt extreme standards of fragility, and thereby make away with redundant causation altogether Even if a man is shot dead by a firing squad, presumably it would have made *some* minute difference to the time and manner of his death if there had been seven bullets instead of eight So if you fired one of the eight bullets, that made some difference, so if his death is taken to be very fragile indeed, then it would not have occurred without your act Under sufficiently extreme standards of fragility, the redundancy vanishes Even this turns out to be a case in which the effect depends on each of several joint causes Likewise for other stock examples of redundancy

(Suppose we did follow this strategy wherever we could Wouldn't we still have residual cases of redundancy, in which it makes *absolutely* no difference to the effect whether both of the redundant causes occur or only one? Maybe so, but probably those residual cases would be mere possibilities, far-fetched and contrary to the ways of this world Then we could happily leave them as spoils to the victor For we could plausibly suggest that common sense is misled its habits of thought are

formed by a world where every little thing that happens spreads its little traces far and wide, and nothing that happens thereafter is quite the same as it would have been after a different past.)

Extreme standards of fragility would not fit a lot of our explicit talk about events. We do say—within limits!—that an event could have been postponed and could have happened differently. But this is not a decisive objection. The standards that apply within the analysis of causation might differ from those that apply in explicit talk.

What matters more is that extreme standards would not fit a lot of our negative judgements about causation itself. Extreme fragility of effects would make for spurious causal dependence in many quite ordinary cases. It would make more trouble than it cures.[21]

For instance, suppose there was a gentle soldier on the firing squad, and he did not shoot. If the minute difference made by eight bullets instead of seven is enough to make a different event, then so is the minute difference made by eight instead of nine. So if the victim's death is so very fragile that it would not have occurred without your act, equally it is so fragile that it would not have occurred without the gentle soldier's omission. If by reason of fragility the death depends causally on your act, then equally it depends causally on the omission. So the gentle soldier caused the death by *not* shooting, quite as much as you caused it by shooting! This is a *reductio*.

That case may puzzle us because it involves at least an appearance of redundancy, and also because it involves causation by omission. But the problem arises for cases without these complications. Boddie eats a big dinner, and then the poisoned chocolates. Poison taken on a full stomach passes more slowly into the blood, which slightly affects the time and manner of the death. If the death is extremely fragile, then one of its causes is the eating of the dinner. Not so.

To be sure, resolution of vagueness is influenced by context, and I can imagine a special context in which we might after all agree that the eating is a cause of the death. Pleased that Boddie is dead but horrified that the death was lingering, the poisoner says: if only he hadn't eaten, *this* wouldn't have happened—and by "this" he means the death, taken as very fragile. Maybe indeed that context makes it right to say that the eating caused the death. But it is also right, certainly in other contexts and probably even in this one, to say what is true under more lenient

[21] I owe this point to Ken Kress, *circa* 1968.

and more ordinary standards of fragility namely, that the eating did *not* cause the death [22]

So if we wanted to make away with the stock examples of redundant causation, what we would require is not a uniformly stringent standard of fragility, but rather a double standard—extremely stringent when we were trying to show that an effect really depends on its alleged redundant causes, but much more lenient when we were trying to agree with common-sense judgements that an effect is not caused by just anything that slightly affects its time and manner It is not out of the question that there should be such a double standard But if there is, an adequate theory of causation really ought to say how it works (The changes of standard noted above, brought on by contextual pressures, are not the ones we want—they cut across cases with and without apparent redundancy) To say how the double standard works may not be a hopeless project, but for the present it is not so much unfinished as unbegun

Extreme fragility of effects might get rid of all but some far-fetched cases of redundant causation, but it leads to trouble that we don't know how to control Moderate fragility gets rid of some cases and casts doubt on others, but plenty are left Our topic has not disappeared

So I return now to genuine redundant causation, including the doubtful cases when taken under standards of fragility that make them genuine I divide it into *preemption* and (*symmetrical*) *overdetermination* [23] In a case of preemption, the redundant causes are not on a par It seems clear that one of them, the *preempting cause*, does the causing, while the other, the *preempted alternative*, waits in reserve The alternative is not a cause, though it could and would have been one, if it had not been preempted There is the beginning of a causal process running from the preempted alternative to the effect But this process does not go to completion One effect of the preempting cause is to cut it off In a case of overdetermination, on the other hand, there is no such asymmetry between the redundant causes It may or may not be

[22] How can it ever be right to say *A*, and equally right to say not-*A*?—Because sometimes what you say is itself the decisive part of the context that resolves vagueness and sets the standards whereby the truth value of what you say is determined Say *A* and thereby you set standards under which *A* is true, so you speak truly But say not-*A* instead, and you speak just as truly for in that case you set standards under which *A* is false See Scorekeeping in a Language Game in my *Philosophical Papers*, Volume I

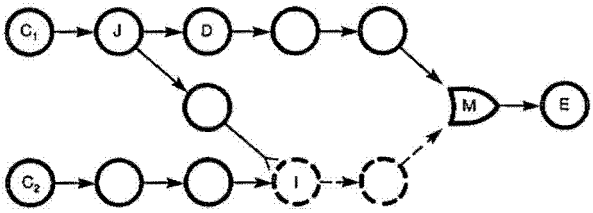
[23] I shall use the word overdetermination narrowly, to imply symmetry and exclude cases of causal preemption

clear whether either is a cause, but it is clear at least that their claims are equal There is nothing to choose between them Both or neither must count as causes

First, preemption It is clear what answer we want—the preempting cause is a cause, the preempted alternative is not—and any analysis that does not yield that answer is in bad trouble It is easy for me to say why the preempted alternative is not a cause the effect does not depend on it My problem is to say why the preempting cause *is* a cause, when the effect does not depend on it either (A regularity analysis of causation has the opposite problem why is the preempted alternative not a cause, when it is part of a set of conditions jointly sufficient for the effect?)

I subdivide preemption into *early* and *late* In early preemption, the process running from the preempted alternative is cut off well before the main process running from the preempting cause has gone to completion Then somewhere along that main process, not too early and not too late, we can find an intermediate event to complete a causal chain in two steps from the preempting cause to the final effect The effect depends on the intermediate, which depends in turn on the preempting cause (Or, in cases with more than one preempted alternative, we might need more steps) We have a causal chain of stepwise dependence between the cause and the effect, even if not dependence *simpliciter*, and since causation is transitive, we take the ancestral of dependence Thus I say that c is a cause of e if there is a sequence $c,$ $,e$ of events, consisting of c and e and zero or more intermediates, with each event in the sequence except the first depending on the one before (Normally all these events would be distinct, and in temporal order, but I do not require this See Postcript F, below)

This is the variety of preemption that I discussed in the paper To illustrate it, let us have another system of neurons



Besides stimulatory synapses from one neuron to another, as before, we now have an inhibitory synapse as well (shown by a backward arrowhead) A neuron normally fires if stimulated, but not if it is in-

hibited at the same time Neurons C_1 and C_2 fire, thereby starting two processes of firing which make their separate ways toward neuron E The main process, which begins with the firing of C_1, goes to completion But the alternative process, which begins with the firing of C_2, is cut short because neuron I is inhibited, the neurons shown dotted never fire There is also a branch process, diverging from the main process The junction event where it diverges is the firing of neuron J It is this branch process that cuts off the alternative process by inhibiting neuron I The main and alternative processes—the one actual, the other partly unactualized—merge with the firing of neuron M, and proceed thence to the final effect, the firing of neuron E

Thus the firing of C_2 is the preempted alternative It is not a cause of the firing of E because there is no direct dependence, and neither is there any stepwise dependence via an intermediate The firing of C_1 is the preempting cause The firing of D is our intermediate event It depends counterfactually on the firing of C_1, the firing of E depends on it, and thereby we have our two-step chain of dependence from the preempting cause to the effect For by the time of the firing of D, the alternative process was already doomed The alternative process was doomed as soon as neuron J fired, though it was not yet cut off, the branch process that was going to cut it off had already diverged from the main process So if the firing of D had not occurred, *both* processes would have failed, and the firing of E also would not have occurred

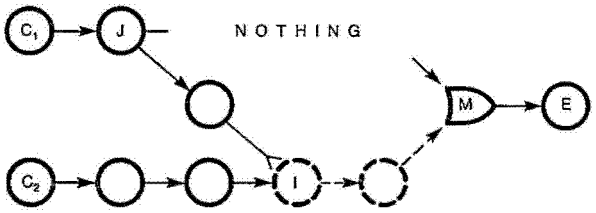
Don't say that if D had not fired, that would mean that it had not been stimulated, and that would mean that the neurons to its left on the main process would not have fired, and so neuron I would not have been inhibited, and so the alternative process would have gone to completion and E would have fired after all That is backtracking, and backtracking counterfactuals, however legitimate in other contexts, are out of place in tracing causal dependence (See "Counterfactual Dependence and Time's Arrow" in this volume) Of course it is not just to deal with early preemption that we must avoid backtracking, as is explained in the paper, the avoidance of backtracking is needed also to solve the problems of effects and of epiphenomena

We have some choice which event goes along the main process to take as our intermediate The firing of J comes too early the effect does not depend on it, since without it the alternative process would not have been cut off The firing of M comes too late it lies on the unactualized alternative process as well as on the main process, and so does not depend on the preempting cause But anything in between would do What makes the solution possible is that there exists some

intermediate event in the gap between too early and too late And so it is, generally, in cases of early preemption Thus we distinguish the genuine cause from its preempted alternative, as we should, even though either one by itself would have sufficed to cause the effect [24]

Late preemption is harder Our solution cannot succeed unless there is a sufficient gap between too early and too late, if not-too-early is already too late, there is no place for an intermediate event to complete a chain of stepwise dependence

There are two far-fetched ways in which this problem might arise The first way involves action at a temporal distance Suppose that in our previous example, we remove all the neurons between J (too early) and M (too late)



In their place, suppose we have some law of delayed action that directly connects the firings of J and M Iff J fires, then M fires a certain time later (as in the original example) but that is absolutely all there is to it—there is no connection betwen the two neurons, and no continuous causal process between their two firings That is possible, I take it, though it goes against what we take to be the ways of this world In such a case, we have no intermediate event to complete our chain of dependence

The second way involves infinite multiple preemption We have infinitely many preempted alternatives, and infinitely many cut-off alternative processes Suppose for simplicity that the main process and its unactualized alternatives merge only at the final effect (Otherwise the problem would be the same, but with the point of merging in place of the final effect) Then any other event on the main process is not too late to depend on earlier events along that process The problem is to

[24] See Bruce LeCatt Censored Vision, *Australasian Journal of Philosophy* 60 (1982) 158–62, for further examples of stepwise dependence in cases of early preemption

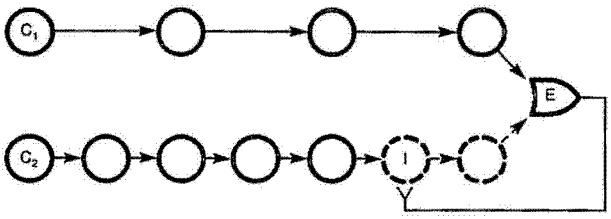
find an intermediate event that is not too early to take the penultimate place in our chain of stepwise dependence—that is, to find an event on which the final effect depends Such an event has to come late enough that by the time it occurs, all of the infinitely many alternative processes are doomed Any one of the alternative processes is eventually doomed, so there is an event that comes late enough so far as it is concerned Likewise for any finite set But since there are infinitely many alternatives, there may be no event before the final effect that comes after *all* the alternative processes are doomed Suppose one of them is doomed 128 seconds before the final effect, another only 64 seconds before, another only 32 seconds before, Then at no time before the final effect are all of them doomed Then there is no intermediate event on which the final effect depends Our causal chains of stepwise dependence can get as close as we like to the final effect, but they never can reach it Then there is no stepwise dependence between the effect and what seems to be its preempting cause [25]

I do not worry about either of these far-fetched cases They both go against what we take to be the ways of this world, they violate the presuppositions of our habits of thought, it would be no surprise if our common-sense judgements about them therefore went astray—spoils to the victor! Common sense does judge them to be cases of causal preemption, in which what seems to be a preempting cause is indeed a cause, despite the lack of either direct or stepwise dependence But an analysis that disagrees may nevertheless be accepted It would be better to agree with common sense about these cases, to be sure, but that is not an urgent goal

Unfortunately there is another variety of late preemption, quite commonplace and not at all far-fetched, and there it *is* an urgent goal to agree with common sense Again we have what seems to be a preempting cause, hence a cause *simpliciter*, but no dependence and no stepwise dependence Here my analysis seems to be in trouble These are cases in which an alternative process is doomed only when the final effect itself occurs The alternative is cut off not by a branch process that diverges from the main process at a junction event before the effect is reached, but rather by a continuation of the main process beyond the effect Shooting a man stalked by seven other gunmen would be a case of this kind, if it is a case of redundant causation at all, and if the other gunmen desist only when they see him dead Another case would be

[25] See William K Goosens, Causal Chains and Counterfactuals *Journal of Philosophy* 76 (1979) 489–95

this system of neurons Again we start with the simultaneous firings of neurons C_1 and C_2, which redundantly cause the firing of E



I ignored such cases when I wrote the paper, and for many years afterward My reason must have been that there is a ready-made solution fragility of the effect If the alternative process is only doomed by the effect itself, and if at the time of the effect it is not yet complete, then the alternative process must run more slowly So if it had been left to produce the effect, the effect would have been delayed Without the firing of neuron C_1 (the seeming main cause) the firing of neuron E would have been delayed by the time it takes for three extra neurons to fire, if you had not shot the man on Tuesday morning, he would not have died until Wednesday evening, and so on, for all such cases We can devise cases in which the delay is very short, but we can never get rid of it altogether (Or not without resort to instantaneous or backward causation But then the case becomes far-fetched, not worrisome, spoils to the victor) If the effect is taken to be fragile, then the delay would suffice to give us a numerically different event instead of the effect that actually occurred We would have causal dependence without redundancy, thus agreeing with common sense that your shooting the man on Tuesday, or the firing of C_1, or whatever, is indeed a cause

But my reason for ignoring these cases was a bad reason, because the ready-made solution is a bad solution Fragility of the effect is no better as a remedy for these cases of late preemption than it is as a remedy for redundant causation generally To deal with all the cases, including those where the delay is very short and there is not much difference in manner to go with it, we need extreme standards of fragility, uniformly extreme standards are no good because they will give us lots of spurious causal dependence, so we need a double standard, and that might be workable, for all we know, but we don't know how to make it work There are two problems One is that a double standard must be principled We need some definite rule to tell us when we should raise the standard when is dependence among fragile versions relevant,

and when is it not, to causation among the original robust events? The second problem is that a stringent standard may give the wrong answer Let c_1 be a preempting cause of e, and let c_2 be the preempted alternative, in a case of late preemption Without c_1, e would have been delayed, and so a more fragile version of e would not have occurred at all So far, so good But it may also be that some side effect of c_2 substantially influences the time and manner of e, in which case, unfortunately, a version of e that is fragile enough to depend on c_1 may depend on c_2 as well Indeed, it may take more fragility to give us the dependence on c_1 that yields the right answer than it does to give us the dependence on c_2 that yields the wrong answer Though I don't reject the fragility approach out of hand, I don't see how to make it work [26] So I am inclined to prefer a different solution, though it is more of a departure from my original analysis in the paper

Leaving the problem of late preemption in abeyance, consider this question Suppose we have processes—courses of events, which may or may not be causally connected—going on in two distinct spatiotemporal regions, regions of the same or of different possible worlds Disregarding the surroundings of the two regions, and disregarding any irrelevant events that may be occurring in either region without being part of the process in question, what goes on in the two regions is exactly alike Suppose further that the laws of nature that govern the two regions are exactly the same Then can it be that we have a causal process in one of the regions but not the other? It seems not Intuitively, whether the process going on in a region is causal depends only on the intrinsic character of the process itself, and on the relevant laws The surroundings, and even other events in the region, are irrelevant Maybe the laws of nature are relevant without being intrinsic to the region (if some sort of regularity theory of lawhood is true) but nothing else is

Intuitions of what is intrinsic are to be mistrusted, I think They too often get in the way of otherwise satisfactory philosophical theories Nevertheless, there is some slight presumption in favor of respecting them Let us see where this one leads us

A process in a region may exhibit a pattern of counterfactual dependence that makes it causal, according to my original analysis Its later parts may depend counterfactually on its earlier parts (later and earlier in time, normally, but all I require is that there be dependence with

[26] I am indebted to discussion with D H Rice, who has persuaded me that it would be premature to give up on fragility solutions without a good deal of further investigation

respect to some order), and in particular, its last event may depend on its first (We will provide for stepwise dependence later)

Now suppose that some process in some region does not itself exhibit this pattern of dependence, but suppose that in its intrinsic character it is just like processes in other regions (of the same world, or other worlds with the same laws) situated in various surroundings And suppose that among these processes in other regions, the great majority—as measured by variety of the surroundings—do exhibit the proper pattern of dependence This means that the intrinsic character of the given process is right, and the laws are right, for the proper pattern of dependence—if only the surroundings were different, and different in any of many ways According to my original analysis, this process is nevertheless not causal Thanks to its special bad surroundings, it is a mere imitation of genuine causal processes elsewhere But that goes against our motivating intuition

So we might extend the analysis Suppose that there exists some actually occurring process of the kind just described, and that two distinct events *c* and *e* are the first and last in that process Then let us say that *e quasi-depends* on *c* We might wish to count that as one kind of causation, based derivatively on counterfactual dependence even though there is no dependence between those two events themselves As before, we must take an ancestral to ensure that causation will come out transitive, thereby providing not only for chains of stepwise dependence, but also for chains of stepwise quasi-dependence, or mixed chains To this end we could redefine a *causal chain* as a sequence of two or more events, with either dependence or quasi-dependence at each step And as always, one event is a *cause* of another iff there is a causal chain from one to the other

That would solve the problem of late preemption, both in the commonplace cases that worry me and in the far-fetched cases that do not For the problem is that we seem to have a causal process starting with a preempting cause, and ending with the final effect, and yet this process does not exhibit the proper pattern of counterfactual dependence, not even if we count stepwise dependence Segments of it do exhibit dependence, but we cannot patch these segments together to make a chain that reaches all the way to the effect What spoils the dependence is something extraneous the presence alongside the main process of one or more preempted alternatives Without them, all would be well Hold fixed the laws but change the surroundings, in any of many ways, and we would have the dependence that my original analysis requires for causation But as is, we have quasi-dependence instead of

dependence So if we extend the analysis, and allow causation by quasi-dependence, that solves our problem We then can agree with common sense that we have genuine preemption, and genuine causation by the preempting cause [27]

The extended analysis, which allows causation by quasi-dependence, is more complicated than my original analysis, and it is less purely a counterfactual analysis, though of course counterfactual dependence still plays a central role The complication would be objectionable if it were just a hoky gimmick to deal with late preemption, but it is not just that For what it is worth, we also have independent motivation in the intuition of intrinsicness While I would still welcome a different solution to the problem of late preemption, within my original analysis, I now think that the extended analysis may well be preferable

This completes my discussion of preemption I now turn to the other variety of redundant causation overdetermination, with nothing to break the symmetry between the redundant causes When I wrote

[27] If we admit causation by quasi dependence, it would be nice if that could buy us some simplification elsewhere Could we perhaps drop the part of the analysis in which we take an ancestral to ensure that causation turns out transitive? I think not, in view of a case suggested by John Etchemendy Suppose we have a case of preemption with this peculiarity there is no way, given the laws of nature, that the preempting cause could fail to have been accompanied by the preempted alternative Any lawful way of producing one must produce the other as well It seems that we have a main causal process running from a preempting cause *c* to its final effect *e* Because of the preemption, *e* does not depend directly on *c* And neither do we have direct quasi-dependence any process just like it, and under the same laws, must likewise have its dependence destroyed by preemption The problem comes not from an accident of circumstances, but from the laws themselves So if we admit causation by quasi-dependence but we do not take an ancestral, we still get the wrong answer

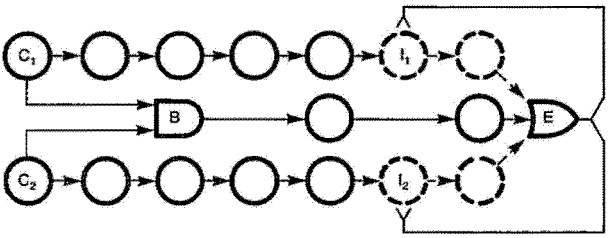
But take some intermediate event *d* along the main process from *c* to *e*, before the point where it merges with the alternative process For the first step, we have causation by dependence *d* does depend on *c* For the second step, we may have causal dependence of *e* on *d* if the preemption is early We may not, if the preemption is late, but even so, assuming that *d* could have been produced without also producing the preempted alternative, we at least have quasi-dependence of *e* on *d* So we have a chain from *c* to *d* to *e*, with dependence or quasi dependence at both steps Then if we take an ancestral to ensure that causation comes out transitive, we get the right answer

What if there is *no* intermediate that could lawfully have been produced without also producing a preempted alternative? That makes the case very peculiar indeed It is central to the way we ordinarily think about preemption that we can regard the main and the alternative processes as distinct and separable So if the laws forbid us to have even a part of the one process without the corresponding part of the other, that goes badly against our habitual presuppositions If so such common sense opinions as we may have need not be respected—spoils to the victor

the paper, I thought that all such cases were alike, that a counterfactual analysis would inevitably deny that the redundant causes in overdetermination are causes *simpliciter*, and that it did not matter much what the analysis said, since all such cases were spoils to the victor for lack of firm common-sense judgements

All that is wrong An important paper by Martin Bunzl changes the picture greatly [28] Bunzl observes that when we examine stock examples of overdetermination in detail, we can very often find an intermediate event—call it a *Bunzl event*—that satisfies two conditions First, the Bunzl event is jointly caused, without redundancy, by the same events that are redundant causes of the final effect Second, the Bunzl event seems clearly to be a cause (often a preempting cause) of the final effect Cases of overdetermination are not all alike, because there are different kinds of Bunzl events (at least three) and also because there are some possible cases, far-fetched perhaps, with no Bunzl events at all A counterfactual analysis does not deny that the redundant causes are causes *simpliciter* of the final effect, provided it can agree that they are causes of a Bunzl event and that the Bunzl event in turn is a cause of the effect The cases should not all be left as spoils to the victor, because once a Bunzl event is noticed, it becomes clear to common sense that we have genuine causation

One kind of Bunzl event is a preempting cause in a case of late preemption This system of neurons illustrates it Here B is an especially lethargic neuron It will not fire if singly stimulated, but it will if doubly stimulated



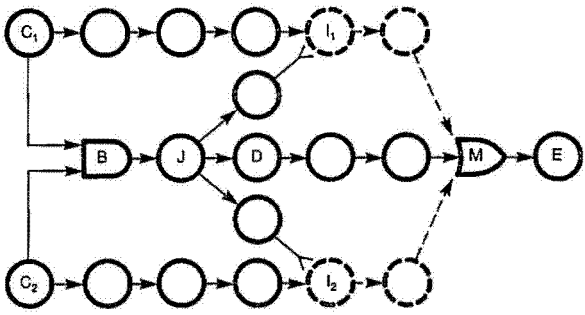
As usual, the simultaneous firings of C_1 and C_2 are redundant causes of the firing of E But also they are joint causes, without redundancy, of a Bunzl event namely, the firing of B And that is a preempting cause of the final effect The preemption is late the two alternative processes,

[28] Causal Overdetermination, *Journal of Philosophy* 76 (1979) 134–50

those that run from the firings of C_1 and C_2 taken separately, are cut off only because of the effect itself It is the firing of E itself that inhibits neurons I_1 and I_2 We must apply whatever solution we favor for late preemption generally If somehow we had a double standard of fragility, we might say that the firing of E is extremely fragile, and would not have occurred (though E would still have fired) without the firing of B Or, probably better, we might say that despite a lack of direct or stepwise dependence, we have causation by quasi-dependence Thanks to the intrinsic character of the course of events running from the firing of B to the firing of E, we would have had counterfactual dependence if just such a course of events had occurred under the same laws but in any of various different surroundings According to the extended analysis, that means that the firing of B is a cause of the firing of E Either way, we say as we should that the firing of B causes the firing of E, and therefore, by transitivity, the firing of C_1 and C_2 are both causes of E

Not only is the firing of B a preempting cause, we can also think of the entire course of events as a case of *self*-preemption The firings of C_1 and C_2, our redundant causes, both preempt and are preempted Taken together as joint causes, they preempt themselves taken separately

A second kind of Bunzl event is a preempting cause in a case of early preemption, as in the following system of neurons

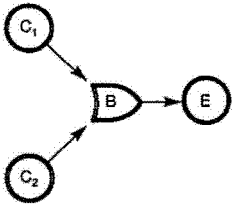


Again the simultaneous firings of C_1 and C_2 are redundant causes of the firing of E, and joint causes without redundancy of the firing of the lethargic neuron B As in simpler cases of early preemption, the firing of D completes a chain of stepwise dependence it depends on the firing of B, and in turn the firing of E depends on the firing of D So there is a three-step causal chain of dependence from the firing of C_1 to the firing

of E, and likewise from the firing of C_2. More simply, there is a two-step chain, since the firing of D also depends directly on the firing of C_1, and likewise on the firing of C_2. The firing of B is a Bunzl event, so is the firing of D, and so are various other intermediate events on the chain. Again we have self-preemption by our redundant causes: the firings of C_1 and C_2 taken jointly preempt themselves, taken separately.

This looks complicated. But just the same sort of early self-preemption can happen in much simpler cases of overdetermination, as follows.

The third kind of Bunzl event is a fragile intermediate. Earlier, we considered a case of fragility of the effect, involving a neuron that would fire vigorously if doubly stimulated, feebly if singly stimulated. We considered that under moderate and reasonable standards of fragility, hence without any problematic double standard, we might say that the vigorous firing and the feeble firing would differ enough in manner to make them numerically different events. If we place the fragile vigorous firing as an effect, what we have is not redundant causation at all. But if we place it as an intermediate, it can be the Bunzl event in a case of overdetermination. Here is such a case, with B as the neuron that may fire either vigorously or feebly.



The vigorous firing of B that actually occurs depends on both of the simultaneous firings of C_1 and C_2. Without either one of these causes it would not have occurred. The feeble firing of B that would have occurred with only one of them would not have been the same event. But also the firing of E depends on the firing of B. So each of our redundant causes is connected to the final effect by a two-step causal chain of dependence. Not by direct dependence: if only one of C_1 and C_2 had fired, so that B fired feebly, E would still have been stimulated and its firing would have been very little different. This is not a case that can be treated by fragility of the effect, or not under moderate standards of fragility.

(My solution depends on assuming that if the intermediate event—the vigorous firing of B—had not occurred, then B would not have fired at all. It isn't that the vigorous firing would have been replaced by

a feeble firing, differing only just enough not to be numerically the same That may seem to go against a similarity theory of counterfactuals—wouldn't the minimal change to get rid of an event be one that replaces it with a barely different event? Not so, a similarity theory needn't suppose that just any sort of similarity we can think of has nonzero weight It is fair to discover the appropriate standards of similarity from the counterfactuals they make true, rather than *vice versa* (See "Counterfactual Dependence and Time's Arrow" in this volume) And we certainly do not want counterfactuals saying that if a certain event had not occurred, a barely different event would have taken its place They sound false, and they would make trouble for a counterfactual analysis of causation not just here, but quite generally)

The case looks simpler than the self-preemption cases above, but it is really much the same The process from the redundant causes jointly through the vigorous firing to the effect goes to completion The two alternative processes from the redundant causes taken singly through the feeble firing to the effect are cut short when the feeble firing does not occur The feeble firing is prevented by the double stimulation of B, and that is an event in the main process

Still there is one important difference from previous cases When we have a fragile intermediate, as opposed to the sorts of Bunzl events considered above, there is room for serious indeterminacy Just as our vague and shifty standards of fragility may leave it unsettled whether we have a fragile effect, so they may leave it unsettled whether we have a fragile intermediate Then they may leave it unsettled whether we have overdetermination with or without a Bunzl event If that is what decides whether the redundant causes are causes *simpliciter*, that question too may have no right answer

So I turn to the last variety of redundant causation overdetermination without a Bunzl event, including doubtful cases when taken under standards of fragility that give no relevant fragility either in the effect or in the intermediates According to my original analysis, the redundant causes in such a case are not causes *simpliciter*, because there is neither direct nor stepwise dependence But the extended analysis would disagree There is quasi-dependence of the effect on each of the two redundant causes, and if we allowed causation by quasi-dependence, that would make the redundant causes count as genuine causes of the effect

Also, the original analysis will say that in cases where it is doubtful whether there is a fragile effect or intermediate, then it is likewise doubtful whether the redundant causes are causes *simpliciter* Whereas

the extended analysis would say that in such cases the redundant causes are causes, though the reason why is left doubtful The first analysis would be better suited to explain indecision and controversy, the second would be better suited to explain positive judgements

I used to think that all cases of overdetermination, as opposed to preemption, could be left as spoils to the victor, and that is what I still think about these residual cases All the more so, given Bunzl's discussion of what we find when we look at realistic cases in microscopic detail, without simplifying idealizations For it seems that cases without Bunzl events require phenomena with perfectly sharp thresholds, whereas thresholds under the laws of this world are imperfectly sharp Thus I am content to say that these cases may go one way or the other The decision will depend on what strategy emerges as victor in the cases that really matter—namely, the commonplace cases of late preemption

I should dispel one worry that if we ever decline to count redundant causes as genuine causes, then we will be left with gaps in our causal histories—no cause at all, at the time when the redundant causes occur, for a redundantly caused event That is not a problem For consider the larger event composed of the two redundant causes (I mean their mereological sum *Not* their disjunction—I do not know how a genuine event could be the disjunction of two events both of which actually occur It would have to occur in any region where either disjunct occurs Hence it would have to occur twice over in one world, which a particular event cannot do See "Events" in this volume) Whether or not the redundant causes themselves are genuine causes, this larger event will be there to cause the effect For without it—if it were completely absent, with neither of its parts still present, and not replaced by some barely different event—the effect would not occur For *ex hypothesi* the effect would not occur if both redundant causes were absent, and to suppose away both of them is just the same as to suppose away the larger event that is composed of them

F SELF-CAUSATION

My requirement that cause and effect be distinct applies to causal dependence, but not to causation generally Two events are distinct if they have nothing in common they are not identical, neither is a proper part of the other, nor do they have any common part Despite the truth of the appropriate counterfactuals, no event depends causally

on itself, or on any other event from which it is not distinct However, I do allow that an event may cause itself by way of a two-step chain of causal dependence *c* depends on *d* which depends in turn on *c*, where *d* and *c* are distinct Likewise for longer closed causal loops, or for loops that lead from an event back not to itself but to another event from which it is not distinct Thus I have taken care not to rule out the sort of self-causation which appears in time-travel stories that I take to be possible (See "The Paradoxes of Time Travel" in this volume)

But no event can be self-caused unless it is caused by some event distinct from it Indeed, no event can be caused at all unless it is caused by some event distinct from it Likewise no event can cause anything unless it causes some event distinct from it

Suppose we think of the entire history of the world as one big event It is not caused by any event distinct from it, else that distinct event both would and would not be part of the entire history Likewise it does not cause any event distinct from it So it has no causes or effects at all Not as a whole, anyway Its parts, of course, do all the causing there is in the world

Some philosophers wish to believe only in entities that have some causal efficacy [29] Either they must reject such totalities as the big event which is the whole of history, or else they should correct their principle They might admit those inefficacious things that could have been efficacious if, for instance, there had been more of history than there actually was Or, more simply, they might admit those inefficacious things that are composed entirely of efficacious parts

[29] For instance, see D M Armstrong, *Universals and Scientific Realism*, Volume I (Cambridge Cambridge University Press, 1978) pp 128–32

TWENTY-TWO

Causal Explanation*

I CAUSAL HISTORIES

Any particular event that we might wish to explain stands at the end of a long and complicated causal history We might imagine a world where causal histories are short and simple, but in the world as we know it, the only question is whether they are infinite or merely enormous

An explanandum event has its causes These act jointly We have the icy road, the bald tire, the drunk driver, the blind corner, the approaching car, and more Together, these cause the crash Jointly they suffice to make the crash inevitable, or at least highly probable, or at least much more probable than it would otherwise have been And the crash depends on each Without any one it would not have happened, or at least it would have been very much less probable than it was

But these are by no means all the causes of the crash For one thing, each of these causes in turn has its causes, and those too are causes of the crash So in turn are their causes, and so, perhaps, *ad infinitum* The crash is the culmination of countless distinct, converging causal chains

This paper is descended, distantly, from my Hagerstrom Lectures in Uppsala in 1977, and more directly from my Howison Lectures in Berkeley in 1979

Roughly speaking, a causal history has the structure of a tree But not quite the chains may diverge as well as converge The roots in childhood of our driver's reckless disposition, for example, are part of the causal chains via his drunkenness, and also are part of other chains via his bald tire

Further, causal chains are dense (Not necessarily, perhaps—time might be discrete—but in the world as we mostly believe it to be) A causal chain may go back as far as it can go and still not be complete, since it may leave out intermediate links The blind corner and the oncoming car were not immediate causes of the crash They caused a swerve, that and the bald tire and icy road caused a skid, that and the driver's drunkenness caused him to apply the brake, which only made matters worse And still we have mentioned only a few of the most salient stages in the last second of the causal history of the crash The causal process was in fact a continuous one

Finally, several causes may be lumped together into one big cause Or one cause may be divisible into parts Some of these parts may themselves be causes of the explanandum event, or of parts of it (Indeed, some parts of the explanandum event itself may be causes of others) The baldness of the tire consists of the baldness of the inner half plus the baldness of the outer half, the driver's drunkenness consists of many different disabilities, of which several may have contributed in different ways to the crash There is no one right way—though there may be more or less natural ways—of carving up a causal history

The multiplicity of causes and the complexity of causal histories are obscured when we speak, as we sometimes do, of *the* cause of something That suggests that there is only one But in fact it is commonplace to speak of "the *X*" when we know that there are many *X*'s, and even many *X*'s in our domain of discourse, as witness McCawley's sentence "the dog got in a fight with another dog " If someone says that the bald tire was the cause of the crash, another says that the driver's drunkenness was the cause, and still another says that the cause was the bad upbringing which made him so reckless, I do not think any of them disagree with me when I say that the causal history includes all three They disagree only about which part of the causal history is most salient for the purposes of some particular inquiry They may be looking for the most remarkable part, the most remediable or blameworthy part, the least obvious of the discoverable parts, Some parts will be salient in some contexts, others in others Some will not be at all salient in any likely context, but they

belong to the causal history all the same the availability of petrol, the birth of the driver's paternal grandmother, the building of the fatal road, the position and velocity of the car a split second before the impact [1]

(It is sometimes thought that only an aggregate of conditions inclusive enough to be sufficient all by itself—Mill's "whole cause"—deserves to be called "the cause " But even on this eccentric usage, we still have many deserving candidates for the title For if we have a whole cause at one time, then also we have other whole causes at later times, and perhaps at earlier times as well)

A causal history is a relational structure Its *relata* are *events* local matters of particular fact, of the sorts that may cause or be caused I have in mind events in the most ordinary sense of the word flashes, battles, conversations, impacts, strolls, deaths, touchdowns, falls, kisses, But also I mean to include events in a broader sense a moving object's continuing to move, the retention of a trace, the presence of copper in a sample (See my "Events," in this volume)

These events may stand in various relations, for instance spatiotemporal relations and relations of part to whole But it is their causal relations that make a causal history In particular, I am concerned with relations of causal dependence An event depends on others, which depend in turn on yet others, , and the events to which an event is thus linked, either directly or stepwise, I take to be its causes Given the full structure of causal dependence, all other causal relations are given Further, I take causal dependence itself to be counterfactual dependence, of a suitably non-backtracking sort, between distinct events In Hume's words, "if the first had not been, the second never had existed "[2] (See "Causation," in this volume) But this paper is not meant to rely on my views about the analysis of causation

[1] On definite descriptions that do not imply uniqueness, see Scorekeeping in a Language Game in my *Philosophical Papers*, Volume I, and James McCawley Presupposition and Discourse Structure, in *Syntax and Semantics* 11, ed by David Dineen and Choon-kyu Oh (New York Academic Press, 1979) On causal selection, see Morton G White, *Foundations of Historical Knowledge* (New York Harper & Row, 1965), Chapter IV Peter Unger, in The Uniqueness of Causation, *American Philosophical Quarterly* 14 (1977) 177–88 has noted that not only the cause of but also the verb caused may be used selectively There is something odd—inconsistent, he thinks—in saying with emphasis that each of two distinct things caused something Even a cause of may carry some hint of selectivity It would be strange, though I think not false to say in any ordinary context that the availability of petrol was a cause of the crash

[2] *An Enquiry Concerning Human Understanding* Section VII

Whatever causation may be, there are still causal histories, and what I shall say about causal explanation should still apply [3]

I include relations of probabilistic causal dependence Those who know of the strong scientific case for saying that our world is an indeterministic one, and that most events therein are to some extent matters of chance, never seriously renounce the commonsensical view that there is plenty of causation in the world (They may preach the "downfall of causality" in their philosophical moments But whatever that may mean, evidently it does not imply any shortage of causation) For instance, they would never dream of agreeing with those ignorant tribes who disbelieve that pregnancies are caused by events of sexual intercourse The causation they believe in must be probabilistic And if, as seems likely, our world is indeed thoroughly indeterministic and chancy, its causal histories must be largely or entirely structures of probabilistic causal dependence I take such dependence to obtain when the objective chances of some events depend counterfactually upon other events if the cause had not been, the effect would have been very much less probable than it actually was (See Postscript B to "Causation," in this volume) But again, what is said in this paper should be compatible with any analysis of probabilistic causation

The causal history of a particular event includes that event itself, and all events which are part of it Further, it is closed under causal dependence anything on which an event in the history depends is itself an event in the history (A causal history need not be closed under the converse relation Normally plenty of omitted events will depend on included ones) Finally, a causal history includes no more than it must to meet these conditions

II EXPLANATION AS INFORMATION

Here is my main thesis *to explain an event is to provide some information about its causal history*

In an act of explaining, someone who is in possession of some infor-

[3] One author who connects explanation and causation in much the same way that I do but builds on a very different account of causation, is Wesley C Salmon See his Theoretical Explanation, in *Explanation*, ed by Stephen Korner (New Haven Yale University Press 1975), A Third Dogma of Empiricism, in *Basic Problems in Methodology and Linguistics*, ed by R Butts and J Hintikka (Dordrecht Reidel, 1977), and Why Ask Why? ? *Proceedings of the American Philosophical Association* 51 (1978) 683–705

mation about the causal history of some event—*explanatory information*, I shall call it—tries to convey it to someone else Normally, to someone who is thought not to possess it already, but there are exceptions examination answers and the like Afterward, if the recipient understands and believes what he is told, he too will possess the information The why-question concerning a particular event is a request for explanatory information, and hence a request that an act of explaining be performed

In one sense of the word, an explanation of an event is such an act of explaining To quote Sylvain Bromberger, "an explanation may be something about which it makes sense to ask How long did it take? Was it interrupted at any point? Who gave it? When? Where? What were the exact words used? For whose benefit was it given?"[4] But it is not clear whether just any act of explaining counts as an explanation Some acts of explaining are unsatisfactory, for instance the explanatory information provided might be incorrect, or there might not be enough of it, or it might be stale news If so, do we say that the performance was no explanation at all? Or that it was an unsatisfactory explanation? The answer, I think, is that we will gladly say either— thereby making life hard for those who want to settle, once and for all, the necessary and sufficient conditions for something to count as an explanation Fortunately that is a project we needn't undertake

Bromberger goes on to say that an explanation "may be something about which none of [the previous] questions makes sense, but about which it makes sense to ask Does anyone know it? Who thought of it first? Is it very complicated?" An explanation in this second sense of the word is not an act of explaining It is a chunk of explanatory information—information that may once, or often, or never, have been conveyed in an act of explaining (It might even be information that never could be conveyed, for it might have no finite expression in any language we could ever use) It is a proposition about the causal history of the explanandum event Again it is unclear—and again we needn't make it clear—what to say about an unsatisfactory chunk of explanatory information, say one that is incorrect or one that is too small to suit us We may call it a bad explanation, or no explanation at all

Among the true propositions about the causal history of an event, one is maximal in strength It is the whole truth on the subject—the biggest chunk of explanatory information that is free of error We

[4] An Approach to Explanation, in *Analytical Philosophy Second Series*, ed by R J Butler (Oxford Blackwell, 1965)

might call this the *whole* explanation of the explanandum event, or simply *the* explanation (But "the explanation" might also denote that one out of many explanations, in either sense, that is most salient in a certain context) It is, of course, very unlikely that so much explanatory information ever could be known, or conveyed to anyone in some tremendous act of explaining!

One who explains may provide not another, but rather himself, with explanatory information He may think up some hypothesis about the causal history of the explanandum event, which hypothesis he then accepts Thus Holmes has explained the clues (correctly or not, as the case may be) when he has solved the crime to his satisfaction, even if he keeps his solution to himself His achievement in this case probably could not be called "an explanation", though the chunk of explanatory information he has provided himself might be so called, especially if it is a satisfactory one

Not only a person, but other sorts of things as well, may explain A theory or a hypothesis, or more generally any collection of premises, may provide explanatory information (correct or incorrect) by implying it That is so whether or not anyone draws the inference, whether or not anyone accepts or even thinks of the theory in question, and whether or not the theory is true Thus we may wonder whether our theories explain more than we will ever realize, or whether other undreamt-of theories explain more than the theories we accept

Explanatory information comes in many shapes and sizes Most simply, an explainer might give information about the causal history of the explanandum by saying that a certain particular event is included therein That is, he might specify one of the causes of the explanandum Or he might specify several And if so, they might comprise all or part of a cross-section of the causal history several events, more or less simultaneous and causally independent of one another, that jointly cause the explanandum Alternatively, he might trace a causal chain He might specify a sequence of events in the history, ending with the explanandum, each of which is among the causes of the next Or he might trace a more complicated, branching structure that is likewise embedded in the complete history

An explainer well might be unable to specify fully any particular event in the history, but might be in a position to make existential statements He might say, for instance, that the history includes an event of such-and-such kind Or he might say that the history includes several events of such-and-such kinds, related to one another in such-and-such ways In other words, he might make an existential statement

to the effect that the history includes a pattern of events of a certain sort (Such a pattern might be regarded, at least in some cases, as one complex and scattered event with smaller events as parts.) He might say that the causal history has a certain sort of cross-section, for instance, or that it includes a certain sort of causal chain.

If someone says that the causal history includes a pattern of events having such-and-such description, there are various sorts of description that he might give. A detailed structural specification might be given, listing the kinds and relations of the events that comprise the pattern. But that is not the only case. The explainer might instead say that the pattern that occupies a certain place in the causal history is some biological, as opposed to merely chemical, process. Or he might say that it has some global structural feature: it is a case of underdamped negative feedback, a dialectical triad, or a resonance phenomenon. (And he might have reason to say this even if he has no idea, for instance, what sort of thing it is that plays the role of a damper in the system in question.) Or he might say that it is a process analogous to some other, familiar process. (So in this special case, at least, there is something to the idea that we may explain by analogizing the unfamiliar to the familiar. At this point I am indebted to David Velleman.) Or he might say that the causal process, whatever it may be, is of a sort that tends in general to produce a certain kind of effect. I say "we have lungs because they keep us alive", my point being that lungs were produced by that process, whatever it may be, that can and does produce all manner of life-sustaining organs. (In conveying that point by those words, of course I am relying on the shared presupposition that such a process exists. In explaining, as in other communication, literal meaning and background work together.) And I might say this much, whether or not I have definite opinions about what sort of process it is that produces life-sustaining organs. My statement is neutral between evolution, creation, vital forces, or what have you, it is also neutral between opinionation and agnosticism.

In short, information about what the causal history includes may range from the very specific to the very abstract. But we are still not done. There is also negative information: information about what the causal history does *not* include. "Why was the CIA man there when His Excellency dropped dead?—Just coincidence, believe it or not." Here the information given is negative, to the effect that a certain sort of pattern of events—namely, a plot—does not figure in the causal history. (At least, not in that fairly recent part where one might have been suspected. Various ancient plots doubtless figure in the causal histories of all current events, this one included.)

A final example The patient takes opium and straightway falls asleep, the doctor explains that opium has a dormitive virtue Doubtless the doctor's statement was not as informative as we might have wished, but observe that it is not altogether devoid of explanatory information The test is that it suffices to rule out at least some hypotheses about the causal history of the explanandum It rules out this one the opium merchants know that opium is an inert substance, yet they wish to market it as a soporific So they keep close watch, and whenever they see a patient take opium, they sneak in and administer a genuine soporific The doctor has implied that this hypothesis, at least, is false, whatever the truth may be, at least it somehow involves distinctive intrinsic properties of the opium

Of course I do not say that all explanatory information is of equal worth, or that all of it equally deserves the honorific name "explanation " My point is simply that we should be aware of the variety of explanatory information We should not suppose that the only possible way to give some information about how an event was caused is to name one or more of its causes

III NON-CAUSAL EXPLANATION?

It seems quite safe to say that the provision of information about causal histories figures very prominently in the explaining of particular events What is not so clear is that it is the whole story Besides the causal explanation that I am discussing, is there also any such thing as non-causal explanation of particular events? My main thesis says there is not I shall consider three apparent cases of it, one discussed by Hempel and two suggested to me by Peter Railton [5]

First case We have a block of glass of varying refractive index A beam of light enters at point *A* and leaves at point *B* In between, it passes through point *C* Why? Because *C* falls on the path from *A* to *B* that takes light the least time to traverse, and according to Fermat's principle of

[5] Carl G Hempel, *Aspects of Scientific Explanation and other Essays in the Philosophy of Science* (New York Free Press, 1965), p 353 Peter Railton, *Explaining Explanation* (Ph D dissertation, Princeton University 1979) I am much indebted to Railton throughout this paper, both where he and I agree and where we do not For his own views on explanation, see also his A Deductive Nomological Model of Probabilistic Explanation *Philosophy of Science* 45 (1978) 206–26, and Probability, Explanation, and Information, *Synthese* 48 (1981) 233–56

least time, that is the path that any light going from *A* to *B* must follow. That seems non-causal. The light does not get to *C* because it looks ahead, calculates the path of least time to its destination *B*, and steers accordingly! The refractive index in parts of the glass that the light has not yet reached has nothing to do with causing it to get to *C*, but that is part of what makes it so that *C* is on the path of least time from *A* to *B*.

I reply that it is by no means clear that the light's passing through *C* has been explained. But if it has, that is because this explanation combines with information that its recipient already possesses to imply something about the causal history of the explanandum. Any likely recipient of an explanation that mentions Fermat's principle must already know a good deal about the propagation of light. He probably knows that the bending of the beam at any point depends causally on the local variation of refractive index around that point. He probably knows, or at least can guess, that Fermat's principle is somehow provable from some law describing that dependence together with some law relating refractive index to speed of light. Then he knows this: (1) the pattern of variation of the refractive index along some path from *A* to *C* is part of the causal history of the light's passing through *C*, and (2) the pattern is such that it, together with a pattern of variation elsewhere that is not part of the causal history, makes the path from *A* to *C* be part of a path of least time from *A* to *B*. To know this much is not to know just what the pattern that enters into the causal history looks like, but it is to know something—something relational—about that pattern. So the explanation does indeed provide a peculiar kind of information about the causal history of the explanandum, on condition that the recipient is able to supply the extra premises needed.

Second case. A star has been collapsing, but the collapse stops. Why? Because it's gone as far as it can go. Any more collapsed state would violate the Pauli Exclusion Principle. It's not that anything caused it to stop—there was no countervailing pressure, or anything like that. There was nothing to keep it out of a more collapsed state. Rather, there just was no such state for it to get into. The state-space of physical possibilities gave out. (If ordinary space had boundaries, a similar example could be given in which ordinary space gives out and something stops at the edge.)

I reply that information about the causal history of the stopping has indeed been provided, but it was information of an unexpectedly negative sort. It was the information that the stopping had no causes at all, except for all the causes of the collapse which was a precondition of the stopping. Negative information is still information. If you request

information about arctic penguins, the best information I can give you is that there aren't any

Third case Walt is immune to smallpox Why? Because he possesses antibodies capable of killing off any smallpox virus that might come along But his possession of antibodies doesn't *cause* his immunity It *is* his immunity Immunity is a disposition, to have a disposition is to have something or other that occupies a certain causal role, and in Walt's case what occupies the role is his possession of antibodies

I reply that it's as if we'd said it this way Walt has some property that protects him from smallpox Why? Because he possesses antibodies, and possession of antibodies is a property that protects him from smallpox Schematically Why is it that something is F? Because A is F An existential quantification is explained by providing an instance I agree that something has been explained, and not by providing information about its causal history But I don't agree that any particular event has been non-causally explained The case is outside the scope of my thesis That which protects Walt—namely, his possession of antibodies—is indeed a particular event It is an element of causal histories, it causes and is caused But that was not the explanandum We could no more explain that just by saying that Walt possesses antibodies than we could explain an event just by saying that it took place What we did explain was something else the fact that something or other protects Walt The obtaining of this existential fact is not an event It cannot be caused Rather, events that would provide it with a truth-making instance can be caused We explain the existential fact by identifying the truth-making instance, by providing information about the causal history thereof, or both (For further discussion of explanation of facts involving the existence of patterns of events, see Section VIII of "Events," in this volume)

What more we say about the case depends on our theory of dispositions [6] I take for granted that a disposition requires a causal basis one has the disposition iff one has a property that occupies a certain

[6] See the discussions of dispositions and their bases in D M Armstrong, *A Materialist Theory of the Mind* (London Routledge & Kegan Paul, 1968), pp 85–88, Armstrong, *Belief Truth and Knowledge* (Cambridge Cambridge University Press, 1973), pp 11–16, Elizabeth W Prior, Robert Pargetter, and Frank Jackson, Three Theses about Dispositions, *American Philosophical Quarterly* 19 (1982) 251–57, and Elizabeth W Prior, *Dispositions* (Aberdeen Aberdeen University Press, 1985) See also Section VIII of Events, in this volume Parallel issues arise for functionalist theories of mind See my An Argument for the Identity Theory and Mad Pain and Martian Pain, in *Philosophical Papers*, Volume I and Jackson, Pargetter, and Prior Functionalism and Type-Type Identity Theories, *Philosophical Studies* 42 (1982) 209–25

causal role (I would be inclined to require that this be an intrinsic property, but that is controversial.) Shall we then identify the disposition with its basis? That would make the disposition a cause of its manifestations, since the basis is. But the identification might vary from case to case. (It surely would, if we count the unactualized cases.) For there might be different bases in different cases. Walt might be disposed to remain healthy if exposed to virus on the basis of his possession of antibodies, but Milt might be so disposed on the basis of his possession of dormant antibody-makers. Then if the disposition is the basis, immunity is different properties in the cases of Walt and Milt. Or better: "immunity" denotes different properties in the two cases, and there is no property of immunity *simpliciter* that Walt and Milt share.

That is disagreeably odd. But Walt and Milt do at least share something: the existential property of having some basis or other. This is the property such that, necessarily, it belongs to an individual X iff X has some property that occupies the appropriate role in X's case. So perhaps we should distinguish the disposition from its various bases, and identify it rather with the existential property. That way, "immunity" could indeed name a property shared by Walt and Milt. But this alternative has a disagreeable oddity of its own. The existential property, unlike the various bases, is too disjunctive and too extrinsic to occupy any causal role. There is no event that is essentially a having of the existential property, *a fortiori*, no such event ever causes anything. (Compare the absurd double-counting of causes that would ensue if we said, for instance, that when a match struck in the evening lights, one of the causes of the lighting is an event that essentially involves the property of being struck in the evening or twirled in the morning. I say there is no such event.) So if the disposition is the existential property, then it is causally impotent. On this theory, we are mistaken whenever we ascribe effects to dispositions.

Fortunately we needn't decide between the two theories. Though they differ on the analysis of disposition-names like "immunity," they agree about what entities there are. There is one genuine event—Walt's possession of antibodies. There is a truth about Walt to the effect that he has the existential property. But there is no second event that is essentially a having of the existential property, but is not essentially a having of it in any particular way. Whatever "Walt's immunity" may denote, it does not denote such an event. And since there is no such event at all, there is no such event to be non-causally explained.

IV GENERAL EXPLANATION

My main thesis concerns the explanation of particular events As it stands, it says nothing about what it is to explain general kinds of events However, it has a natural extension All the events of a given kind have their causal histories, and these histories may to some extent be alike Especially, the final parts of the histories may be much the same from one case to the next, however much the earlier parts may differ Then information may be provided about what is common to all the parallel causal histories—call it *general explanatory information* about events of the given kind To explain a kind of event is to provide some general explanatory information about events of that kind

Thus explaining why struck matches light in general is not so very different from explaining why some particular struck match lit In general, and in the particular case, the causal history involves friction, small hot spots, liberation of oxygen from a compound that decomposes when hot, local combustion of a heated inflammable substance facilitated by this extra oxygen, further heat produced by this combustion, and so on

There are intermediate degrees of generality If we are not prepared to say that every event of such-and-such kind, without exception, has a causal history with so-and-so features, we need not therefore abjure generality altogether and stick to explaining events one at a time We may generalize modestly, without laying claim to universality, and say just that quite often an event of such-and-such kind has a causal history with so-and-so features Or we may get a bit more ambitious and say that it is so in most cases, or at least in most cases that are likely to arise under the circumstances that prevail hereabouts Such modest generality may be especially characteristic of history and the social sciences, but it appears also in the physical sciences of complex systems, such as meteorology and geology We may be short of known laws to the effect that storms with feature X always do Y, or always have a certain definite probability of doing Y Presumably there are such laws, but they are too complicated to discover either directly or by derivation from first principles But we do have a great deal of general knowledge of the sorts of causal processes that commonly go on in storms

The pursuit of general explanations may be very much more widespread in science than the pursuit of general laws And not necessarily because we doubt that there are general laws to pursue Even if the scientific community unanimously believed in the existence of power-

ful general laws that govern all the causal processes of nature, and whether or not those laws were yet known, meteorologists and geologists and physiologists and historians and engineers and laymen would still want general knowledge about the sorts of causal processes that go on in the systems they study

V EXPLAINING WELL AND BADLY

An act of explaining may be more or less satisfactory, in several different ways It will be instructive to list them It will *not* be instructive to fuss about whether an unsatisfactory act of explaining, or an unsatisfactory chunk of explanatory information, deserves to be so-called, and I shall leave all such questions unsettled

1 An act of explaining may be unsatisfactory because the explanatory information provided is unsatisfactory In particular, it might be misinformation it might be a false proposition about the causal history of the explanandum This defect admits of degree False is false, but a false proposition may or may not be close to the truth [7] If it has a natural division into conjuncts, more or fewer of them may be true If it has some especially salient consequences, more or fewer of those may be true The world as it is may be more or less similar to the world as it would be if the falsehood were true

2 The explanatory information provided may be correct, but there may not be very much of it It might be a true but weak proposition, one that excludes few (with respect to some suitable measure) of the alternative possible ways the causal history of the explanandum might be Or the information provided might be both true and strong, but unduly disjunctive The alternative possibilities left open might be too widely scattered, too different from one another These defects too

[7] The analysis of verisimilitude has been much debated A good survey is Ilkka Niiniluoto, Truthlikeness Comments on Recent Discussion, *Synthese* 38 (1978) 281–329 Some plausible analyses have failed disastrously others conflict with one another One conclusion that emerges is that it is probably a bad move to try to define a single virtue of verisimilitude-cum-strength It s hard to say whether strength is a virtue in the case of false information especially if we have no uniquely natural way of splitting the misinformation into true and false parts Another conclusion is that even if this lumping together is avoided, verisimilitude still seems to consist of several distinguishable virtues

admit of degree Other things being equal, it is better if more correct explanatory information is provided, and it is better if that information is less disjunctive, up to the unattainable limit in which the whole explanation is provided and there is nothing true and relevant left to add

3 The explanatory information provided may be correct, but not thanks to the explainer He may have said what he did not know and had no very good reason to believe If so, the act of explaining is not fully satisfactory, even if the information provided happens to be satisfactory

4 The information provided, even if satisfactory in itself, may be stale news It may add little or nothing to the information the recipient possesses already

5 The information provided may not be of the sort the recipient most wants He may be especially interested in certain parts of the causal history, or in certain questions about its overall structure If so, no amount of explanatory information that addresses itself to the wrong questions will satisfy his wants, even if it is correct and strong and not already in his possession

6 Explanatory information may be provided in such a way that the recipient has difficulty in assimilating it, or in disentangling the sort of information he wants from all the rest He may be given more than he can handle, or he may be given it in a disorganized jumble [8] Or he may be given it in so unconvincing a way that he doesn't believe what he's told If he is hard to convince, just telling him may not be an effective way to provide him with information You may have to argue for what you tell him, so that he will have reason to believe you

7 The recipient may start out with some explanatory misinformation, and the explainer may fail to set him right

This list covers much that philosophers have said about the merits and demerits of explanations, or about what does and what doesn't deserve the name And yet I have not been talking specifically about explanation at all! What I have been saying applies just as well to acts of providing information about *any* large and complicated structure It might as well have been the rail and tram network of Melbourne rather than the causal history of some explanandum event The information provided, and the act of providing it, can be satisfactory or not in pre-

[8] As in the square peg example of Hilary Putnam, Philosophy and our Mental Life, in his *Mind Language and Reality* (Cambridge Cambridge University Press, 1975) pp 295–97

cisely the same ways There is no special subject pragmatics of explanation

Philosophers have proposed further *desiderata* A good explanation ought to show that the explanandum event had to happen, given the laws and the circumstances, or at least that it was highly probable, and could therefore have been expected if we had known enough ahead of time, or at least that it was less surprising than it may have seemed A good explanation ought to show that the causal processes at work are of familiar kinds, or that they are analogous to familiar processes, or that they are governed by simple and powerful laws, or that they are not too miscellaneous But I say that a good explanation ought to show none of these things unless they are true If one of these things is false in a given case, and if the recipient is interested in the question of whether it is true, or mistakenly thinks that it is true, then a good explanation ought to show that it is false But that is nothing special it falls under points 1, 5, and 7 of my list

It is as if someone thought that a good explanation of any current event had to be one that revealed the sinister doings of the CIA When the CIA really does play a part in the causal history, we would do well to tell him about it we thereby provide correct explanatory information about the part of the causal history that interests him most But in case the CIA had nothing to do with it, we ought not to tell him that it did Rather we ought to tell him that it didn't Telling him what he hopes to hear is not even a merit to be balanced off against the demerit of falsehood In itself it has no merit at all What does have merit is addressing the right question

This much is true We are, and we ought to be, biased in favor of believing hypotheses according to which what happens is probable, is governed by simple laws, and so forth That is relevant to the credibility of explanatory information But credibility is not a separate merit alongside truth, rather, it is what we go for in seeking truth as best we can

Another proposed *desideratum* is that a good explanation ought to produce understanding If understanding involves seeing the causal history of the explanandum as simple, familiar, or whatnot, I have already registered my objection But understanding why an event took place might, I think, just mean possession of explanatory information about it—the more of that you possess, the better you understand If so, of course a good explanation produces understanding It produces possession of that which it provides But this *desideratum*, so construed, is empty It adds nothing to our understanding of explanation

VI WHY-QUESTIONS, PLAIN AND CONTRASTIVE

A why-question, I said, is a request for explanatory information All questions are requests for information of some or other sort [9] But there is a distinction to be made Every question has a maximal true answer the whole truth about the subject matter on which information is requested, to which nothing could be added without irrelevancy or error In some cases it is feasible to provide these maximal answers Then we can reasonably hope for them, request them, and settle for nothing less "Who done it?—Professor Plum " There's no more to say

In other cases it isn't feasible to provide maximal true answers There's just too much true information of the requested sort to know or to tell Then we do not hope for maximal answers and do not request them, and we always settle for less The feasible answers do not divide sharply into complete and partial They're all partial, but some are more partial than others There's only a fuzzy line between enough and not enough of the requested information "What's going on here?"—No need to mention that you're digesting your dinner "Who is Bob Hawke?"—No need to write the definitive biography Less will be a perfectly good answer Why-questions, of course, are among the questions that inevitably get partial answers

When partial answers are the order of the day, questioners have their ways of indicating how much information they want, or what sort "In a word, what food do penguins eat?" "Why, in economic terms, is there no significant American socialist party?"

One way to indicate what sort of explanatory information is wanted is through the use of contrastive why-questions Sometimes there is an explicit "rather than " Then what is wanted is information about the causal historyof the explanandum event, not including information that would also have applied to the causal histories of alternative events, of the sorts indicated, if one of them had taken place instead In other words, information is requested about the difference between the actualized causal history of the explanandum and the unactualized causal histories of its unactualized alternatives Why did I visit Melbourne in 1979, rather than Oxford or Uppsala or Wellington? Because Monash University invited me That is part of the causal

[9] Except perhaps for questions that take imperative answers What do I do now, Boss?

history of my visiting Melbourne, and if I had gone to one of the other places instead, presumably that would not have been part of the causal history of my going there It would have been wrong to answer Because I like going to places with good friends, good philosophy, cool weather, nice scenery, and plenty of trains That liking is also part of the causal history of my visiting Melbourne, but it would equally have been part of the causal history of my visiting any of the other places, had I done so

The same effect can be achieved by means of contrastive stress Why did I *fly* to Brisbane when last I went there? I had my reasons for wanting to get there, but I won't mention those because they would have been part of the causal history no matter how I'd travelled Instead I'll say that I had too little time to go by train If I had gone by train, my having too little time could not have been part of the causal history of my so doing

If we distinguish plain from contrastive why-questions, we can escape a dilemma about explanation under indeterminism On the one hand, we seem quite prepared to offer explanations of chance events Those of us who think that chance is all-pervasive (as well as those who suspend judgment) are no less willing than the staunchest determinist to explain the events that chance to happen [10] On the other hand, we balk at the very idea of explaining why a chance event took place—for is it not the very essence of chance that one thing happens rather than another for no reason whatsoever? Are we of two minds?

No, I think we are right to explain chance events, yet we are right also to deny that we can ever explain why a chance process yields one outcome rather than another According to what I've already said, indeed we cannot explain why one happened *rather than the other* (That is so regardless of the respective probabilities of the two) The actual causal history of the actual chance outcome does not differ at all

[10] A treatment of explanation in daily life, or in history, dare not set aside the explanation of chance events as a peculiarity arising only in quantum physics If current scientific theory is to be trusted, chance events are far from exceptional The misguided hope that determinism might prevail in history if not in physics well deserves Railton's mockery All but the most basic regularities of the universe stand forever in peril of being interrupted or upset by intrusion of the effects of random processes The success of a social revolution might appear to be explained by its overwhelming popular support but this is to overlook the revolutionaries luck if all the naturally unstable nuclides on earth had commenced spontaneous nuclear fission in rapid succession the triumph of the people would never have come to pass (A Deductive Nomological Model of Probabilistic Explanation, pp 223–24) On the same point, see my Postscript B to A Subjectivist s Guide to Objective Chance, in this volume

from the unactualized causal history that the other outcome would have had, if that outcome had happened A contrastive why-question with "rather" requests information about the features that differentiate the actual causal history from its counterfactual alternative There are no such features, so the question can have no positive answer Thus we are right to call chance events inexplicable, if it is contrastive explanation that we have in mind (Likewise, we can never explain why a chance event *had* to happen, because it didn't have to) But take away the "rather" (and the "had") and explanation becomes possible Even a chance event has a causal history There is information about that causal history to be provided in answer to a plain why-question And thus we are right to proceed as we all do in explaining what we take to be chance events

VII THE COVERING-LAW MODEL

The covering-law model of explanation has long been the leading approach As developed in the work of Hempel and others, it is an elegant and powerful theory How much of it is compatible with what I have said?

Proponents of the covering-law model do not give a central place to the thesis that we explain by providing information about causes But neither do they say much against it They may complain that the ordinary notion of causation has resisted precise analysis, they may say that mere mention of a cause provides less in the way of explanation than might be wished, they may insist that there are a few special cases in which we have good non-causal explanations of particular occurrences But when they give us their intended examples of covering-law explanation, they almost always pick examples in which—as they willingly agree—the covering-law explanation does include a list of joint causes of the explanandum event, and thereby provides information about its causal history

The foremost version of the covering-law model is Hempel's treatment of explanation in the non-probabilistic case [11] He proposes that an explanation of a particular event consists, ideally, of a correct deductive-nomological (henceforth D–N) argument There are law premises and particular-fact premises and no others The conclusion

[11] For a full presentation of Hempel s views, see the title essay in his *Aspects of Scientific Explanation*

says that the explanandum event took place The argument is valid, in the sense that the premises could not all be true and the conclusion false (We might instead define validity in syntactic terms If so, we should be prepared to included mathematical, and perhaps definitional, truths among the premises) No premise could be deleted without destroying the validity of the argument The premises are all true

Hempel also offers a treatment for the probabilistic case, but it differs significantly from his deductive-nomological model, and also it has two unwelcome consequences (1) An improbable event cannot be explained at all (2) One requirement for a correct explanation—"maximal specificity"—is relative to our state of knowledge, so that our ignorance can make correct an explanation that would be incorrect if we knew more Surely what's true is rather that ignorance can make an explanation seem to be correct when really it is not Therefore, instead of Hempel's treatment of the probabilistic case, I prefer to consider Railton's "deductive-nomological model of probabilistic explanation" [12] This closely parallels Hempel's D–N model for the non-probabilistic case, and it avoids both the difficulties just mentioned Admittedly, Railton's treatment is available only if we are prepared to speak of chances—single-case objective probabilities But that is no price at all if we have to pay it anyway And we do, if we want to respect the apparent content of science (Which is not the same as

[12] See Railton s paper of the same name In what follows I shall simplify Railton s position in two respects (1) I shall ignore his division of a D-N argument for a probabilistic conclusion into two parts, the first deriving a law of uniform chances from some broader theory and the second applying that law to the case at hand (2) I shall pretend until further notice, that Railton differs from Hempel only in his treatment of probabilistic explanation in fact there are other important differences, to be noted shortly

It is important to distinguish Railton s proposal from a different way of using single-case chances in a covering-law model of explanation, proposed in James H Fetzer, A Single Case Propensity Theory of Explanation, *Synthese* 28 (1974) pp 171–98 For Fetzer, as for Railton, the covering laws are universal generalizations about single case chances But for Fetzer, as for Hempel, the explanatory argument without any addendum, is the whole of the explanation it is inductive, not deductive, and its conclusion says outright that the explanandum took place, not that it had a certain chance This theory shares some of the merits of Railton s However, it has one quite peculiar consequence For Fetzer, as for Hempel, an explanation is an argument, however a good explanation is not necessarily a good argument Fetzer, like Railton, wants to have explanations even when the explanandum is extremely improbable But in that case a good explanation is an extremely bad argument It is an inductive argument whose premises not only fail to give us any good reason to believe the conclusion but in fact give us very good reason to *dis*believe the conclusion

respecting the positivist philosophy popular among scientists) Frequencies—finite or limiting, actual or counterfactual—are fine things in their own right So are degrees of rational belief But they just do not fit our ordinary conception of objective chance, as exemplified when we say that any radon-222 atom at any moment has a 50% chance of decaying within the next 3 825 days If chances are good enough for theorists of radioactive decay, they are good enough for philosophers of science

Railton proposes that an explanation of a particular chance event consists, ideally, of two parts The first part is a D–N argument, satisfying the same constraints that we would impose in the nonprobabilistic case, to a conclusion that the explanandum event had a certain specified chance of taking place The chance can be anything very high, middling, or even very low The D–N argument will have probabilistic laws among its premises—preferably, laws drawn from some powerful and general theory—and these laws will take the form of universal generalizations concerning single-case chances The second part of the explanation is an addendum—not part of the argument—which says that the event did in fact take place The explanation is correct if both parts are correct if the premises of the D–N argument are all true, and the addendum also is true

Suppose we have a D–N argument, either to the explanandum event itself or to the conclusion that it has a certain chance And suppose that each of the particular-fact premises says, of a certain particular event, that it took place Then those events are jointly sufficient, given the laws cited, for the event or for the chance In a sense, they are a minimal jointly sufficient set, but a proper subset might suffice given a different selection of true law premises, and also it might be possible to carve off parts of the events and get a set of the remnants that is still sufficient under the original laws To perform an act of explaining by producing such an argument and committing oneself to its correctness is, in effect, to make two claims (1) that certain events are jointly sufficient, under the prevailing laws, for the explanandum event or for a certain chance of it, and (2) that only certain of the laws are needed to establish that sufficiency

It would make for reconciliation between my account and the covering-law model if we had a covering-law model of causation to go with our covering-law model of explanation Then we could rest assured that the jointly sufficient set presented in a D–N argument was a set of causes of the explanandum event Unfortunately, that assurance is not to be had Often, a member of the jointly sufficient set pre-

sented in a D–N argument will indeed be one of the causes of the explanandum event But it may not be The counterexamples are well known, I need only list them

1 An irrelevant non-cause might belong to a non-minimal jointly sufficient set Requiring minimality is not an adequate remedy, we can get an artificial minimality by gratuitously citing weak laws and leaving stronger relevant laws uncited That is the lesson of Salmon's famous example of the man who escapes pregnancy because he takes birth control pills, where the only cited law says that nobody who takes the pills becomes pregnant, and hence the premise that the man takes pills cannot be left out without spoiling the validity of the argument [13]

2 A member of a jointly sufficient set may be something other than an event For instance, a particular-fact premise might say that something has a highly extrinsic or disjunctive property I claim that such a premise cannot specify a genuine event, see "Events," in this volume

3 An effect might belong to a set jointly sufficient for its cause, as when there are laws saying that a certain kind of effect can be produced in only one way That set might be in some appropriate sense minimal, and might be a set of events That would not suffice to make the effect be a cause of its cause

4 Such an effect might also belong to a set jointly sufficient for another effect, perhaps a later effect, of the same cause Suppose that, given the laws and circumstances, the appearance of a beer ad on my television could only have been caused by a broadcast which would also cause a beer ad to appear on your television Then the first appearance may be a member of a jointly sufficient set for the second, still, these are not cause and effect Rather they are two effects of a common cause

5 A preempted potential cause might belong to a set jointly sufficient for the effect it would have caused, since there might be nothing that could have stopped it from causing that effect without itself causing the same effect

In view of these examples, we must conclude that the jointly sufficient set presented in a D–N argument may or may not be a set of causes We do not, at least not yet, have a D–N analysis of causation All the same, a D–N argument may present causes If it does, or rather

[13] See Wesley C Salmon et al, *Statistical Explanation and Statistical Relevance* (Pittsburgh University of Pittsburgh Press, 1971), p 34

if it appears to the explainer and audience that it does, then on my view it ought to look explanatory That is the typical case with sample D–N arguments produced by advocates of the covering-law model

If the D–N argument does not appear to present causes, and it looks explanatory anyway, that is a problem for me In Section III, I discussed three such problem cases, the alleged non-causal explanations there considered could readily have been cast as D–N arguments, and indeed I took them from Hempel's and Railton's writings on covering-law explanation In some cases, I concluded that information was after all given about how the explanandum was caused, even if it happened in a more roundabout way than by straightforward presentation of causes In other cases, I concluded that what was explained was not really a particular event Either way, I'm in the clear

If the D–N argument does not appear to present causes, and therefore fails to look explanatory, that is a problem for the covering-law theorist He might just insist that it *ought* to look explanatory, and that our customary standards of explanation need reform To the extent that he takes this high-handed line, I lose interest in trying to agree with as much of his theory as I can But a more likely response is to impose constraints designed to disqualify the offending D–N arguments Most simply, he might say that an explanation is a D–N argument of the sort that does present a set of causes, or that provides information in some more roundabout way about how the explanandum was caused Or he might seek some other constraint to the same effect, thereby continuing the pursuit of a D–N analysis of causation itself Railton is one covering-law theorist who acknowledges that not just any correct D–N argument (or probabilistic D–N argument with addendum) is explanatory, further constraints are needed to single out the ones that are In sketching these further constraints, he does not avoid speaking in causal terms (He has no reason to, since he is not attempting an analysis of causation itself) For instance, he distinguishes D–N arguments that provide an "account of the mechanism" that leads up to the explanandum event, by which he means, I take it, that there ought to be some tracing of causal chains He does not make this an inescapable requirement, however, because he thinks that not all covering-law explanation is causal [14]

A D–N argument may explain by presenting causes, or otherwise giving information about the causal history of the explanandum, is it

[14] See his *Explaining Explanation* A Deductive Nomological Model of Probabilistic Explanation, and Probability, Explanation and Information

also true that any causal history can be characterized completely by means of the information that can be built into D–N arguments? That would be so if every cause of an event belongs to some set of causes that are jointly sufficient for it, given the laws, or, in the probabilistic case, that are jointly sufficient under the laws for some definite chance of it Is it so that causes fall into jointly sufficient sets of one or the other sort? That does not follow, so far as I can tell, from the counterfactual analysis of causation that I favor It may nevertheless be true, at least in a world governed by a sufficiently powerful system of (strict or probabilistic) laws, and this may be such a world If it is true, then the whole of a causal history could in principle be mapped by means of D–N arguments (with addenda in the probabilistic case) of the explanatory sort

In short, if explanatory information is information about causal histories, as I say it is, then one way to provide it is by means of D–N arguments Moreover, under the hypothesis just advanced, there is no explanatory information that could not in principle be provided in that way To that extent the covering-law model is dead right

But even when we acknowledge the need to distinguish explanatory D–N arguments from others, perhaps by means of explicitly causal constraints, there is something else wrong It is this The D–N argument—correct, explanatory, and fully explicit—is represented as the ideal serving of explanatory information It is the right shape and the right size It is enough, anything less is not enough, and anything more is more than enough

Nobody thinks that real-life explainers commonly serve up full D–N arguments which they hope are correct We very seldom do And we seldom could—it's not just that we save our breath by leaving out the obvious parts We don't know enough Just try it Choose some event you think you understand pretty well, and produce a fully explicit D–N argument, one that you can be moderately sure is correct and not just almost correct, that provides some non-trivial explanatory information about it Consult any science book you like Usually the most we can do, given our limited knowledge, is to make existential claims [15] We can venture to claim that there exists some (correct, etc)

[15] In *Foundations of Historical Knowledge*, Chapter III, Morton White suggests that because -statements should be seen as existential claims You assert the existence of an explanatory argument which includes a given premise, even though you may be unable to produce the argument This is certainly a step in the right direction However it seems to underestimate the variety of existential statements that might be made, and also it incorporates a suspect D-N analysis of causation

D–N argument for the explanandum that goes more or less like this, or that includes this among its premises, or that draws its premises from this scientific theory, or that derives its conclusion from its premise with the aid of this bit of mathematics, or I would commend these existential statements as explanatory, to the extent—and only to the extent—that they do a good job of giving information about the causal history of the explanandum But if a proper explanation is a complete and correct D–N argument (perhaps plus addendum), then these existential statements are not yet proper explanations Just in virtue of their form, they fail to meet the standard of how much information is enough

Hempel writes "To the extent that a statement of individual causation leaves the relevant antecedent conditions, and thus also the requisite explanatory laws, indefinite it is like a note saying that there is a treasure hidden somewhere "[16] The note will help you find the treasure provided you go on working, but so long as you have only the note you have no treasure at all, and if you find the treasure you will find it all at once I say it is not like that A shipwreck has spread the treasure over the bottom of the sea and you will never find it all Every dubloon you find is one more dubloon in your pocket, and also it is a clue to where the next dubloons may be You may or may not want to look for them, depending on how many you have so far and on how much you want to be how rich

If you have anything less than a full D–N argument, there is more to be found out Your explanatory information is only partial Yes *And so is any serving of explanatory information we will ever get*, even if it consists of ever so many perfect D–N arguments piled one upon the other There is always more to know A D–N argument presents only one small part—a cross section, so to speak—of the causal history There are very many other causes of the explanandum that are left out Those might be the ones we especially want to know about We might want to know about causes earlier than those presented Or we might want to know about causes intermediate between those presented and the explanandum We might want to learn the mechanisms involved by tracing particular causal chains in some detail (The premises of a D–N argument might tell us that the explanandum would come about through one or the other of two very different causal chains, but not tell us which one) A D–N argument might give us far from enough explanatory information, considering what sort of information we

[16] *Aspects of Scientific Explanation*, p 349

want and what we possess already On the other hand, it might give us too much Or it might be the wrong shape, and give us not enough and too much at the same time, for it might give us explanatory information of a sort we do not especially want The cross-section it presents might tell us a lot about the side of the causal history we're content to take for granted, and nothing but stale news about the side we urgently want to know more about

Is a (correct, etc) D–N argument in *any* sense a complete serving of explanatory information? Yes in this sense, and this sense alone it completes a jointly sufficient set of causes (And other servings complete seventeen-membered sets, still others complete sets going back to the nineteenth century) The completeness of the jointly sufficient set has nothing to do with the sort of enoughness that we pursue There is nothing ideal about it, in general Other shapes and sizes of partial servings may be very much better—and perhaps also better within our reach

It is not that I have some different idea about what is the unit of explanation We should not demand a unit, and that demand has distorted the subject badly It's not that explanations are things we may or may not have one of, rather, explanation is something we may have more or less of

One bad effect of an unsuitable standard of enoughness is that it may foster disrespect for the explanatory knowledge of our forefathers Suppose, as may be true, that seldom or never did they get the laws quite right Then seldom or never did they possess complete and correct D–N arguments Did they therefore lack explanatory knowledge? Did they have only some notes, and not yet any of the treasure? Surely not! And the reason, say I, is that whatever they may not have known about the laws, they knew a lot about how things were caused

But once again, the covering-law model needn't have the drawback of which I have been complaining, and once again it is Railton who has proposed the remedy [17] His picture is similar to mine Associated with each explanandum we have a vast and complicated structure, explanatory information is information about this structure, an act of explaining is an act of conveying some of this information, more or less information may be conveyed, and in general the act of explaining may be more or less satisfactory in whatever ways any act of conveying information about a large and complicated structure may be more or

[17] See *Explaining Explanation* and Probability, Explanation, and Information

less satisfactory The only difference is that whereas for me the vast structure consists of events connected by causal dependence, for Railton it is an enormous "ideal text" consisting of D–N arguments—correct, satisfying whatever constraints need be imposed to make them explanatory, and with addenda as needed—strung together They fit together like proofs in a mathematics text, with the conclusion of one feeding in as a premise to another, and in the end we reach arguments to the occurrence, or at least a chance, of the explanandum itself It is unobjectionable to let the subject matter come in units of one argument each, so long as the activity of giving information about it needn't be broken artificially into corresponding units

By now, little is left in dispute Both sides agree that explaining is a matter of giving information, and no standard unit need be completed The covering-law theorist has abandoned any commitment he may once have had to a D–N analysis of causation, he agrees that not just any correct D–N argument is explanatory, he goes some distance toward agreeing that the explanatory ones give information about how the explanandum is caused, and he does not claim that we normally, or even ideally, explain by producing arguments For my part, I agree that one way to explain would be to produce explanatory D–N arguments, and further, that an explainer may have to argue for what he says in order to be believed Explanation as argument versus explanation as information is a spurious contrast More important, I would never deny the relevance of laws to causation, and therefore to explanation, for when we ask what would have happened in the absence of a supposed cause, a first thing to say is that the world would then have evolved lawfully The covering-law theorist is committed, as I am not, to the thesis that all explanatory information can be incorporated into D–N arguments, however, I do not deny it, at least not for a world like ours with a powerful system of laws I am committed, as he is not, to the thesis that all explaining of particular events gives some or other sort of information about how they are caused, but when we see how many varieties of causal information there are, and how indirect they can get, perhaps this disagreement too will seem much diminished

One disagreement remains, central but elusive It can be agreed that information about the prevailing laws is at least highly relevant to causal information, and *vice versa*, so that the pursuit of explanation and the investigation of laws are inseparable in practice But still we can ask whether information about the covering laws is itself *part* of explanatory information The covering law theorist says yes, I say no But this looks like a question that would be impossible to settle, given

that there is no practical prospect of seeking or gaining information about causes without information about laws, or information about laws without information about causes We can ask whether the work of explaining would be done if we knew all the causes and none of the laws We can ask, but there is little point trying to answer, since intuitive judgments about such preposterous situations needn't command respect

TWENTY THREE

Events*

I INTRODUCTION

Events are not much of a topic in their own right They earn their keep in the discussion of other topics sometimes the semantics of nominalisations and adverbial modification, sometimes the analysis of causation and causal explanation There is no guarantee that events made for semantics are the same as the events that are causes and effects It seems unlikely, in some cases at least A certain mathematical sequence converges There is some entity or other that we may call the converging of the sequence The sequence converges rapidly iff, in some sense, this entity is rapid I have no objection to that, but I insist that the converging of the sequence, whatever it may be, is nothing like any event that causes or is caused (The so-called "events" of probability theory are something else again—propositions, or properties of things at times) My present interest is in events as causes and effects Therefore I shall not follow the popular strategy of approaching events by way of nominalisations Events made in the image of nominalisations are right for some purposes, but not for mine When I introduce nominalisations to denote events, as I shall, it will not be analysis of natural language but mere stipulative definition

* I am much indebted to discussions with Jonathan Bennett, Alison McIntyre, and Mark Johnston

In the two previous papers, I put forward several theses about causation and explanation. (1) Causal dependence is counterfactual dependence between distinct events. Event *e* depends causally on the distinct event *c* iff, if *c* had not occured, *e* would not have occurred—or at any rate, *e*'s chance of occurring would have been very much less than it actually was. (We must take care to use the right kind of counterfactuals: no backtrackers. See "Counterfactual Dependence and Time's Arrow," in this volume.) (2) Causation is the ancestral of causal dependence: event *c* causes event *e* iff either *e* depends on *c*, or *e* depends on an intermediate event *d* which in turn depends on *c*, or.... Causation without direct causal dependence is exceptional, but it occurs in cases of causal preemption. (See Postscript E to "Causation," in this volume.) (3) Any event has a causal history: a vast branching structure consisting of that event and all the events which cause it, together with all the relations of causal dependence among these events. (4) To explain why an event occurs is to give information about its causal history. Such information is inevitably partial. An explanation may specify part of the causal history of the explanandum event, or it may just provide structural information of one or another sort about the causal history. Goodness of explanation is governed by the pragmatic standards that apply to information-giving generally.

Since these four theses concern causation among events, their meaning cannot be entirely clear until I provide a theory of events to go with them. Not just any theory will do. If a theory posits too many distinct events, then many instances of counterfactual dependence between its allegedly distinct alleged events will clearly not be causal.[1] This difficulty will arise, for instance, on a theory that posits an abundance of distinct events to match the abundance of nonequivalent predicates in nominalisations. A theory that allows unlimited Boolean combination of events also will generate alleged events that enter into relations of counterfactual dependence, but that do not seem intuitively suited to cause or to be caused. On the other hand, a sparse theory may posit too few events, forcing us to go beyond the events it countenances in order to complete our causal histories. This difficulty will arise, for instance, on a theory that limits itself to events falling under event-nouns of ordinary language: flashes, bangs, thumps, bumps, lectures, kisses, battles.... More generally, it will arise on a theory that provides no events to fill those stretches of time that we call "uneventful."

[1] As is shown by Jaegwon Kim in his "Causes and Counterfactuals," *Journal of Philosophy* 70 (1973): 570–72; and "Noncausal Connections," *Nous* 8 (1974): 41–52.

A theory that gives events unduly rich and fragile essences also will make trouble, as we shall see

In this paper I shall consider what sort of theory of events I need to go with my theses about causation If none could be found, that would be reason to reject what I say about causation But I think a suitable theory can be found—or at least sketched—and I think it is a reasonably attractive theory in its own right What other purposes it might serve, if any, I cannot say

II EVENTS ARE PROPERTIES OF SPATIOTEMPORAL REGIONS

An event is a localised matter of contingent fact It occurs It is contingent that it occurs, no event occurs at every possible world Hence we have nonvacuous counterfactuals about what would have been the case if a given event had not occurred, as we must if we are to place that event in a history of causal dependence An event occurs in a particular spatiotemporal region Its region might be small or large, there are collisions of point particles and there are condensations of galaxies, but even the latter occupy regions small by astronomical standards

(Perhaps not just any region is a region in which an event can occur A smallish, connected, convex region may seem a more likely candidate than a widely scattered part of spacetime But I leave this question unsettled, for lack of clear test cases If all of this year's VFL football comprises one big event, that event occurs in a scattered region, bits of it occur in various parks on various afternoons But does it all comprise one event? Intuition is silent, and, so far as I can tell, the needs of my account of causation could be met either way)

An event occurs in exactly one region of the world, if it occurs at all If an event occurs in a region, it does not occur in any proper part of that region The whole of the event occupies the whole of its region Parts of it, but not the whole of it, may occur in parts of its region Also, an event is unrepeated it does not occur in two different regions of the world

Thus an "annual event" such as the Grand Final is not an event in the sense of the present theory As is only right and proper the Grand Final does not cause or get caused *simpliciter* It has different causes and effects in different years, which is to say that the different, unrepeated Grand Finals of the different years are what really have the causes and effects

I distinguish occurring *in* a region and *within* a region An event occurs within every region that includes the region in which it occurs, and it occurs in the region that is the intersection of all regions within which it occurs We might also say that it *is occurring in* every region that is part of the region in which it occurs, and that it occurs in the region that is the mereological sum of all regions in which it is occurring

To any event there corresponds a property of regions the property that belongs to all and only those spatiotemporal regions, of this or any other possible world, in which that event occurs Such a property belongs to exactly one region of any world where the event occurs, and there are some such worlds It belongs to no region of any world where the event does not occur, and there are some of those worlds also If a property of regions satisfies the conditions just stated, it may or may not correspond to an event But at least it is *formally eligible* to do so

By a *property* I mean simply a class—any class [2] To have the property is to belong to the class *All* the things that have the property, whether actual or merely possible, belong My point in using the word "property" is simply to emphasise that we are dealing with a class that may have otherworldly things, unactualised *possibilia*, among its members (It might even, in the case of mathematical properties such as oddness, have *un*worldly members that are not part of any world) The property that corresponds to an event, then, is the class of all regions—at most one per world—where that event occurs

[2] A property is *not* a universal, and properties are no substitute for universals It also seems fairly unlikely that universals would do as a substitute for properties The existence of properties, in my sense of the word, and the existence of universals are independent questions I am committed to properties, neutral on universals If there are universals, they are sparse the question which of them exist is an important scientific issue they are wholly present whenever present at all, their sharing makes for resemblance, and things that have exactly the same ones are perfect qualitative duplicates None of the above is true for properties Properties are abundant, numbering at least beth 3 for properties of individuals alone, they are suited to serve as semantic values of arbitrarily complex predicates and gerunds, and as values of second and higher order variables, they are shared in equal multiplicity in cases of perfect duplication and in cases of utter dissimilarity If there are universals they make certain properties special—namely, those that belong to exactly the things that share a universal If not, it remains true that certain properties are special, but their specialness must be otherwise explained or left as primitive See D M Armstrong, *Universals and Scientific Realism* (Cambridge Cambridge University Press, 1978), and my New Work for a Theory of Universals *Australasian Journal of Philosophy* 61 (1983) 343–77

Two events can occur in exactly the same region An electron's presence in a field can cause its acceleration, radiation of two frequencies can reverberate throughout the same cavity, two chemical reactions can go on in the same flask More fancifully, there might be goblins made of a sort of matter that passes through our sort without any interaction, and a battle of goblins might occur in the very same region as this conference (the 1981 Conference of the Australasian Association of Philosophy, where this paper was first read)

But in each case it would have been possible for one of the two events to occur without the other It would have been contrary to law, in some cases, but I take it that the laws of nature themselves are contingent However, I can think of no plausible case of two events such that, necessarily, for any region, one occurs in that region iff the other does Two such inseparables would be causally indistinguishable on a counterfactual analysis of causation, so it is hard to see how my treatment of causation could possibly need them both I shall therefore take it that for any two events there is some region of some world where one occurs and the other does not That region has the property corresponding to one event It lacks the property corresponding to the other So the two events correspond to two different properties Our correspondence between events and properties of regions is therefore one-to-one

A one-to-one correspondence is an opportunity for reduction, and I see no reason why events are needed as irreducible elements of being Therefore, I propose to identify events with their corresponding properties An event is a property, or in other words a class, of spatio-temporal regions It satisfies my conditions of formal eligibility by containing one region each from some worlds, none from others, and never more than one from the same world It occurs if and where and when there is a region that is a member of it

Not just any property meeting the conditions given is to count as an event I have said what kind of things events are—namely, formally eligible properties of regions—but not which things of that kind are events The latter parts of the paper will address that question, though all I say will still fall sadly short of a precise necessary and sufficient condition for eventhood

I am relying on the assumptions (1) that regions are individuals which are parts of possible worlds, and (2) that no region is part of two different worlds Those assumptions are controversial, but I need not defend them here Those who doubt them need only retreat to a more complicated set-theoretic construction of properties—as functions that

assign to each world the set of things that have that property at that world—whereupon what I say will go through essentially unchanged

My proposal resembles that of Richard Montague, on which events are taken as certain properties of times [3] The event occurs at a certain time at a certain world iff the property which is that event belongs, at that world, to that time Thus he identifies the event with the property of being a time when that event occurs I think my proposal has two minor advantages (1) In view of Relativity, it is not altogether clear what sort of thing a time is (2) Given that a Montague event occurs at a certain time at a certain world, we must work to recover the place where it occurs, given that one of my events occurs in a region of a world, its place is given immediately

My proposal also resembles the suggestion that events might simply be identified with regions, or perhaps with regions including all that occupies them That has been suggested by several authors, usually with some acknowledgement that what they suggest does not conform to normal usage [4] We do usually think that two different events might occur in the very same region, not so, of course, if we identify events with their regions If this conference is its region, and a battle of goblins is its region, and the conference and the battle are the same region, then the battle and the conference are a single event You might like the idea of treating events as regions, and yet you might insist on distinguishing these two events If so, you might want to say that one event is the region *qua* conference venue, the other is the very same region *qua* battlefield And you might want to say that they are identical, yet they are to be distinguished In a similar frame of mind, you might want to say that Russell *qua* philosopher and Russell *qua* politician are identical, yet they are to be distinguished You really should

[3] On the Nature of Certain Philosophical Entities, *Monist* 53 (1969) 159–94 reprinted in Montague, *Formal Philosophy* (New Haven Yale University Press, 1974) See also the treatment of events briefly stated in M J Cresswell, *Logics and Languages* (London Methuen, 1973), p 95, Cresswell s treatment resembles mine even more closely than Montague s does, but differs in taking space-time points to be identical across worlds

[4] E J Lemmon, Comments on Davidson, The Logical Form of Action Sentences , in *The Logic of Decision and Action* ed by Nicholas Rescher (Pittsburgh University of Pittsburgh Press 1967), W V Quine, *Philosophy of Logic* (Englewood Cliffs, New Jersey Prentice Hall, 1970) pp 31–32, Quine, *Roots of Reference* (La Salle, Illinois Open Court 1974), pp 5 and 131–32, Quine Things and Their Place in Theories, in his *Theories and Things* (Cambridge, Massachusetts Harvard University Press, 1981), and J J C Smart, Further Thoughts on the Identity Theory, *Monist* 56 (1972) 149–62

not say such things nothing may be distinguished in any way whatsoever from itself Yet we may sympathise with your inclination, and provide for it legitimately as follows Russell *qua* philosopher is Russell-taken-in-intension the Russell of our world, taken together with the Russells of other worlds who are his philosophical counterparts Likewise for Russell *qua* politician, except that instead of the philosophical counterparts we take the political counterparts So far as this world is concerned, there is no difference between the two Russells-in-intension, their thisworldly members really are identical But the otherworldly members differ, since our Russell has many philosophical-but-not-political and political-but-not-philosophical counterparts So the two Russells-in-intension, taken entire as classes spread over many worlds, really do differ Likewise for a pair of regions-in-intension such as this region *qua* conference venue and the same region *qua* battlefield their thisworldly members are identical, but they differ by having different otherworldly members My events are exactly such regions-in-intension, consisting of regions of many worlds united by suitable relations of similarity

If events are properties, understood as classes with members from many worlds, then you might wish to say that an event exists whether or not it occurs like a number, it exists necessarily, from the standpoint of every world, though it is not part of any world Or you might instead say that an event exists at just those worlds where it occurs, or even that it does not fully exist at any world, since no world holds all of it I think these are merely verbal questions We may as well leave them unsettled, nothing hangs on them Never mind whether it is contingent that an event *exists*, it is at any rate contingent that an event *occurs* Also, I do not think it would be appropriate to deny the name "event" to those events that do not occur in our own world After all, they are of a kind with the events that do actually occur Others may prefer a different terminology, on which "events" are the ones that actually occur, and the rest are called something else It matters not—the terminologies are intertranslatable

III EVENTS ARE DESCRIBED ESSENTIALLY AND ACCIDENTALLY

Events have their essences built in, in the form of necessary conditions for their occurrence We may classify events by their essences, stating conditions that a region must satisfy if that event is to occur there For

instance, an event is essentially a change iff, necessarily, that event occurs in a region only if something changes throughout that region [5] Likewise an event is essentially a death iff, necessarily, that event occurs in a region only if someone dies throughout that region, and not throughout any larger one Such a region will be the location of a final temporal segment of the victim, beginning just when he starts to die (The vagueness of just when he starts to die infects our classification of events with vagueness, but there's no harm in that) And so on for a wide range of essential classifications These will include some made from single verbs, and others made from more complex predicate phrases—even infinitely complex ones Thus we say what it would mean (whether or not it is ever true) that some event essentially is a vibrating-of-a-steel-gong-of-so-and-so-size-and-shape-at-so-and-so-frequency

(When I use predicate phrases to define essential classifications of events, I am not making any claim of "conceptual priority," whatever that might mean If the predicates in turn are definable in terms of the classification of events, we have nothing worse than a circle of interdefinables Such circles do not suffice to eliminate all the interdefined terms at once, of course, but they may nevertheless be useful collections of analytic truths)

We can also say what it would mean (whether or not it is ever true) for an event to essentially involve Socrates It does so iff, necessarily, it occurs in a region only if Socrates is present there—either the Socrates of our world, or else some otherworldly Socrates who is a counterpart of ours [6] (The vagueness of the counterpart relation infects the classification with vagueness, but again we needn't mind) Likewise, combining classification by predicates and by individuals, an event would be essentially a death of Socrates iff, necessarily, it occurs in a region only if Socrates dies throughout that region Perhaps we should say (for reasons to be considered later) that the individuals essentially involved in events are not whole persisting people but temporal segments of them But if so, that doesn't change the definition of involvement for instance an event essentially involves a certain temporal segment of

[5] Compare the definition of a change by essence and accident in Michael A Slote, *Metaphysics and Essence* (Oxford Blackwell, 1974), p 16 I borrow Slote s method, but I apply it less ambitiously since I have specified the category of the definiendum beforehand

[6] See Counterpart Theory and Quantified Modal Logic, in my *Philosophical Papers*, Volume I

Socrates iff, necessarily, that event occurs in a region only if that segment, or a counterpart of it, is present there

We can also say what it would mean for an event to have its location, or a partial specification of its location, essentially It essentially occurs in region *R* iff, necessarily, it occurs in a region only if that region is *R*, or a counterpart of *R* It essentially occurs within the twentieth century iff, necessarily, it occurs only within the twentieth century, or a counterpart thereof

Not only does an event have built-in necessary conditions for its occurrence, it has built-in necessary *and sufficient* conditions That is just to say that there is a property that belongs to all *and only* the regions of this and other worlds where it occurs, and that is just to say that these regions comprise a class If we could manage to express that property, and thus state necessary and sufficient conditions for the occurrence of an event, then not only could we classify that event by its essence, we could specify it uniquely It would be the unique event such that, necessarily, it occurs in a region *R* iff *R*

You might hope that an essential specification of an event could easily be extracted from the sort of nominalisation whereby we standardly denote it Suppose we denote an event by a nominalisation "the *F*-ing of *A* at *T*" Let *f* be the property expressed by the predicate *F*, let *a* be the individual denoted by *A*, and let *t* be the time denoted by *T* (The denoting needn't be rigid) The nominalisation denotes an event by way of the "constitutive" triple of *f*, *a*, and *t*, further, the occurrence of that event is somehow connected with the fact that property *f* belongs to individual *a* at time *t*[7] (How does a property belong to an individual at a time? Perhaps because it is really a property of time-slices, or perhaps it is really a relation of individuals to times) Then it is all too easy to assume that the triple gives us an essential specification of the denoted event That is, we have the hypothesis that "the *F*-ing of *A* at *T*" denotes the event such that, necessarily, it occurs iff *f* belongs to *a* at *t* (Iff so, then presumably it occurs in the region occupied by *a* at *t*)

I think this will not do, at least not given the needs of a counterfactual analysis of causation Sometimes, perhaps, an event can indeed be essentially specified in this way by means of a constitutive property,

[7] On the specifying of events by way of such triples, see Jaegwon Kim, Causation, Nomic Subsumption, and the Concept of Event, *Journal of Philosophy* 70 (1973) 217–36 Kim is not committed to the view that such specifications are essential, despite the suggestion conveyed by his term constitutive

individual, and time But it is not so in general for the events we denote by nominalisations, and it is not so in general for causes and effects The trouble is that an event with such a rich essence is a fragile thing It is hard to change it without destroying it It cannot occur at any but its constitutive time, it cannot involve any but its constitutive individual, and it cannot occur without something being an instance of its constitutive property The causes and effects whereof we ordinarily speak are more robust than that

The clearest difficulty concerns the alleged constitutive time It is one thing to postpone an event, another to cancel it A cause without which it would have occurred later, or sooner, is not a cause without which it would not have occurred at all Who would dare be a doctor, if the hypothesis under consideration were right? You might manage to keep your patient alive until 4 12, when otherwise he would have died at 4 08 You would then have caused his death For his death was, in fact, his death at 4 12 If that time is essential, his death is an event that would not have occurred had he died at 4 08, as he would have done without your action That will not do (The point is due to Ken Kress For further discussion, see Postscript E to "Causation," in this volume)

Of course, we should not bounce off to the other extreme, and suppose that the death of the patient is an event such that, necessarily, it occurs iff he dies, never mind when and how That would mean that the only way to cause someone's death would be to rob him of immortality, which is quite wrong also Someone could die any of many different deaths, but not just any difference of time is enough to make the difference

The alleged constitutive individual also is problematic It is by no means clear that an event involving an individual always involves that individual essentially Sometimes we are entitled to think of individuals as interchangeable parts One member only of the firing squad got live ammunition, the rest fired blanks The shooting was in fact done by Ted, but it could very well have been done by Ned instead A cause without which someone else would have fired the fatal shot is not, or at any rate not clearly, a cause without which that very shooting—the one that was in fact a shooting by Ted—would not have taken place It is not something on which that event depends

Even the alleged constitutive property is not beyond suspicion Perhaps any change, or any death, or any shooting, is such essentially Perhaps not But what if some much more specific, detailed predicate appears in the nominalisation? Sebastian strolled because he had plenty

of time Had he been delayed, the walking that was in fact a strolling might rather have been a striding It might not even have been a walking, but rather a running That is not to say, not clearly, that it would not have occurred at all

So, while it is clear enough what it would mean to specify events essentially, often that does not seem to be what we really do At any rate, it is not what we do when we specify events by means of our standard nominalisations

Indeed, it may be no easy thing to refer to events by means of essential specifications of them It would be downright impossible, if the event occurs in one but not both of two absolutely indiscernible regions of some world, and any event that could occur in a world of eternal recurrence is an event that occurs in two such regions We might restrict our ambitions, ignore such especially troublesome worlds, and hope we could state a condition for the occurrence of an event which would be necessary and sufficient so far as the better-behaved worlds are concerned That would not quite be an essential specification, but it would approximate to one Even that much would, I fear, be a tall order, though the more we restrict ourselves, the easier it gets Ignoring the worlds infected by indiscernibility might be only a beginning If we restrict our attention to a small range of worlds sufficiently similar to actuality, we might have some hope of success at stating a condition that would be necessary and sufficient, so far as worlds in the range were concerned, for occurrence of a certain event Of course, it would not distinguish events that differ only with respect to regions of worlds outside the range Thus it could not be used to specify one event determinately But often we find it tolerable to leave some indeterminacy in our specifications of things Where exactly does the outback begin? Nobody knows, not because it's a secret, but because we've never bothered to settle exactly what "the outback" denotes And yet we know, near enough, what we're talking about It might be the same way with our best feasible approximations to essential specifications of events the specification might be ambiguous between many events that more-or-less coincide throughout nearby worlds but differ at more distant worlds That might be near enough to determinate reference to meet (some of) our needs We might specify events that way, but then again we might not Often, for instance when we denote events by means of the standard nominalisations, our specifications do not even approximate to being essential

Likewise, while it is clear enough what it would mean to classify events essentially, it seems that very often we do not do that either We

specify and we classify actually occurring events in part by descriptions that fit these events accidentally The event does fit the description, but that very event might have occurred without fitting the description The event is a class consisting of one region of this world together with various regions of other worlds wherein the event might have occurred What goes on in the former region fits the description, but what goes on in some of the latter regions does not The description may imply something about the essential classification of the event it describes, but it is not exhausted by that information

There are many ways an event might be accidentally classified, and I cannot hope to give a complete inventory (1) For events, as for anything else, we can always hoke up thoroughly artificial descriptions "the event that is the Big Bang if Essendon will win the Grand Final, the birth of Calvin Coolidge if not " (2) We might classify it in part by its causes or its effects "Fred's sunstroke," "what Fred did to bring it about that the window is open " (3) We might classify it with reference to its place in a conventional system "Fred's signalling for a left turn," "Fred's signing of the cheque " (4) We might conjoin an accidental circumstance to an essential classification A certain famous event was essentially a fiddling, let us assume, but only accidentally was it a fiddling while Rome burned

I have already suggested that classifications by "constitutive" triples may be accidental, indeed that all three terms of the triple may be inessential (5) We can classify or specify an event by its time, or more generally by the (exact or approximate) location of the region in which it actually occurs, even if it could have occurred at a somewhat different time and place (6) Though there may be some events that involve "constitutive individuals" essentially, I have argued that others—as in the case of the firing squad—involve individuals accidentally (7) Likewise for "constitutive properties " I don't deny that some event with a richly detailed essence might be essentially a strolling, so that necessarily it occurs only in a region wherein someone strolls But some less fragile event might be only accidentally a strolling, it might be a strolling that could have been a striding At the end of the next section, I shall consider the relation between essential and accidental strollings

(8) If an event essentially involves one individual, it may thereby accidentally involve another Suppose an event essentially involves a certain soldier, who happens to belong to a certain army This event cannot occur in regions where there is no counterpart of that soldier, but it can occur where there is a counterpart of the soldier who does

not belong to a counterpart of that army Then the event accidentally involves that army, by way of its soldier Similarly, suppose an event essentially involves a certain temporal person-segment, which is accidentally a segment of Socrates, that would be one way for an event to accidentally involve Socrates

(9) Another possibility for accidental classification turns on nonrigid designation of properties I persist in thinking that "heat" nonrigidly designates whatever phenomenon it is that occupies a certain role and presents itself to us by causing certain manifestations [8] In fact, this is molecular motion, but it might have been something else A world where caloric fluid causes those manifestations is a world where the hot things are the ones with lots of caloric fluid Then "the loss of heat by the poker" may denote an actually occurring event that is essentially a decreasing of molecular motion, and is only accidentally a loss of heat This same event might have occurred at a world where caloric fluid is what presents itself as heat, in a region where the poker absorbed caloric fluid while its molecular motion decreased

In any case of accidental classification or specification, the event actually described is one that might have occurred without fitting the given description Whenever that is so, we must take care with our causal counterfactuals Consider whether the event of a certain description would or would not have occurred under some counterfactual supposition It is one thing to say that the event itself would not have occurred, it is a different thing to say that no event fitting the given description would have occurred For the event might have occurred without fitting the description, or not that event, but another event fitting the description, might have occurred

Many authors, most prominently Davidson, have noted that sentences which do not explicitly mention events often are equivalent to sentences which assert that there occurs (or exists) an event of such-and-such description [9] We must at least agree that many such equiva-

[8] *Contra* the widely shared view of Saul Kripke, *Naming and Necessity* (Oxford Blackwell, 1980), pp 132–33 Likewise for the case of denotationless theoretical terms caloric fluid for instance, would nonrigidly designate anything that both occupied the appropriate role and also was a kind of fluid matter See How to Define Theoretical Terms, in my *Philosophical Papers* Volume I

[9] See Hans Reichenbach, *Elements of Symbolic Logic* (New York Macmillan, 1947) Section 48 Donald Davidson The Logical Form of Action Sentences, Causal Relations and The Individuation of Events, in his *Essays on Actions and Events* (Oxford Oxford University Press, 1980) Judith J Thomson *Acts and Other Events* (Ithaca Cornell University Press 1977)

Sometimes we can ignore the modal dimension of events, and formulate useful equi

lences hold, whether or not we regard them as somehow revealing underlying logical forms, and whether or not we think they are delivered wholesale in accordance with some nice general schema The present treatment of events should not be expected to deliver them wholesale, as witness the mathematical sequence that converges although there occurs no event—in the sense I have in mind—which is its converging But many such equivalences do hold, and these should guide us in considering the variety of ways in which events are classified One reason to insist on accidental classification of events is that with essential classification alone, we do not get the proper equivalences We get one direction easily enough, if there occurs an event that is essentially an *F*-ing, then in some region of this world an event occurs such that, necessarily, it occurs in a region only if something *F*-s there, so something *F*-s But the converse direction will often fail If something *F*-s, it does not in general follow that there occurs an event that is essentially an *F*-ing The most that follows (and even that not in full generality, as witness the converging sequence) is that there occurs an event which, in some way or other, essentially or accidentally, is an *F*-ing If Nero fiddles while Rome burns, I agree that an event occurs which is essentially a fiddling, and an event occurs which is accidentally a fiddling while Rome burns, since the aforementioned fiddling does occur while Rome burns But for reasons to be considered later, I doubt that there occurs any event which is *essentially* a fiddling while Rome burns, and which could not have occurred under happier circumstances

The foregoing discussion contributes only in a negative way to answering the question which ones among the formally eligible properties of regions are the events We cannot answer that question correctly by first investigating our event descriptions, then taking events to be those properties of regions that such descriptions can specify essentially, because our specifications of events may be largely accidental

valences in terms of their thisworldly regions alone Arabella walks across the meadow iff, roughly there is a spatiotemporal region throughout which Arabella walks, and which crosses the meadow Even if we treat adverbial phrases generally as predicate modifiers not hidden predicates in underlying logical forms still we can make use of such equivalences in the analyses of particular adverbs and adverb making prepositions as is done by M J Cresswell in his Prepositions and Points of View *Linguistics and Philosophy* 2 (1977) 1–41, and elsewhere

IV EVENTS STAND IN LOGICAL RELATIONS

Let us say that event *e implies* event *f* iff, necessarily, if *e* occurs in a region then also *f* occurs in that region Considered as classes, event *e* is a subclass included in class *f*

Just because we can define a relation, it doesn't follow that there are any instances of it A theory of events might include the thesis that no events imply other events That was my own view until recently, but now I think it should be rejected

John says "Hello " He says it rather too loudly Arguably there is one event that occurs which is essentially a saying-"Hello" and only accidentally loud, it would have occurred even if John had spoken softly Arguably there is a second event that implies, but is not implied by, the first This event is essentially a saying-"Hello"-loudly, and it would not have occurred if John had said "Hello" but said it softly Both events actually occur, but the second could not have occurred without the first

We have two descriptions "John's saying 'Hello'" and "John's saying 'Hello' loudly " But it does not follow from this alone that we have two events to describe The second description as well as the first might denote the first event, since the second description might describe the first event in part accidentally Alternatively, the first description as well as the second might denote the second event, since the first description might describe the second event by less than the whole of its essence Indeed, even if there are two different events, it still does not follow that one description denotes one and the other denotes the other If both descriptions are somewhat vague or ambiguous, it could be that both denote both

The real reason why we need both events, regardless of which description denotes which, is that they differ causally An adequate causal account of what happens cannot limit itself to either one of the two The first event (the weak one) caused Fred to greet John in return The second one (the strong one) didn't If the second one had not occurred—if John hadn't said "Hello" so loudly—the first one still might have, in which case Fred still would have returned John's greeting Also there is a difference on the side of causes the second event was, and the first wasn't, caused *inter alia* by John's state of tension [10]

[10] See Alvin I Goldman, *A Theory of Human Action* (Englewood Cliffs, New Jersey Prentice-Hall, 1970), p 3

We have two different events, causally distinguished *Different*, but we needn't count them as *distinct* "Distinct" does not mean "non-identical " I and my nose are not identical, but neither are we distinct My nose is part-identical to me, identical to part of me There is a clear sense in which our second event is part of the first the subclass is part of the class, they are neither identical nor distinct (Confusingly, by the inverse variation of intension and extension, there is also a sense in which the first event is part of the second) Indeed, we dare not count the two as distinct For their distinctness, plus my theses about causation, would together imply what is surely false that the first event causes the second For if the first had not occurred, then the second which implies the first would not have occurred either Here is a case of non-causal counterfactual dependence—but *not* between distinct events We may take it as a general principle that when one event implies another, then they are not distinct and their counterfactual dependence is not causal

There is a persuasive intuition—I was long persuaded by it— that it is wrong to count both the first and the second event because if we do, we count something twice over I now think that we do this intuition sufficient justice when we say that the first and second events, though not identical, also are not distinct Compare the equally persuasive intuition that it is double-counting to include both atoms and molecules in our inventory of being—an adequate answer is that the molecules and their atoms are not distinct

There might be two occurrent events that are both implied by some third occurrent event, but that are independent of each other Like this the first event is essentially a saying-"Hello"-loudly and is accidentally abrupt, the second is essentially a saying-"Hello"-abruptly, and is accidentally loud, the third event, which implies both the first and the second, is essentially a saying-"Hello"-loudly-and-abruptly In this case also, the first and second should not be counted as distinct Therefore they are not eligible to stand in causal dependence

But we must beware Suppose we had an occurrent event with a very rich essence the unit class of a region of our world It would imply all the events that occur in that region, none of them would then be distinct, and they would be ineligible to depend causally on one another That is wrong If this conference and the battle of goblins occurred in the same region, they would nevertheless be intuitively distinct, and they might perhaps stand in causal dependence The acceleration of the electron does depend causally on its presence in the field Therefore there must be a limit on how rich the essence of a genuine event can be

Mere unit classes of regions are ruled out, and I should think a good deal else besides

I have argued that we might have two events, one implying the other, such that one is essentially a saying-"Hello"-loudly, and the other is only accidentally a saying-"Hello"-loudly I suggest that this illustrates one important way in which events come by their accidental classifications We have, so to speak, a more and a less detailed version of what happens in a region Both are occurrent events The more detailed version has a richer essence, the otherworldly regions included in it are fewer and less varied, it is more tightly unified by similarity, there is less variety in the ways it could have occurred The more detailed version is one, but only one, of the ways in which the less detailed version could have occurred But it, unlike alternative more detailed versions, happens to be the way the less detailed version actually did occur In such a case, the essential classification of the occurrent more detailed version carries over to become an accidental classification of the less detailed version Likewise, I said above that some events are essentially strollings and some are accidentally strollings The event that is accidentally a strolling is so because it is implied by another event, its more detailed version, which actually occurs and which is essentially a strolling The accidental strolling has alternative more detailed versions, for instance, it is also implied by a non-occurrent event which is essentially a striding If that one of the implying events had been the one that occurred, then the implied event would have been accidentally a striding rather than a strolling

(Above, I noted this implication if there occurs an event that is essentially an *F*-ing, then it follows that something *F*-s Now we can drop the "essentially," though subject to a proviso For if there occurs an event that is accidentally an *F*-ing, provided it comes by this accidental classification in the way just discussed, then there occurs another event that is essentially an *F*-ing, and again it follows that something *F*-s)

Similarly, if some occurrent event essentially involves some individual, then an event that it implies—its less detailed version—accidentally involves that individual Consider the shooting that was done by Ted, but might have been done by Ned How does it involve Ted? Perhaps (but this is not the only possible answer) because it is implied by an occurrent event, its more detailed version, that was a shooting essentially done by Ted

Can it be, as I thought for a time, that accidental classifications of events *always* work this way? Can we say in general that they are

essential classifications of the occurrent more detailed version, carried over via implication? If so, then every accidental classification of one event would have to be an essential classification of another. But some classifications seem so very accidental that no event could have them essentially. Consider accidental classifications in terms of circumstances. There is an event that is accidentally a fiddling while Rome burns, but I doubt that any event is essentially a fiddling while Rome burns. And the example can be made even more extreme. There is an event that is accidentally classifiable as follows: it is a fiddling in the presence of a boy whose grandson will first set foot on the moon. Surely no event is essentially *that*! So we get no unified theory; there must be other ways for events to come by their accidental classifications. For accidental classifications in terms of circumstances, at any rate, it is no mystery how they manage to do it.

Again we would be in trouble if we had events with overly rich essences. If the unit classes of regions were counted as events, to take the worst case, then the accidental descriptions of events that happen to occur in the same region would coalesce. If this conference occurs in the same region as a battle of goblins, it would follow that the conference *is*, albeit accidentally, a battle of goblins. For it would be implied by an occurrent event—namely, the unit class—that is so essentially. That would never do.

V EVENTS HAVE A SPATIOTEMPORAL MEREOLOGY

We have seen how events may be, in a sense, logical parts of one another. If events are classes, as I propose, then they have a mereology in the way that all classes do: the parts of a class are its subclasses.

However, there is a second sense in which events have a mereology, and that will be our business in this section. Regions may be spatiotemporal parts of one another, events are classes of regions, the mereology of the members carries over to the classes, giving us a sense in which events also may be spatiotemporal parts of one another. Each of Sebastian's steps is a spatiotemporal part of his stroll, so is the entire half-stroll performed by the left half of him. Small events that occur in subregions are parts of the big event that occurs in the big region.

Let us say that event e is *essentially part* of event f, iff, necessarily, if f occurs in a region, then also e occurs in a subregion included in that region. (Not necessarily a proper subregion. Therefore an implied event is, in the sense of this definition, essentially part of the implying

event, though in another sense, the subclass sense, the implying event is part of the implied one We have already noted this ambiguity) However, events need not have their spatiotemporal parts essentially Sebastian's stroll might have consisted of more steps, or fewer, or the same number but not the very same steps A war might have consisted of different battles, though the scope for difference is limited We need to provide for accidental, as well as essential, spatiotemporal mereology of events [11]

To do so, we may imitate the proposal of the previous section let the essential mereology of the more detailed occurrent versions carry over via implication to become the accidental mereology of the less detailed versions Let us say that occurrent event *e* is *part* of occurrent event *f* iff some occurrent event that implies *e* is essentially part of some occurrent event that implies *f* This covers not only the case in which *e* is essentially part of *f*, but also the case in which *e* is only accidentally part of *f*, and *f* could have occurred without having *e* as a part

When one event is part of another, whether essentially or accidentally, they are not identical but they are not distinct either (Again, they are partly identical) The same is true when two events share a part in common though neither is a part of the other Kim gives the case of someone who writes "Larry," and as part of that event writes "rr "[12] I add that he writes "Larr" and he writes "rry," these being two overlapping events No two of these four events are distinct As in the case considered previously of events related by implication (which case is subsumed under the present one), events that are not distinct cannot stand in causal dependence If the writing of "rr" had not occurred, the writing of "Larry" would not have, but that does not make those two events be cause and effect Nor can the whole of the writing of "Larr" be a cause on which the writing of "rry" depends, though the first part of it well might be

(Often we do say loosely that event *c* causes event *e* when what's true is that one or more parts of *c* cause one or more parts of *e* Thus we might speak of some prolonged self-perpetuating process as its own cause, when really there is no self-causation and it is earlier parts that cause later parts See Postscript A to "Causation," in this volume)

Once more we would be in trouble if we had events with overly rich

[11] The distinction between essential and accidental mereology of events is noted in W R Carter On Transworld Event Identity *Philosophical Review* 88 (1979) 443–52 and in Michael Smith Actions Attempts and Internal Events, *Analysis* 43 (1983) 142–46

[12] Causes and Counterfactuals

essences If the unit classes of regions were counted as events, it would turn out that whenever one event occurs within the region in which another occurs, the former event is part of the latter

Once we have defined the part-whole relation for events, we can go on as usual to define other mereological notions Thus events overlap iff they have some event as a common part, an atomic event is one that has no events except itself as parts, an event e is the mereological sum of events $f_1, f_2,$ iff e overlaps all and only those events that overlap at least one of the f's, and so on Then we can ask what principles this mereology of events obeys Is it like the unrestricted mereology of individuals, in which several individuals always have another individual as their sum? Or is it like the restricted mereology of chairs, in which several chairs seldom, if ever, have another chair as their sum? Or is it in between? I suggest that events are at any rate more amenable to summation than chairs are a war may be the sum of its battles, a conference may be the sum of its sessions But I leave open the question whether several events, however miscellaneous, always have another event as their sum If there is unrestricted summation, then there can be no limit on how large and disconnected and disunified an event may be, whereas if events must have some unity to them, then some attempted summations would fail to yield a genuine event (Maybe they yield a property of regions which is formally eligible, but not an event) It is hard to find arguments to settle the question Our events are meant to serve as causes and effects, but it seems hard to tell when we can be content to say only that several events are joint causes and separate effects, and when we must also insist on a single event that is their sum

Another question is this given any subregion of the region in which an event occurs, is there a part of that event which occurs in that subregion? I am inclined to think that there is in the case of a suitable subregion—one with boundaries that are reasonably simple in shape, or that match the boundaries of something within the region—but not in the case of an arbitrary subregion But again I cannot find arguments to settle the question

VI THE HISTORY OF EVENTS IS THE WHOLE OF HISTORY

Suppose we are given the complete history of the world's events exactly which events occur in exactly which regions, throughout all of space and time Then no historical information is lacking No two

possible worlds could be exactly alike in their histories of events, yet unlike elsewhere in their histories of matters of manifest particular fact The total history of events implies every historical truth and contradicts every historical falsehood

It must be so, given the four theses about causation and explanation that I took as my starting point To explain is to give information about the causal history of the explanandum event, and that history is a structure of causally related events If history is a patchwork of events and nonevents, and if the nonevents are not implicitly given by the history of events, then the nonevents are left out of causal histories Then they never enter into the explanations of events But is there really anything in history that is thus isolated, that never plays any part in explaining why any event took place? I do not see what it could be

Not all events involve change We cannot afford to count the unchanges as nonevents, for the unchanges may be needed to complete causal histories Indeed, the causal history of an event—an uncontroversial event, an abrupt change—might consist entirely of unchanges An isolated particle has existed from all eternity, it is unstable, and has at every moment a chance of decaying, eventually it does decay The decay was caused (probabilistically) by the previous presence, at all earlier times, of the unchanging particle The causal history that explains this event is entirely changeless It is a thoroughly uneventful course of events

The need for unchanges as events is urgent in connection with causal theories of perception, memory, persistence over time, and so on The causal chains required by such theories often will consist simply of something continuing to exist a travelling signal, a memory trace, a surviving person or a persisting lump of matter The need is urgent also in connection with causal preemption The preempting cause may cause its effect not because the effect depends directly on the cause, but because the effect depends causally on some intermediate event which in turn depends causally on the preempted cause (See Postscript E to "Causation," in this volume) If we could not count unchanges, there would be realistic cases in which we could find no suitable intermediate event

Terminology is not the issue If it is abuse of language to call unchanges "events," so be it The point is that we must have them as causes and effects

This section does not contribute directly to answering the question which among the formally eligible properties of regions are events But it is indirectly relevant no adequate demarcation of the events can be

too restrictive, else it leaves us with not enough events to make up the whole of history

Given the history of events, the whole of history is implicitly given, but not, perhaps, the whole truth about contingent matters It is an independent, and difficult, question whether two worlds exactly alike in their histories would have to be exactly alike in every way in their chances, their laws, their modal truths and counterfactuals, their causal relations, and so on (See the introduction to this volume) My present point does not touch that question I say just that total history supervenes on the history of events, whatever else may or may not supervene in turn on total history

VII EVENTS ARE PREDOMINANTLY INTRINSIC

Consider the alleged event such that, necessarily, it occurs in a region iff (1) that region is located at a certain time t, (2) Xanthippe is present in that region, and (3) at time t, someone dies who has been married to her until that time [13] I shall call this alleged event *the widowing of Xanthippe*, since I need a name for it, but I do not mean to suggest (and do not believe) that this is what the phrase would ordinarily denote (Maybe it would denote the death of Socrates, or maybe no event at all, but rather a certain fact, that is a certain true proposition) Condition (3) is entirely extrinsic It has nothing to do with the qualitative character of the region in question, and everything to do with what goes on at other times and places The other conditions are at least partly intrinsic But they do not go very far toward picking out those regions where the alleged event allegedly occurs

We can devise a more extreme, but also more artificial, case in which a supposed event is almost purely extrinsic in its essence Begin with some genuine event this conference, let us say Define its *centennial* to be that event—if such it be—such that, necessarily, it occurs in a region R iff the original event occurs in the region which results from shifting R exactly 100 years backward, holding fixed its place, size, and shape A centennial in this artificial sense is not at all intrinsic Whether it occurs in a region is completely independent of what things are there, how they are arranged, what they are like, and what they are doing It is quite unlike a centennial in the ordinary sense a genuine event wherein some previous event is remembered and celebrated The qualitative

[13] See Kim, Noncausal Connections

character of the region could be anything (With one exception it must have a size and shape which the original event could possibly have)

We can define an *intrinsic* property of a region as one such that, whenever two possible regions are perfect duplicates, the property belongs to both or neither Likewise a *purely extrinsic* property is one such that, for any possible region, there is some possible region which is a perfect duplicate of it and has the property [14]

I think there are no such events as the widowing of Xanthippe or the centennial of this conference We have the properties of regions, right enough, and they are formally eligible But they are not events because they are purely or predominantly extrinsic, whereas the properties of regions that are genuine events are predominantly intrinsic

Extrinsic events—or, more generally, events not predominantly intrinsic—are objectionable on three counts (1) They offend our sense of economy We would seem to count the death of Socrates twice over in our inventory of events, once as itself and again as (what I am calling) the widowing of Xanthippe Still more clearly, the centennial of this conference is but a shadow of the conference itself (2) They stand in relations of noncausal counterfactual dependence to those genuine events in virtue of which they occur Without the death of Socrates, the widowing of Xanthippe would not have occurred (She might still have been widowed sooner or later But recall that the widowing of Xanthippe, as I defined it, had its time essentially) Without this conference, its centennial would not occur What's worse without its centennial, the conference would not occur None of these is a genuine case of causal dependence (3) They also stand in relations of noncausal counterfactual dependence to other genuine events, events logically independent of them Without the widowing of Xanthippe, the subsequent cooling of Socrates's body would not have occurred (For in that case he would not have died when he did) Without the centennial of this conference 100 years hence, our homeward departures a few days from now would not occur These also are not genuine cases of causal dependence Instantaneous and backward causation are not so very easy!

The first two objections might be answered by claiming that the events in question are in some sense not distinct It is a bit hard to say

[14] We could go the other way, and define *duplication* as sharing of all intrinsic properties Here we are dealing with a substantial circle of interdefinables and so have a choice of alternative primitives For further discussion see my Extrinsic Properties *Philosophical Studies* 44 (1983) 197–200 and New Work for a Theory of Universals 355–57

why not in the case of the death of Socrates and the widowing of Xanthippe, either could have occurred without the other (They are logically, though not counterfactually, independent) Further, they occur in nonoverlapping regions But suppose that some suitable sense of distinctness could be found That would do nothing at all to answer the third objection, which is therefore the decisive one

Should I say that genuine events must be entirely, not just predominantly, intrinsic? I think not For if an event were an entirely intrinsic property of regions, then it would have to occur in any duplicate of any region in which it occurs Suppose it is an event that could occur within some epoch of some world of eternal recurrence—that's not a very strong condition, especially not on the hypothesis that events are entirely intrinsic But every region within an epoch has duplicates within all the other epochs So this event occurs many times over in a single world, *contra* our stipulation that we are talking about particular events That stipulation applies even to a world of eternal recurrence, because surely we want to distinguish the events of different epochs, indiscernible though they may be A similar problem arises from duplication of regions within more ordinary worlds It may be hard for big regions to be exactly alike in what goes on in them, but remember that submicroscopic events go on in submicroscopic regions, and electron-sized regions will have less opportunity to differ than larger regions do We don't want a collision of two electrons to be one event that occurs in all regions of the same world where two electrons collide in just that way, rather, we want to distinguish different collisions that occur once each If the regions do not differ intrinsically, then an event that occurs in one but not all of them cannot be entirely intrinsic It can be predominantly intrinsic, and that is all I should require

The rejection of overly extrinsic events allows me to return to some unfinished business Earlier, I asked how an event comes to be accidentally classified as a fiddling while Rome burns, not, I said, because it has an occurrent more detailed version that is essentially a fiddling while Rome burns Now I can give the reason the alleged event that is essentially a fiddling while Rome burns would be too extrinsic

I also avoided committing myself to the existence of genuine events which essentially involve Socrates, that is, which cannot occur except in a region where our Socrates, or a counterpart of him, is present Therefore, I also could not commit myself to the existence of events that involve Socrates accidentally by having occurrent more detailed versions that involve him essentially Now I can give the reason being a counterpart of Socrates is rather an extrinsic matter Counterparts are

united by similarity, and the similarity in question may be largely extrinsic This is so especially on counterpart relations which stress match of origins Socrates might have been—some counterpart of him is—a dimwitted and taciturn politician, most noted for his good looks and slender build, but only if his origins had been more or less exactly as they actually were I am less of an enthusiast for match of origins than most philosophers of modality, nevertheless, I must agree that it often carries a lot of weight And if similarities in respect of one's role in society and in history also carry weight in making counterparts, as I think they sometimes do, those too are extrinsic The origins and role of Socrates are not intrinsic even to Socrates taken as a whole Still less are they intrinsic to the temporal part of him whereby he is present in a region and involved there in some event So the property of having present in it a counterpart of Socrates is not an intrinsic property of a region, so an alleged event that essentially involved Socrates would be to that extent extrinsic I am not sure whether it would be extrinsic enough to deserve rejection

(One worry about it is needless, I think We might fear that any extrinsic element in an event will give rise to cases of noncausal counterfactual dependence Suppose that Socrates is essentially the outcome of a certain event of conception, and that his death is essentially his Then his death is essentially preceded by his conception Had his conception not occurred, his death would not have—because then no death could have been *his* But this time the dependence does seem causal, so we have no problem Is there perhaps a new problem causal dependence between *in*distinct events? No, there is a sense in which the death implies the conception, and I did say that when one event implies another in a certain sense they are not distinct and not eligible to stand in causal dependence, but the two senses are not the same Nothing I said detracts from the distinctness, and eligibility for dependence, of two events that occur in nonoverlapping regions)

Can we afford to reject events that essentially involve Socrates? That depends on whether we can do without them Certainly some events involve Socrates at least accidentally But can we find any way for an event to involve Socrates accidentally if none involves him essentially? Of course it will not do to say that he is involved just by being present in the region where the event actually occurs—what if those goblins are present in the region of this conference? Hence my rather cumbersome suggestion about involvement via a temporal segment It is not necessary that counterpart relations must work in the same way for the parts as for the whole, or that the counterparts of the parts must be

parts of the counterpart of the whole I find it somewhat plausible that a counterpart relation for temporal segments of people could work more by intrinsic similarity and less by origins or roles than a counterpart relation for whole people does If so, I have less difficulty with an event that essentially involves a person-segment than with one that essentially involves a person Now suppose we have a death that involves a certain person-segment (Whether essentially or accidentally doesn't matter, if we have one that involves it accidentally, we have another that involves it essentially) Now suppose that in fact this segment is part of Socrates Accidentally so not all its counterparts are parts of his Now at last we've got Socrates involved in his own death, and in a way that bypasses any unduly extrinsic events I wouldn't say no to a simpler method[1]

VIII EVENTS ARE NOT DISJUNCTIVE

Let us call event e the *disjunction* of events f_1, f_2, iff, necessarily, event e occurs in a region iff either f_1 or f_2 or occurs there (There may be finitely or infinitely many disjuncts) Equivalently, e is the disjunction of the f's iff each of the f's implies e, and e implies any other event of which the same is true Considered as classes, a disjunction of events is simply their union

I do not deny that some events are disjunctions of others An event that is essentially a stamping might, for instance, be the disjunction of one event that is essentially a stamping-the-left-foot and another that is essentially a stamping-the-right-foot But I do not think that just any class will have a disjunction The disjuncts must not be too miscellaneously varied In calling an alleged event *disjunctive*, as opposed to saying that it is the disjunction *of* such-and-such, I mean that it would be the disjunction of some disjuncts that are overly varied An example might be the supposed disjunction of one event that is essentially a walking and another that is essentially a talking

Disjunctive events are *prima facie* open to the same three objections that I raised against extrinsic events (1) They offend against economy To count both disjunction and disjunct looks like counting the same thing twice But to this an adequate reply is that the two are not distinct, by definitions already given (2) They stand in relations of noncausal counterfactual dependence with their logical relatives, namely their disjuncts Without the disjunction, no disjunct could have occurred Without the occurrent disjunct, the disjunction would not

have occurred (except in special cases where some other disjunct would have occurred instead) It is an adequate reply that these are not cases of counterfactual dependence between distinct events (It is just as well that these two objections fail They threaten to prove too much—namely, that no event is *ever* the disjunction of others) (3) They stand in relations of noncausal counterfactual dependence with other events that are not their logical relatives, and are clearly distinct from them This objection is the decisive one

Fred talks, and his talking causes Ted to laugh Suppose that besides Fred's talking there is another event, the disjunctive event of Fred's talking-or-walking Without it, Fred's talking would not have occurred, and neither would Ted's laughing So this disjunctive event also causes Ted to laugh That is intuitively wrong No such event causes Ted's laughing, or anything else Given the theses I took as my starting point, that can only be because there is no such event Hence disjunctive events are to be rejected

This may seem too hasty It may seem that there are some disjunctive events that we dare not lose, because they are indispensable as causes and effects Suppose that a certain poker cools down, that is, it loses heat Suppose also that I am right to say that "heat" is not a rigid designator, heat is whatever property it is that occupies a certain causal role, and so might be one property or might be another So there are many quite different ways that the poker might lose heat, depending on what sort of world it is in (It or its counterpart) Its molecular motion might decrease, in a world where molecular motion is what occupies the role, or it might lose caloric fluid, in a world where caloric fluid is what occupies the role, or So it seems that losing heat is quite a disjunctive affair, and what's worse, extrinsic, since whether one property or another occupies the heat-role depends on what goes on throughout the world in question, not just on the region of it where the poker is All the same, isn't the loss of heat by the poker a perfectly good event? Isn't it a cause of the poker's slight contraction, and an effect of my taking the poker out of the fire?

I agree that the loss of heat by the poker is a perfectly good event But it is not disjunctive, so we needn't fear to lose it if we reject disjunctive events For it is not the disjunction of one event which is essentially a decreasing of molecular motion in a world where that occupies the role, and another which is essentially a loss of caloric fluid in a world where that occupies the role, and That would be an event that was *essentially* a losing of heat (And that had no relevant narrower essential classification, so that some disjuncts would be eli-

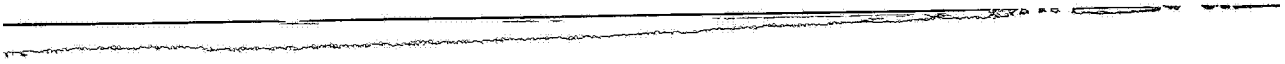
minated, let me omit this complication henceforth) That alleged event does indeed deserve rejection Twice over both because it is disjunctive and because the disjuncts are excessively extrinsic But the perfectly good event which is the loss of heat by the poker is specified as such *accidentally*, not essentially It is essentially a decreasing of molecular motion This event does not occur in otherworldly regions where the poker loses heat by losing caloric fluid But it does occur in regions where the molecular motion decreases and yet the poker does not lose heat, these being regions of worlds where something besides molecular motion is what occupies the heat-role In those worlds, of course, the event fails to fits its thisworldly accidental description as a loss of heat

Likewise in general Whenever some term nonrigidly designates the occupant of a role, and that role could be occupied in a variety of ways, the term becomes unsuitable for essential specification of events If being fragile means having some or another basis for a disposition to break when struck, and if many different properties could serve as such bases (under this- or otherworldly laws), then no genuine event is essentially classifiable as the window's being fragile There is a genuine event which is accidentally classifiable in terms of fragility, essentially, however, it is a possession of such-and-such molecular structure, that being the actual basis of the window's fragility (This event is an unchange, but I haven't rejected those) I think this observation gives the sense in which, as Prior *et al* say, dispositions are "inefficacious "[15] And if I am right to think that mental states are definable as occupants of causal roles, then no genuine event is essentially classifiable as my being in pain There are pain events, no doubt of it, but they are pain events only accidentally, just as pain itself is a property that only contingently occupies its role and deserves its name Essentially, the events are firings of neurons, perhaps—unless "firing" and "neuron" also are terms for occupants of roles, in which case we must get more physical before we finally reach an essential classification

If there are no extrinsic or disjunctive events to be caused, still there are extrinsic or disjunctive truths about regions to be explained They can be explained, of course And their explanations can be mostly or entirely causal, even if my theses about causal explanation of events do not apply directly The explanandum truth is made true by a pattern of

[15] See Elizabeth W Prior, Robert Pargetter, and Frank Jackson Three Theses about Dispositions, *American Philosophical Quarterly* 19 (1982) 251–57, also Section III of Causal Explanation, in this volume

genuine, occurrent events (This making true is logical, not causal) These events have their causal histories Explanatory information about the explanandum truth consists in part of noncausal information about the truth-making pattern itself what sort of pattern it is, and what events comprise it And it consists in part of information about the causal histories of the events that comprise the pattern As usual, explaining means providing some explanatory information The serving provided may consist of noncausal information about the pattern, or causal information about its events, or some of each Why did Xanthippe become a widow? Because she was married to Socrates at the time of his death (Noncausal) Because Socrates was made to drink hemlock (Causal, with the noncausal background most likely presupposed) Why did Fred talk or walk then? Because he talked (noncausal) and he did that because he had just heard a joke he couldn't keep to himself (causal)



PART SEVEN

Dependence and Decision

TWENTY-FOUR

Veridical Hallucination and Prosthetic Vision

I

I see Before my eyes various things are present and various things are going on The scene before my eyes causes a certain sort of visual experience in me, thanks to a causal process involving light, the retina, the optic nerve, and the brain The visual experience so caused more or less matches the scene before my eyes All this goes on in much the same way in my case as in the case of other people who see And it goes on in much the same way that it would have if the scene before my eyes had been visibly different, though in that case the visual experience produced would have been different

How much of all this is essential to seeing?

II

It is not far wrong to say simply that someone sees if and only if the scene before his eyes causes matching visual experience So far as I know, there are no counterexamples to this in our ordinary life Shortly we shall consider some that arise under extraordinary circumstances

But first, what do we mean by "matching visual experience"? What goes on in the brain (or perhaps the soul) is not very much like what goes on before the eyes They cannot match in the way that a scale model matches its prototype, or anything like that Rather, visual experience has informational content about the scene before the eyes, and it matches the scene to the extent that its content is correct

Visual experience is a state characterised by its typical causal role, and its role is to participate in a double causal dependence Visual experience depends on the scene before the eyes,[1] and the subject's beliefs about that scene depend in turn partly on his visual experience The content of the experience is, roughly, the content of the belief it tends to produce

The matter is more complicated, however The same visual experience will have a different impact on the beliefs of different subjects, depending on what they believed beforehand (And on other differences between them, e g differences of intelligence) Holmes will believe more on the basis of a given visual experience than Watson, and Watson in turn will believe more than someone will who suspects that he has fallen victim to a field linguist no less powerful than deceitful [2] We should take the range of prior states that actually exist among us, and ask what is common to the impact of a given visual experience on all these states Only if a certain belief would be produced in almost every case may we take its content as part of the content of the visual experience (The more stringently we take "almost every," the more we cut down the content of the visual experience and the more of its impact we attribute to unconscious inference, for our purposes, we need not consider how that line ought to be drawn)

Beliefs produced by visual experience are in large part self-ascriptive the subject believes not only that the world is a certain way but also that he himself is situated in the world in a certain way To believe that the scene before my eyes is stormy is the same as to believe that I am facing a stormy part of the world Elsewhere[3] I have argued that the objects of such beliefs should be taken, and that the objects of

[1] I shall have more to say about this dependence in what follows So although my concern here is with the analysis of seeing in terms of visual experience, what I say would also figure in a prior analysis of visual experience in terms of its definitive causal role

[2] The problem of the suspicious subject is raised in Frank Jackson, *Perception A Representative Theory* (Cambridge Cambridge University Press, 1977) pp 37–42

[3] Attitudes *De Dicto* and *De Se* *The Philosophical Review*, LXXXVIII (1979) pp 513–543

all beliefs may be taken, as properties which the subject self-ascribes Hence the content of visual experience likewise consists of properties—properties which the subject will self-ascribe if the visual experience produces its characteristic sort of belief The content is correct, and the visual experience matches the scene before the eyes, to the extent that the subject has the properties that comprise the content of his visual experience

Equivalently we might follow Hintikka's scheme and take the content of visual experience as a set of alternative possibilities [4] A modification is desirable, however, in view of the self-ascriptive character of visually produced belief We should take these visual alternatives not as possible worlds but as possible individuals-situated-in-worlds The visual experience characteristically produces in the subject the belief that he himself belongs to this set of alternative possible individuals Matching then means that the subject is, or at least closely resembles, a member of his alternative set

Not all of the content of visual experience can be characterised in terms of the beliefs it tends to produce It is part of the content that the duck-rabbit look like a duck or a rabbit, but the belief produced is that there is no duck and no rabbit but only paper and ink However, aspects of the content that do not show up in the produced belief also are irrelevant to our task of saying what it is for visual experience to match the scene before the eyes We can therefore ignore them

III

I shall not dwell on the question whether it is possible to see even if the scene before the eyes does not cause matching visual experience Three sorts of examples come to mind (1) Perhaps someone could see without having visual experience He would need something that more or less played the role of visual experience, but this substitute might not be visual experience, either because it played the role quite imperfectly[5] or because it is not what normally plays the role in human

[4] Jaakko Hintikka On the Logic of Perception , in his *Models for Modalities Selected Essays* (Dordrecht Reidel, 1969) The proposed modification solves (by theft rather than toil) a problem for Hintikka s important idea of perceptual cross-identification where do we get the cross-identification of the perceiving subject himself, in relation to whom we perceptually cross identify the things that surround him?

[5] As in cases of blind sight in which the subject claims to have no visual experience and yet acquires information about the scene before his eyes just as if he did

beings (or in some other natural kind to which the subject in question belongs) [6] (2) Perhaps someone could see in whom the scene before the eyes causes non-matching visual experience, provided that the failure of match is systematic and that the subject knows how to infer information about the scene before the eyes from this non-matching visual experience (3) Perhaps someone could see in whom the scene elsewhere than before the eyes causes visual experience matching that scene, but not matching the scene before the eyes (if such there be) I do not find these examples clear one way or the other, and therefore I shall consider them no further They will not meet the conditions for seeing that follow, wherefore I claim only sufficiency and not necessity for those conditions

Two further preliminaries (1) My analysandum is seeing in a strong sense that requires a relation to the external scene Someone whose visual experience is entirely hallucinatory does not see in this strong sense I take it that he can be said to see in a weaker, phenomenal sense—he sees what isn't there—and this is to say just that he has visual experience (2) My analysandum is seeing in the intransitive sense, not seeing such-and-such particular thing The latter analysandum poses all the problems of the former, and more besides it raises the questions whether something is seen if it makes a suitable causal contribution to visual experience but it is not noticed separately from its background, and whether something is seen when part of it—for instance, its front surface—makes a causal contribution to visual experience [7]

IV

My first stab is good enough to deal with some familiar counterexamples to causal analyses of seeing they are not cases of seeing because they are not cases in which the scene before the eyes causes matching visual experience [8]

[6] See my Mad Pain and Martian Pain , in Ned Block, ed *Readings in the Philosophy of Psychology* Vol 1 (Cambridge, Massachusetts Harvard University Press, 1980)

[7] Alvin Goldman considers transitive seeing in his Discrimination and Perceptual Knowledge , *Journal of Philosophy*, LXXIII (1976), pp 771–791 Despite the difference of analysandum, I have followed his treatment to a considerable extent

[8] Example 1 and an auditory version of Example 2 are due to P F Strawson, Causation in Perception , in his *Freedom and Resentment and other essays* (London Methuen, 1974), pp 77–78

Example 1 The Brain I hallucinate at random, by chance I seem to see a brain floating before my eyes, my own brain happens to look just like the one I seem to see, my brain is causing my visual experience, which matches it I do not see No problem my brain is no part of the scene before my eyes

Example 2 The Memory I hallucinate not at random, visual memory influences the process, thus I seem to see again a scene from long ago, this past scene causes visual experience which matches it I do not see No problem the past scene is not part of the scene before my eyes [9]

However, more difficult cases are possible They are cases of *veridical hallucination*, in which the scene before the eyes causes matching visual experience, and still one does not see They show that what I have said so far does not provide a sufficient condition for seeing

Example 3 The Brain Before the Eyes As in Example 1, I hallucinate at random, I seem to see a brain before my eyes, my own brain looks just like the one I seem to see, and my brain is causing my visual experience But this time my brain is before my eyes It has been carefully removed from my skull The nerves and blood vessels that connect it to the rest of me have been stretched somehow, not severed It is still working and still hallucinating

Example 4 The Wizard The scene before my eyes consists mostly of a wizard casting a spell His spell causes me to hallucinate at random, and the hallucination so caused happens to match the scene before my eyes

Example 5 The Light Meter I am blind, but electrodes have been implanted in my brain in such a way that when turned on they will

[9] However, it seems that some past things are part of the scene now before my eyes distant stars as they were long ago, to take an extreme case It would be circular to say that they, unlike the past scene in Example 2 are visible now Perhaps the best answer is that the stars, as I now see them, are not straighforwardly past for lightlike connection has as good a claim as simultaneity-in-my-rest-frame to be the legitimate heir to our defunct concept of absolute simultaneity (I owe the problem to D M Armstrong and the answer to Eric Melum)

cause me to have visual experience of a certain sort of landscape A light meter is on my head It is connected to the electrodes in such a way that they are turned on if and only if the average illumination of the scene before my eyes exceeds a certain threshold By chance, just such a landscape is before my eyes, and its illumination is enough to turn on the electrodes

V

Ordinarily, when the scene before the eyes causes matching visual experience, it happens as follows Parts of the scene reflect or emit light in a certain pattern, this light travels to the eye by a more or less straight path, and is focused by the lens to form an image on the retina, the retinal cells are stimulated in proportion to the intensity and spectral distribution of the light that falls on them, these stimulated cells stimulate other cells in turn, and so on, and the stimulations comprise a signal which propagates up the optic nerve into the brain, and finally there is a pattern of stimulation in the brain cells which either is or else causes the subject's visual experience

That is not at all what goes on in our three examples of veridical hallucination Rather, the scene before the eyes causes matching visual experience by peculiar, non-standard causal processes Perhaps, as has been proposed by Grice[10] and others, seeing requires the standard causal process That would explain why Examples 3, 4, and 5 do not qualify as cases of seeing

(The proposal faces a technical dilemma If the standard process is defined as the process in which light is reflected or emitted, etc (as above), then it seems to follow that few of us now (and none in the not-too-distant past) know enough to have the concept of seeing, whereas if the standard process is defined as the most common process by which the scene before the eyes causes matching visual experience, whatever that may be, then it seems to follow that any of our examples of veridical hallucination might have been a case of seeing, and what I am doing now might not have been, if only the frequencies had been a bit different Either conclusion would be absurd However, the

[10] H P Grice, The Causal Theory of Perception, *Proceedings of the Aristotelian Society* Supplementary Volume XXXV (1961) pp 121–152

dilemma can be avoided by appeal to the recent idea of fixing reference by rigidified descriptions)[11]

Unfortunately, requiring the standard process would disqualify good cases along with the bad Some cases in which the scene before the eyes causes matching visual experience by a non-standard process seem fairly clearly to be cases of genuine seeing, not veridical hallucination

Example 6 The Minority It might be found that a few of us have visual systems that work on different principles from other people's The differences might be as extreme as the difference between AM versus FM transmission of signals, analogue versus digital processing, or point-by-point measurement of light versus edge detection If so, would we be prepared to say that the minority don't really see? Would those who belong to the minority be prepared to say it? Surely not

I anticipate the reply that the abnormal process in the minority is not different enough, the boundaries of the standard process should be drawn widely enough to include it But I think this puts the cart before the horse We know which processes to include just because somehow we already know which processes are ones by which someone might see

Example 7 The Prosthetic Eye A prosthetic eye consists of a miniature television camera mounted in, or on, the front of the head, a computer, and an array of electrodes in the brain The computer receives input from the camera and sends signals to the electrodes in such a way as to produce visual experience that matches the scene before the eyes When prosthetic eyes are perfected, the blind will see The standard process will be absent, unless by "standard process" we just mean one that permits seeing, but they will see by a non-standard process

Some prosthetic eyes are more convincing than others as means for genuine seeing (1) It seems better if the computer is surgically implanted rather than carried in a knapsack, but better if it's carried in a knapsack rather than stationary and linked by radio to the camera and electrodes (2) It seems better if the prosthetic eye contains no

[11] See the discussion of the metre and the metre bar in Saul A Kripke, Naming and Necessity , in Donald Davidson and Gilbert Harman *Semantics of Natural Language* (Dordrecht Reidel, 1972), pp 274–275 and 288–289

parts which can be regarded as having wills of their own and cooperating because they want to (3) It seems better if the prosthetic eye works in some uniform way, rather than dealing with different sorts of inputs by significantly different means (4) It seems better if it does not use processes which also figure in the standard processes by which we sometimes hallucinate But if these considerations influence us, presumably it is because they make the prosthetic eye seem a little more like the natural eye (Or so we think—but we just might be wrong about the natural eye, and these properties of a prosthetic eye just might detract from the resemblance) Why should that matter, once we grant that the standard process is not required? I see no real need for any limits on how a prosthetic eye might work Even the least convincing cases of prosthetic vision are quite convincing enough

If you insist that "strictly speaking" prosthetic vision isn't really seeing, then I'm prepared to concede you this much Often we do leave semantic questions unsettled when we have no practical need to settle them Perhaps this is such a case, and you are resolving a genuine indeterminacy in the way you prefer But if you are within your rights, so, I insist, am I I do not really think my favoured usage is at all idiosyncratic But it scarcely matters I would like to understand it whether it is idiosyncratic or not

VI

The trouble with veridical hallucination is not that it involves a non-standard causal process Is it perhaps this that the process involved produces matching visual experience only seldom, perhaps only this once?

No, someone might go on having veridical hallucinations for a long time Veridical hallucinations are improbable, and a long run of them is still more improbable, but that doesn't make it impossible No matter how long they go on, the sorts of occurrences I've classified as cases of veridical hallucination still are that and not seeing

On the other hand, a process that permits genuine seeing might work only seldom, perhaps only this once

Example 8 The Deathbed Cure God might cure a blind man on his deathbed, granting him an instant of sight by means of some suitable non-standard process For an instant he sees exactly as others do Then

he is dead The scene before his eyes produces matching visual experience by a suitable process, but only this once

Example 9 The Loose Wire A prosthetic eye has a loose wire Mostly it flops around, and when it does the eye malfunctions and the subject's visual experience consists of splotches unrelated to the scene before the eyes But sometimes it touches the contact it ought to be bonded to, and as long as it does, the eye functions perfectly and the subject sees Whether he sees has nothing to do with whether the wire touches the contact often, or seldom, or only this once

The proposal isn't far wrong It asks almost the right question when the scene before the eyes causes matching visual experience this time, is that an isolated case or is it part of a range of such cases? The mistake is in asking for a range of actual cases, spread out in time Rather, we need a range of counterfactual alternatives to the case under consideration

VII

What distinguishes our cases of veridical hallucination from genuine seeing—natural or prosthetic, lasting or momentary—is that there is no proper counterfactual dependence of visual experience on the scene before the eyes If the scene had been different, it would not have caused correspondingly different visual experience to match that different scene Any match that occurs is a lucky accident It depends on the scene being just right In genuine seeing, the fact of match is independent of the scene Just as the actual scene causes matching visual experience, so likewise would alternative scenes Different scenes would have produced different visual experience, and thus the subject is in a position to discriminate between the alternatives

This is my proposal if the scene before the eyes causes matching visual experience as part of a suitable pattern of counterfactual dependence, then the subject sees, if the scene before the eyes causes matching visual experience without a suitable pattern of counterfactual dependence, then the subject does not see

An ideal pattern of dependence would be one such that any scene whatever would produce perfectly matching visual experience But that is too much to require Certainly one can see even if the match, actual and counterfactual, is close but imperfect and the content of

visual experience is mostly, but not entirely, correct Perhaps indeed this is our common lot Further, one can see even if there are some alternative scenes that would fail altogether to produce matching visual experience, so long as the actual scene is not one of those ones

Example 10 The Laser Beam I see now, but if the scene before my eyes had included a powerful laser beam straight into my eyes, I would have been instantly struck blind and would not have had matching visual experience even for a moment

Example 11 The Hypnotic Suggestion I must do business with Martians and I can't stand the sight of them The remedy is hypnotic suggestion when a Martian is before my eyes I will seem to see not a Martian but a nice black cat Thus when there are Martians around, the scene before my eyes causes visual experience that does not match the scene very closely But when there are no Martians, I see perfectly well [12]

We cannot require that any two different scenes would produce different visual experience, for they might differ in some invisible respect, in which case the same visual experience would match both equally well Its content would concern only those aspects of the scene in which both are alike For one who sees, *visibly* different scenes would (for the most part) produce different visual experience, but that is unhelpful unless we say which differences are the visible ones, and that seems to be an empirical matter rather than part of the analysis of seeing What can be required analytically is that there be plenty of visible differences of some sort or other, that is, plenty of different alternative scenes that would produce different visual experience and thus be visually discriminable

That would almost follow from a requirement of match over a wide range of alternative scenes But not quite Most of our visual experience is rich in content, but some is poor in content and would match a wide range of alternative scenes equally well Any pitch-dark scene would produce matching visual experience—what content there is would be entirely correct—but it would be the same in every case See-

[12] Adapted from an olfactory example in Robert A Heinlein, *Double Star* (Garden City, New York Doubleday 1956) Ch 3

ing is a capacity to discriminate, so this sort of match over a wide range of alternatives will not suffice

I conclude that the required pattern of counterfactual dependence may be specified as follows There is a large class of alternative possible scenes before the subject's eyes, and there are many mutually exclusive and jointly exhaustive subclasses thereof, such that (1) any scene in the large class would cause visual experience closely matching that scene, and (2) any two scenes in different subclasses would cause different visual experience

The requirement admits of degree in three ways How large a class? How many subclasses? How close a match? The difference between veridical hallucination and genuine seeing is not sharp, on my analysis It is fuzzy, when the requirement of suitable counterfactual dependence is met to some degree, but to a degree that falls short of the standard set by normal seeing, we may expect borderline cases And indeed it is easy to imagine cases of partial blindness, or of rudimentary prosthetic vision, in which the counterfactual dependence is unsatisfactory and it is therefore doubtful whether the subject may be said to see

VIII

A further condition might also be imposed that in the actual case the subject's visual experience must be rich in content, that it must not be the sort of visual experience that would match a wide range of scenes equally well For instance, it must not be the sort of visual experience that we have when it is pitch dark This condition of rich content is needed to explain why we do not see in the dark, even though the scene before the eyes causes matching visual experience as part of a suitable pattern of counterfactual dependence

But we are of two minds on the matter We think we do not see in the dark, but also we think we find things out by sight only when we see, and in the pitch dark, we find out by sight that it is dark How else—by smell? By the very fact that we do not see?—No, for we also do not see in dazzling light or thick fog, and it is by sight that we distinguish various situations in which we do not see

In a sense, we do see in the dark when we see that it is dark In a more common sense, we never see in the dark There is an ambiguity in our concept of seeing, and the condition of rich content is often but not always required When it is, it admits of degree and thus permits still another sort of borderline case of seeing

IX

Given a suitable pattern of counterfactual dependence of visual experience on the scene before the eyes (including both the actual case and its counterfactual alternatives) it is redundant to say as I did that the scene causes, or would cause, the visual experience To make the explicit mention of causation redundant, according to my counterfactual analysis of causation, we need not only a suitable battery of scene-to-visual-experience counterfactuals but also some further counterfactuals Along with each counterfactual saying that if the scene were S the visual experience would be E, we need another saying that if the scene S were entirely absent, the visual experience would not be E Counterfactuals of the latter sort may follow from the battery of scene-to-visual-experience counterfactuals in some cases, but they do not do so generally According to the counterfactual analysis of causation that I have defended elsewhere,[13] any such counterfactual dependence among distinct occurrences is causal dependence, and implies causation of the dependent occurrences by those on which they depend It would suffice if our counterfactuals said just that if the scene before the eyes were so-and-so, then the visual experience would be such-and-such

If we leave the causation implicit, however, then we must take care that the counterfactuals from scene to visual experience are of the proper sort to comprise a causal dependence We must avoid backtrackers those counterfactuals that we would support by arguing that different effects would have to have been produced by different causes [14] Backtracking counterfactual dependence does not imply causal dependence and does not suffice for seeing

Example 12 The Screen I am hallucinating at random My hallucinations at any moment are determined by my precursor brain states a few seconds before My brain states are monitored, and my hallucinations are predicted by a fast computer It controls a battery of lights focused on a screen before my eyes in such a way that the scene before my eyes

[13] Causation , *Journal of Philosophy*, LXX (1973), pp 556–567

[14] This is circular in the context of a counterfactual analysis of causation, but in Counterfactual Dependence and Time s Arrow , *Nous*, XIII (1979), pp 455–476, I have proposed a way to distinguish backtrackers without the circular reference to causation, at least under determinism

is made to match my predicted visual experience at the time It is true in a sense—in the backtracking sense—that whatever might be on the screen, my visual experience would match it But my visual experience does not depend causally on the scene before my eyes Rather, they are independent effects of a common cause, namely my precursor brain states Therefore I do not see

The same example shows that it would not suffice just to require that the laws of nature and the prevailing conditions imply a suitable correspondence between visual experience and the scene before the eyes That could be so without the proper sort of counterfactual, and hence causal, dependence, in which case one would not see

X

The following case (Example 11 carried to extremes) is a hard one It closely resembles cases of genuine seeing, and we might well be tempted to classify it as such According to my analysis, however, it is a case of veridical hallucination The scene before the eyes causes matching visual experience without any pattern of counterfactual dependence whatever, suitable or otherwise

Example 13 The Censor My natural or prosthetic eye is in perfect condition and functioning normally, and by means of it the scene before my eyes causes matching visual experience But if the scene were any different my visual experience would be just the same For there is a censor standing by, ready to see to it that I have precisely that visual experience and no other, whatever the scene may be (Perhaps the censor is external, perhaps it is something in my own brain) So long as the scene is such as to cause the right experience, the censor does nothing But if the scene were any different, the censor would intervene and cause the same experience by other means If so, my eye would not function normally and the scene before my eyes would not cause matching visual experience

The case is one of causal preemption [15] The scene before my eyes is the actual cause of my visual experience the censor is an alternative potential cause of the same effect The actual cause preempts the poten-

[15] See my discussion of preemption in Causation

tial cause, stopping the alternative causal chain that would otherwise have gone to completion

The argument for classifying the case as seeing is that it is just like a clear case of seeing except for the presence of the censor, and, after all, the censor doesn't actually do anything, and if the scene before the eyes were different and the censor nevertheless stood idly by—as in actuality—then the different scene would indeed cause suitably different visual experience

My reply is that the case is really not so very much like the clear case of seeing to which it is compared The censor's idleness is an essential factor in the causal process by which matching visual experience is produced, just as the censor's intervention would be in the alternative process No such factor is present in the comparison case If the scene were different this factor would not be there, so it is wrong to hold it fixed in asking what would happen if the scene were different We cannot uniformly ignore or hold fixed those causal factors which are absences of interventions The standard process might be riddled with them (Think of a circuit built up from exclusive-or-gates every output signal from such a gate is caused partly by the absence of a second input signal) Who knows what would happen in an ordinary case of natural (or prosthetic) vision if the scene were different and all absences of interventions were held fixed? Who cares? We do not in general hold fixed the absences of intervention, and I see no good reason to give the censor's idleness special treatment

The decisive consideration, despite the misleading resemblance of this case to genuine cases of seeing, is that the censor's potential victim has no capacity at all to discriminate by sight Just as in any other case of veridical hallucination, the match that occurs is a lucky accident [16]

[16] I am grateful to seminar audiences at the University of Auckland, Victoria University of Wellington, the University of Sydney and Monash University for valuable comments on earlier versions of this paper and to the New Zealand-United States Educational Foundation and Monash University for making those seminars possible

Postscript to "Veridical Hallucination and Prosthetic Vision"

FURTHER CONSIDERATIONS ON "SUITABILITY"

As I noted, it is in several ways a matter of degree whether my condition for "a suitable pattern of counterfactual dependence" is satisfied That leaves room for borderline cases of seeing Among these borderline cases, some may be better than others I think there are further considerations that never—or hardly ever—make the difference between a clear negative case and a clear positive, but that do influence our judgments that one unclear case is more of a case of seeing than another The general principle is simple we know what happens in the ideal or normal case, and differences from that tend to detract from the claim of other cases to be judged positive

It is in this way, if at all, that considerations of mechanism are relevant I do not think they are ever decisive, or close to decisive, by themselves But they may tend to incline us one way or another in otherwise doubtful cases

A second consideration that might have some weight, but I think much less than decisive weight, comes if we have a probabilistic kind of causal dependence (See Postscript D to "Counterfactual Dependence and Time's Arrow," Postscript C to "Causation," and "Causal Decision Theory," all in this volume) Suppose that what we have are not counterfactuals saying that if there were such-and-such scene then there would definitely be such-and-such matching experience, but rather that there would be a chance distribution over experiences giving significant probability (and much more than there would have been without the scene) to matching experience Other things being equal, the better the chances of matching experience, the better case of seeing Here I have in mind the actual chance given the actual scene, as well as the chances there would be given other scenes *Ex hypothesi* the actual experience does match the actual scene And that is enough, if I am right that a counterfactual with a true antecedent is true iff its conse-

quent is, to give us a non-probabilistic counterfactual for that one, actual scene if there were that scene, there would be that matching experience But there will be a probabilistic counterfactual as well if there were that scene, there would be so-and-so chance of that matching experience, and maybe also some chance of various other experiences If the actual chance of match is substantially below one, then despite the non-probabilistic counterfactual, we have a consideration that detracts somewhat from the claim of the case to be judged positive

John Bigelow has suggested (in discussion, 1980) a third consideration call it the *Island Effect* There are good scenes that would produce matching experience, and bad scenes that would not An ideal pattern would have no bad scenes, but that is too much to demand, so I settled for the requirement that there be a wide range of good scenes Note that scenes may be close together or far apart, they may differ from one another more or less So a good scene might be surrounded by other good scenes, with no bad ones nearby The nearest bad scene to it might differ quite substantially Or at the opposite extreme it might be a tiny island, surrounded by a sea of bad scenes Suppose the actual good scene is such an island My requirement that there be a wide range of good scenes may be satisfied only in virtue of some distant continent (Or in virtue of many other islands, widely scattered) It's a narrow escape the subject sees, on my analysis, but had the scene been just a little different then he wouldn't have done For any scene just a little different would have been a bad scene To make matters still worse, Bigelow considers the case that any nearby scene not only would have been bad, but also would have produced just the same experience that the actual good scene produces Within limits—the distance to the next land—a different scene would have made no difference to visual experience Then does the subject see?

One might go so far as to think that extreme cases of the Island Effect are clear cases of not seeing, even if there is nothing else wrong with them I disagree, but I certainly think that the Island Effect influences comparative judgements about unclear cases, even that it suffices to turn what would otherwise be clearly positive cases into doubtful ones

I said that these secondary considerations would never turn a clear positive into a clear negative—or hardly ever What if all three of the secondary considerations were present to an extreme degree, working together? What if there is an abnormal mechanism, and in addition the chance of matching experience given the actual scene is quite low,

though as it happens there is matching experience, and in addition the actual scene is a tiny island, and my requirement of counterfactual dependence through a wide range of scenes is satisfied only by means of scenes quite different from the actual one? How would we judge that case? It satisfies my main conditions, but the secondary considerations go against it as powerfully as can be

I know how I *did* judge such a case I judged it negative And so perhaps did you, if you read my paper before reading this postscript For it is none other than my example of the wizard I thought it a clear negative case, and cited it in my favor, without ever noticing that my own conditions classified it as positive![1] For the actual scene with the hallucinogenic wizard *does* cause matching experience,[2] and we *do* have a wide range of alternative scenes—namely, ordinary scenes without the wizard—that would cause matching experience in a normal way (Compare scenes without Martians in my example of the hypnotic suggestion)

So the secondary considerations *can* have decisive weight, if they all push together as hard as they can Nothing less would do, I think I would not judge the example of the wizard negative if the spell left the normal mechanism in operation but increased its rate of random errors and thus drastically lowered its probability of success, or if the wizard's presence produced matching experience for that particular scene with high probability, though not by the standard mechanism, or if the scene with the wizard were in the midst of other scenes that would somehow, with significant probability, also produce matching experience

Bruce LeCatt[3] has suggested a consideration that tends in the positive direction *Stepwise Dependence* Take some intermediate stage in the causal process that leads from scene to visual experience It may be that there is a good pattern of counterfactual dependence whereby what goes on at the intermediate stage depends on the scene, and also a

[1] My mistake was pointed out to me by Cliff Landesman in 1984

[2] You might wonder whether the presence of the spell casting wizard really *causes* the matching experience, when the probability of matching experience would have been much better without him Yes, the probability that there would somehow have been a match would have been much better But the probability of *this* experience would have been much lower, and that is what makes it so that the scene causes this experience, an experience which is in fact matching Do not say the scene causes the experience *qua* experience of such-and such, but not *qua* matching no such distinction is part of our concept of causation I take it that causation relative to descriptions is a philosophers invention, motivated by a misguided deductive nomological analysis of causation

[3] Censored Vision, *Australasian Journal of Philosophy* 60 (1982) 158–62

good pattern of counterfactual dependence whereby visual experience depends on what goes on at the intermediate stage Further, this two-fold pattern might link scenes indirectly with matching experience, over a suitably wide and varied range of scenes Even more indirectly, there might be linkage via a threefold pattern of counterfactual dependence involving two intermediate stages, and so on Then we have a suitable pattern of stepwise counterfactual dependence of visual experience on the scene before the eyes It does not follow that we have a suitable pattern of counterfactual dependence *simpliciter*, because counterfactuals are not necessarily transitive [4] In fact my case of the censor is a case of excellent stepwise dependence and no dependence *simpliciter* at all LeCatt suggests, and I agree, that it is the stepwise dependence that accounts for any inclinations we have to judge the case of the censor as a positive case of seeing He further claims that this judgment is correct, but there I do not agree, and I insist that the essential feature of seeing is altogether missing

But there are mixed cases partial or conditional censorship, some dependence *simpliciter* but not much compared with normal cases Then indeed the presence of stepwise dependence might make the difference between better cases and worse

[4] See Counterfactuals and Comparative Possibility and Causation, in this volume

TWENTY-FIVE

Are We Free To Break the Laws?

Soft determinism seems to have an incredible consequence It seems to imply, given certain acceptable further premises, that sometimes we are able to act in such a way that the laws of nature are broken But if we distinguish a strong and a weak version of this incredible consequence, I think we shall find that it is the strong version that is incredible and the weak version that is the consequence

Soft determinism is the doctrine that sometimes one freely does what one is predetermined to do, and that in such a case one is able to act otherwise though past history and the laws of nature determine that one will not act otherwise

Compatibilism is the doctrine that soft determinism may be true A compatibilist might well doubt soft determinism because he doubts on physical grounds that we are ever predetermined to act as we do, or perhaps because he doubts on psychoanalytic grounds that we ever act freely I myself am a compatibilist but no determinist, hence I am obliged to rebut some objections against soft determinism but not others But for the sake of the argument, let me feign to uphold soft determinism, and indeed a particular instance thereof

I have just put my hand down on my desk That, let me claim, was a free but predetermined act I was able to act otherwise, for instance to raise my hand, But there is a true historical proposition *H* about the intrinsic state of the world long ago, and there is a true proposition *L*

specifying the laws of nature that govern our world, such that H and L jointly determine what I did They jointly imply the proposition that I put my hand down They jointly contradict the proposition that I raised my hand Yet I was free, I was able to raise my hand The way in which I was determined not to was not the sort of way that counts as inability

What if I had raised my hand? Then at least one of three things would have been true Contradictions would have been true together, or the historical proposition H would not have been true, or the law proposition L would not have been true Which? Here we need auxiliary premises, but since I accept the premises my opponent requires to make his case, we may proceed Of our three alternatives, we may dismiss the first, for if I had raised my hand, there would still have been no true contradictions Likewise we may dismiss the second, for if I had raised my hand, the intrinsic state of the world long ago would have been no different [1] That leaves the third alternative If I had raised my hand, the law proposition L would not have been true That follows by a principle of the logic of counterfactuals which is almost uncontroversial [2]

$$A \mathbin{\square\!\!\rightarrow} B \vee C \vee D,\ A \mathbin{\square\!\!\rightarrow} -B,\ A \mathbin{\square\!\!\rightarrow} -C, \quad A \mathbin{\square\!\!\rightarrow} D$$

If L had not been true, that implies that some law of nature would have been broken, for L is a specification of the laws That is not to say that anything would have been both a law and broken—that is a contradiction in terms if, as I suppose, any genuine law is at least an absolutely unbroken regularity Rather, if L had not been true, something that is in fact a law, and unbroken, would have been broken, and no law It would at best have been an almost-law

In short, as a (feigned) soft determinist, who accepts the requisite auxiliary premises and principle of counterfactual logic, I am committed to the consequence that if I had done what I was able to do—raise my hand—then some law would have been broken

"That is to say," my opponent paraphrases, "you claim to be able to break the very laws of nature And with so little effort! A marvelous power indeed! Can you also bend spoons?"

Distinguo My opponent's paraphrase is not quite right He has replaced the weak thesis that I accept with a stronger thesis that I join him in rejecting The strong thesis is utterly incredible, but it is no part

[1] I argue for this in [4]

[2] The inference is valid in any system that treats the conditional as a propositionally (or even sententially) indexed family of normal necessities, in the sense of Brian F Chellas ([1])

of soft determinism The weak thesis is controversial, to be sure, but a soft determinist should not mind being committed to it The two theses are as follows

(Weak Thesis) I am able to do something such that, if I did it, a law would be broken

(Strong Thesis) I am able to break a law

To see the difference, consider not a marvelous ability to break a law but a commonplace ability to break a window Perhaps I am able to throw a stone in a certain direction, and perhaps if I did, the stone would hit a certain window and the window would break Then I am able to break a window For starters I am able to do something such that, if I did it, a window would be broken But there is more to be said I am able to do something such that, if I did it, my act would cause a window-breaking event

Or consider a commonplace ability to break a promise Perhaps I am able to throw a stone, and perhaps if I did, I would break my promise never to throw a stone Then I am able to break a promise For starters I am able to do something such that, if I did it, a promise would be broken But there is more to be said I am able to do something such that, if I did it, my act would itself be a promise-breaking event

Next, consider what really would be a marvelous ability to break a law—an ability I could not credibly claim Suppose that I were able to throw a stone very, very hard And suppose that if I did, the stone would fly faster than light, an event contrary to law Then I really would be able to break a law For starters I would be able to do something such that, if I did it, a law would be broken But there is more to be said I would be able to do something such that, if I did it, my act would cause a law-breaking event

Or suppose that I were able to throw a stone so hard that in the course of the throw my own hand would move faster than light Then again I would be able to break a law, regardless of what my act might cause For starters I would be able to do something such that, if I did it, a law would be broken But there is more to be said I would be able to do something such that, if I did it, my act would itself be a law-breaking event

If no act of mine either caused or was a window-, promise-, or law-breaking event, then I think it could not be true that I broke a window, a promise, or a law Therefore I am able to break a window, a promise, or a law only if I am able to do something such that, if I did it, my act

either would cause or would be a window-, promise-, or law-breaking event

Maybe my opponent will contend that according to soft determinism, there is another way of being able to break a law But I see no reason to grant his contention

Now consider the disputed case I am able to raise my hand, although it is predetermined that I will not If I raised my hand, some law would be broken I even grant that a law-breaking event would take place (Here I use the present tense neutrally I mean to imply nothing about *when* a law-breaking event would take place) But is it so that my act of raising my hand would cause any law-breaking event? Is it so that my act of raising my hand would itself be a law-breaking event? Is it so that any other act of mine would cause or would be a law-breaking event? If not, then my ability to raise my hand confers no marvelous ability to break a law, even though a law would be broken if I did it [3]

Had I raised my hand, a law would have been broken beforehand The course of events would have diverged from the actual course of events a little while before I raised my hand, and at the point of divergence there would have been a law-breaking event—a divergence miracle, as I have called it ([4]) But this divergence miracle would not have been caused by my raising my hand If anything, the causation would have been the other way around Nor would the divergence miracle have been my act of raising my hand That act was altogether absent from the actual course of events, so it cannot get under way until there is already some divergence Nor would it have been caused by any other act of mine, earlier or later Nor would it have been any other act of mine Nor is there any reason to say that if I had raised my hand there would have been some other law-breaking event besides the

[3] Up to a point, my strategy here resembles that of Keith Lehrer ([2], p 199) Lehrer grants a weak thesis the agent could have done something such that if he had done it, there would have been a difference in either laws or history He rejects as I would, the step from that to a stronger thesis the agent could have brought about a difference in laws or history So far so good But Lehrer s reason for rejecting the stronger thesis is one I cannot accept His reason is this it is false that if the agent had preferred that there be a difference in laws or history there would have been a difference in laws or history I say, first, that this conditional may not be false Suppose the agent is predetermined to prefer that there be no difference had he preferred otherwise, there would have been a difference (Had anything been otherwise than it was predetermined to be, there would have been a difference in either laws or history) And second, if this conditional is not false, that is not enough to make the stronger thesis true There must be some other reason different from the one Lehrer gives, why the stronger thesis is false

divergence miracle, still less, that some other law-breaking event would have been caused by, or would have been, my act of raising my hand To accommodate my hypothetical raising of my hand while holding fixed all that can and should be held fixed, it is necessary to suppose one divergence miracle, gratuitous to suppose any further law-breaking

Thus I insist that I was able to raise my hand, and I acknowledge that a law would have been broken had I done so, but I deny that I am therefore able to break a law To uphold my instance of soft determinism, I need not claim any incredible powers To uphold the compatibilism that I actually believe, I need not claim that such powers are even possible

I said that if I had raised my hand, the divergence miracle beforehand would not have been caused by my raising my hand That seems right But my opponent might argue *ad hominem* that according to my own analysis of causation ([3]), my raising my hand does turn out to cause the divergence miracle The effect would precede the cause, but I do not object to that We seem to have the right pattern of counterfactual dependence between distinct events (1) if I had raised my hand, the divergence miracle would have occurred, but (2) if I had not raised my hand, it would not have occurred

I reply that we do not have this required pattern, nor would we have had it if I had raised my hand Therefore I am safe in denying that the miracle would have been caused by my act

We do not have the pattern because (1) is false What is true is only that if I had raised my hand, then some or other divergence miracle would have occurred There is no particular divergence miracle that definitely would have occurred, since the divergence might have happened in various ways [4]

If I had raised my hand, (1) would have been true But we still would not have had the right pattern, because in that case (2) would have been false Consider a counterfactual situation in which a divergence miracle beforehand has allowed me to raise my hand Is it so, from the standpoint of that situation, that if I had not raised my hand, the miracle

[4] Cf [4], p 463 At this point I am relying on contingent features of the world as we suppose it to be, as Allen Hazen has pointed out to me, we can imagine a world of discrete processes at which one divergent history in which I raise my hand clearly takes less of a miracle than any of its rivals I think this matters little, since the task of compatibilism is to show how freedom and determinism might coexist at a world that might for all we know, be ours

would not have taken place? No, the miracle might have taken place, only to have its work undone straightway by a second miracle (Even in this doubly counterfactual context, when I speak of a miracle I mean a violation of the actual laws.) What is true, at most, is that if I had not raised my hand, then the first miracle might not have taken place.

My incompatibilist opponent is a creature of fiction, but he has his prototypes in real life. He is modelled partly on Peter van Inwagen ([5], [6], [7]) and partly on myself when I first worried about van Inwagen's argument against compatibilism. He definitely is not van Inwagen, he does not choose his words so carefully. Still I think that for all his care, van Inwagen is in the same boat with my fictitious opponent.

Van Inwagen's argument runs as follows, near enough. (I recast it as a *reductio* against the instance of soft determinism that I feign to uphold.) I did not raise my hand, suppose for *reductio* that I could have raised my hand, although determinism is true. Then it follows, given four premises that I cannot question, that I could have rendered false the conjunction HL of a certain historical proposition H about the state of the world before my birth and a certain law proposition L. If so, then I could have rendered L false. (Premise 5.) But I could not have rendered L false. (Premise 6.) This refutes our supposition.

To this I reply that Premise 5 and Premise 6 are not both true. Which one is true depends on what van Inwagen means by "could have rendered false."

It does not matter what "could have rendered false" means in ordinary language; van Inwagen introduced the phrase as a term of art. It does not even matter what meaning van Inwagen gave it. What matters is whether we can give it any meaning that would meet his needs—any meaning that would make all his premises defensible without circularity. I shall consider two meanings. I think there is nothing in van Inwagen's text to suggest any third meaning that might work better than these two.[5]

[5] Van Inwagen has indicated (personal communication, 1981) that he would adopt a third meaning for "could have rendered false," different from both of the meanings that I discuss here. His definition is roughly as follows: an agent could have rendered a proposition false iff he could have arranged things in a certain way, such that his doing so, plus the whole truth about the past, together strictly imply the falsehood of the proposition. On this definition, Premise 6 simply says that I could not have arranged things in any way such that I was predetermined not to arrange things in that way. It is uninstructive to learn that the soft determinist is committed to denying Premise 6 thus understood.

First, a preliminary definition Let us say that an event would falsify a proposition iff, necessarily, if that event occurs then that proposition is false For instance, an event consisting of a stone's flying faster than light would falsify a law So would an act of throwing in which my hand moves faster than light So would a divergence miracle But my act of throwing a stone would not itself falsify the proposition that the window in the line of fire remains intact, all that is true is that my act would cause another event that would falsify that proposition My act of raising my hand would falsify any sufficiently inclusive conjunction of history and law But it would not itself falsify any law—not if all the requisite law-breaking were over and done with beforehand All that is true is that my act would be preceded by another event—the divergence miracle—that would falsify a law

Let us say that I could have rendered a proposition false in the weak sense iff I was able to do something such that, if I did it, the proposition would have been falsified (though not necessarily by my act, or by any event caused by my act) And let us say that I could have rendered a proposition false in the strong sense iff I was able to do something such that, if I did it, the proposition would have been falsified either by my act itself or by some event caused by my act

The Weak Thesis, which as a soft determinist I accept, is the thesis that I could have rendered a law false in the weak sense The Strong Thesis, which I reject, is the thesis that I could have rendered a law false in the strong sense

The first part of van Inwagen's argument succeeds whichever sense we take If I could have raised my hand despite the fact that determinism is true and I did not raise it, then indeed it is true both in the weak sense and in the strong sense that I could have rendered false the conjunction *HL* of history and law But I could have rendered false the law proposition *L* in the weak sense, though I could not have rendered *L* false in the strong sense So if we take the weak sense throughout the argument, then I deny Premise 6 If instead we take the strong sense, then I deny Premise 5

Van Inwagen supports both premises by considering analogous cases I think the supporting arguments fail because the cases produced are not analogous they are cases in which the weak and strong senses do not diverge In support of Premise 6, he invites us to reject the supposition that a physicist could render a law false by building and operating a machine that would accelerate protons to twice the speed of light Reject that supposition by all means, but that does nothing to support Premise 6 taken in the weak sense, for the rejected supposition

is that the physicist could render a law false in the strong sense In support of Premise 5, he invites us to reject the supposition that a traveler could render false a conjunction of a historical proposition and a proposition about his future travels otherwise than by rendering false the nonhistorical conjunct Reject that supposition by all means, but that does nothing to support Premise 5 taken in the strong sense Given that one could render false, in the strong sense, a conjunction of historical and nonhistorical propositions (and given that, as in the cases under consideration, there is no question of rendering the historical conjunct false by means of time travel or the like), what follows? Does it follow that one could render the nonhistorical conjunct false in the strong sense? That is what would support Premise 5 in the strong sense Or does it only follow, as I think, that one could render the nonhistorical conjunct false in at least the weak sense? The case of the traveler is useless in answering that question, since if the traveler could render the proposition about his future travels false in the weak sense, he could also render it false in the strong sense

REFERENCES

[1] CHELLAS, B F ' Basic conditional logic " *Journal of philosophical logic*, vol 4 (1975), pp 133–153

[2] LEHRER, K "Preferences, conditionals and freedom " In Peter van Inwagen, ed , *Time and cause* Dordrecht Reidel, 1980

[3] LEWIS, D "Causation " *Journal of Philosophy*, vol 70 (1973), pp 556–567

[4] LEWIS, D "Counterfactual dependence and time's arrow " *Nous*, vol 13 (1979), pp 455–476

[5] VAN INWAGEN, P "A formal approach to the problem of free will and determinism " *Theoria*, vol 40 (1974), pp 9–22

[6] VAN INWAGEN, P 'The incompatibility of free will and determinism " *Philosophical studies*, vol 27 (1975), pp 185–199

[7] VAN INWAGEN, P "Reply to Narveson *Philosophical studies*, vol 32 (1977), pp 89–98

TWENTY-SIX

Prisoners' Dilemma Is a Newcomb Problem

Several authors have observed that Prisoners' Dilemma and Newcomb's Problem are related—for instance, in that both involve controversial appeals to dominance [1] But to call them "related" is an understatement Considered as puzzles about rationality, or disagreements between two conceptions thereof, they are one and the same problem Prisoners' Dilemma *is* a Newcomb Problem—or rather, two Newcomb Problems side by side, one per prisoner Only the inessential trappings are different Let us make them the same

You and I, the "prisoners," are separated Each is offered the choice to rat or not to rat (The action of "ratting" is so called because I consider it to be *rat*ional—but that is controversial) Ratting is done as follows one reaches out and takes a transparent box, which is seen to contain a thousand dollars A prisoner who rats gets to keep the thousand (Maybe ratting is construed as an act of confessing and accusing one's partner, much as taking the Queen's shilling was once

[1] Robert Nozick, Newcomb s Problem and Two Principles of Choice, in *Essays in Honor of Carl G Hempel*, ed N Rescher (Dordrecht Reidel, 1969), pp 130–131, Steven J Brams, Newcomb s Problem and Prisoners Dilemma, *Journal of Conflict Resolution* 19 (1975) 596–612 Lawrence H Davis, Prisoners, Paradox, and Rationality, *American Philosophical Quarterly* 14 (1977) 321 and J Howard Sobel, *Chance, Choice, and Action Newcomb s Problem Resolved* (duplicated manuscript, July 1978), pp 167–168

construed as an act of enlisting—but that is irrelevant to the decision problem.) If either prisoner declines to rat, he is not at all rewarded, but his partner is presented with a million dollars, nicely packed in an opaque box. (Maybe each faces a long sentence and a short sentence to be served consecutively, escape from the long sentence costs a million, and escape from the short sentence costs a thousand. But it is irrelevant how the prisoners propose to spend their money.) So the payoff matrix looks like this:

	I rat	I don't rat
You rat	I get $1,000 You get $1,000	I get $0 You get $1,001,000
You don't rat	I get $1,001,000 You get $0	I get $1,000,000 You get $1,000,000

There we have it: a perfectly typical case of Prisoners' Dilemma. My decision problem, in a nutshell, is as follows, yours is exactly similar:

(1) I am offered a thousand—take it or leave it.
(2) Perhaps also I will be given a million, but whether I will or not is causally independent of what I do now. Nothing I can do now will have any effect on whether or not I get my million.
(3) I will get my million if and only if you do not take your thousand.

Newcomb's Problem is the same as regards points (1) and (2). The only difference—if such it be—is that point (3) is replaced by

(3′) I will get my million if and only if it is predicted that I do not take my thousand.

"Predicted" need not mean "predicted in advance." Not so in English: we credit new theories with success in "predicting" phenomena already observed. And not so in Newcomb's Problem. While it dramatizes the problem to think of the million *already there*, or else already not there, in the opaque box in front of me as I deliberate, it is agreed all around that what really matters is (2), and hence that the "predic-

tion" should be causally independent of my decision Making the prediction ahead of time is one good way to secure this causal independence But it is not the only way [2] Provided that I can have no effect on it, the prediction could just as well be made simultaneously with my decision or even afterwards, and the character of Newcomb's Problem would be unchanged [3] Likewise in the case of Prisoners' Dilemma nothing need be assumed—and in my telling of the story, nothing was assumed—about whether the prisoners are put to the test simultaneously or one after the other

Also it is inessential to Newcomb's Problem that any prediction—in advance, or otherwise—should actually take place It is enough that some potentially predictive process should go on, and that whether I get my million is somehow made to depend on the outcome of that process It could all be automated if the predictive computer sends a pulse of current to the money-putting machine I get my million, otherwise not Or there might be people who put the million in the box or not depending on the outcome of the process, but who do not at all think of the outcome as a prediction of my choice, or as warrant for a prediction It makes no difference to my decision problem whether someone—the one who gives the million, or perhaps some bystander—does or doesn't form beliefs about what I will do by inference from the outcome of the predictive process

Eliminating inessentials, then, Newcomb's Problem is characterized by (1), (2), and

(3″) I will get my million if and only if a certain potentially predictive process (which may go on before, during, or after my choice) yields the outcome which could warrant a prediction that I do not take my thousand

The potentially predictive process *par excellence* is *simulation* To predict whether I will take my thousand, make a replica of me, put my replica in a replica of my predicament, and see whether my replica takes *his* thousand And whether or not anybody actually makes a prediction about me by observing my replica, still my replica's decision is a potentially predictive process with respect to mine Disregarding predictive processes other than simulation, if such there be, we have this special case of (3″)

[2] And perhaps not an infallible way See David Lewis, The Paradoxes of Time Travel, *American Philosophical Quarterly* 13 (1976) 145–152

[3] That is noted by Nozick, Newcomb s Problem, p 132, and I have not seen it disputed

(3′′′) I will get my million if and only if my replica does not take his thousand

There are replicas and replicas Some are the same sort of thing that I am, others are less so A flesh-and-blood duplicate made by copying me atom for atom would be one good sort of replica A working scale model of me, smaller perhaps by a ratio of 1 148, also might serve So might a pattern of bits in a computer, or beads on an abacus, or marks on paper, or neuron firings in a brain, even though these things are unlike me and replicate me only by way of some complicated isomorphism

Also, some replicas are more reliable than others There may be grounds for greater or lesser degrees of confidence that my replica and I will decide alike in the matter of the thousand A replica that matches me perfectly in the respects relevant to my decision (whether duplicate or isomorph) will have more predictive power than a less perfect replica, but even a poor replica may have some significant degree of predictive power

As Newcomb's Problem is usually told, the predictive process involved is extremely reliable But that is inessential The disagreement between conceptions of rationality that gives the problem its interest arises even when the reliability of the process, as estimated by the agent, is quite poor—indeed, even when the agent judges that the predictive process will do little better than chance More precisely, define *average estimated reliability* as the average of (A) the agent's conditional degree of belief that the predictive process will predict correctly, given that he takes his thousand, and (B) his conditional degree of belief that the process will predict correctly, given that he does not take his thousand (When the predictive process is a simulation, for instance, we have the average of two conditional degrees of belief that the agent and his replica will decide alike) Let r be the ratio of the value of the thousand to the value of the million 001 if value is proportional to money, perhaps somewhat more under diminishing marginal value We have a disagreement between two conceptions of rationality if and only if the expected value[4] of taking the thousand is less than that of declining it, which is so if and only if the average estimated reliability exceeds $\frac{(1 + r)}{2}$ (That is 5005 if value is pro-

[4] As calculated according to the non-causal sort of decision theory presented for instance in Richard Jeffrey, *The Logic of Decision* (New York McGraw Hill 1965)

portional to money) This is not a very high standard of reliability So there can be a fully problematic case of Newcomb's Problem in which the predictive process consists of simulation by some very imperfect and very unreliable replica

The most readily available sort of replica of me is simply another person, placed in a replica of my predicament For instance you, my fellow prisoner Most likely you are not a very exact replica of me, and your choice is not a very reliable predictive process for mine [5] Still, you might well be reliable enough (in my estimation) for a Newcomb Problem [6] So we have this special case of (3′′′)

(3) I will get my million if and only if you do not take your thousand

Inessential trappings aside, Prisoners' Dilemma is a version of Newcomb's Problem, *quod erat demonstrandum*

Some who discuss Newcomb's Problem think it is rational to decline the thousand if the predictive process is reliable enough Their reason is that they believe, justifiably, that those who decline their thousands will probably get their millions Some who discuss Prisoners' Dilemma think it is rational not to rat if the two partners are enough alike [7] Their reason is that they believe, justifiably, that those who do not rat will probably not be ratted on by their like-thinking partners These two opinions are one opinion in two guises

But some—I, for one—who discuss Newcomb's Problem think it is rational to take the thousand no matter how reliable the predictive process may be Our reason is that one thereby gets a thousand more than he would if he declined, since he would get his million or not regardless of whether he took his thousand And some—I, for one—who discuss Prisoners' Dilemma think it is rational to rat no matter how much

[5] On the other hand you might be an extremely perfect and reliable replica as in the Prisoners Dilemma between twins described by Nozick, Newcomb s Problem, pp 130–131

[6] If you do not meet even the low standard of estimated reliability just considered, either because you are unlike me or because you and I alike are apt to choose at random or because the payoffs are such as to set *r* rather high, then we have a situation with no clash between conceptions of rationality on *any* conception, it is rational to rat But even this non-problem might legitimately be called a version of Newcomb s Problem, since it satisfies conditions (1), (2), and (3″)

[7] For instance Davis, Prisoners, Paradox, and Rationality He considers the case in which the partners are alike because they are both rational, but there is also the case where they are alike because they are given to the same sorts of irrationality

alike the two partners may be, and no matter how certain they may be that they will decide alike Our reason is that one is better off if he rats than he would be if he didn't, since he would be ratted on or not regardless of whether he ratted These two opinions also are one

Some have fended off the lessons of Newcomb's Problem by saying "Let us not have, or let us not rely on, any intuitions about what is rational in goofball cases so unlike the decision problems of real life " But Prisoners' Dilemmas are deplorably common in real life They are the most down-to-earth versions of Newcomb's Problem now available

TWENTY-SEVEN

Causal Decision Theory

1 INTRODUCTION

Decision theory in its best-known form[1] manages to steer clear of the thought that what's best to do is what the agent believes will most tend to cause good results Causal relations and the like go unmentioned The theory is simple, elegant, powerful, and conceptually economical Unfortunately it is not quite right In a class of somewhat peculiar cases, called Newcomb problems, this noncausal decision theory gives the wrong answer It commends an irrational policy of managing the news so as to get good news about matters which you have no control over

I am one of those who have concluded that we need an improved decision theory, more sensitive to causal distinctions Noncausal decision theory will do when the causal relations are right for it, as they very often are, but even then the full story is causal Several versions of causal decision theory are on the market in the works of

[1] As presented, for instance, in Richard C Jeffrey, *The Logic of Decision* (New York McGraw Hill, 1965)

Gibbard and Harper, Skyrms, and Sobel,[2] and I shall put forward a version of my own. But also I shall suggest that we causal decision theorists share one common idea, and differ mainly on matters of emphasis and formulation. The situation is not the chaos of disparate approaches that it may seem.

Of course there are many philosophers who understand the issues very well, and yet disagree with me about which choice in a Newcomb problem is rational. This paper is about a topic that does not arise for them. Noncausal decision theory meets their needs and they want no replacement. I will not enter into debate with them, since that debate is hopelessly deadlocked and I have nothing new to add to it. Rather, I address myself to those who join me in presupposing that Newcomb problems show the need for some sort of causal decision theory, and in asking what form that theory should take.

2 PRELIMINARIES: CREDENCE, VALUE, OPTIONS

Let us assume that a (more or less) rational agent has, at any moment, a *credence* function and a *value* function. These are defined in the first instance over single possible worlds. Each world W has a credence $C(W)$, which measures the agent's degree of belief that W is the actual world. These credences fall on a scale from zero to one, and they sum to one. Also each world W has a value $V(W)$, which measures how satisfactory it seems to the agent for W to be the actual world. These values fall on a linear scale with arbitrary zero and unit.

We may go on to define credence also for sets of worlds. We call such sets *propositions*, and we say that a proposition *holds* at just those worlds which are its members. I shall not distinguish in notation between a world W and a proposition whose sole member is W, so all that is said of propositions shall apply also to single worlds. We sum credences for any proposition X,

$$C(X) =^{df} \Sigma_{W \in X} C(W)$$

[2] Allan Gibbard and William Harper, "Counterfactuals and Two Kinds of Expected Utility", in C. A. Hooker, J. J. Leach, and E. F. McClennen, eds., *Foundations and Applications of Decision Theory*, Volume 1 (Dordrecht, Holland: D. Reidel, 1978), Brian Skyrms, "The Role of Causal Factors in Rational Decision", in his *Causal Necessity* (New Haven: Yale University Press, 1980), and Jordan Howard Sobel, *Probability, Chance, and Choice: A Theory of Rational Agency* (unpublished, presented in part at a workshop on Pragmatics and Conditionals at the University of Western Ontario in May 1978).

We define conditional credences as quotients of credences, defined if the denominator is positive

$$C(X/Y) =^{df} C(XY)/C(Y),$$

where XY is the conjunction (intersection) of the propositions X and Y. If $C(Y)$ is positive, then $C(-/Y)$, the function that assigns to any world W or proposition X the value $C(W/Y)$ or $C(X/Y)$, is itself a credence function. We say that it *comes from* C *by conditionalising on Y*. Conditionalising on one's total evidence is a rational way to learn from experience. I shall proceed on the assumption that it is the only way for a fully rational agent to learn from experience, however, nothing very important will depend on that disputed premise.

We also define (expected) value for propositions. We take credence-weighted averages of values of worlds: for any proposition X,

$$V(X) =^{df} \Sigma_W C(W/X)V(W) = \Sigma_{W \varepsilon X} C(W)V(W)/C(X)$$

A *partition* (or a *partition of X*) is a set of propositions of which exactly one holds at any world (or at any X-world). Let the variable Z range over any partition (in which case the XZ's, for fixed X and varying Z, are a partition of X). Our definitions yield the following *Rules of Additivity* for credence, and for the product of credence and expected value:

(1) $C(X) = \Sigma_Z C(XZ)$,
$C(X)V(X) = \Sigma_Z C(XZ)V(XZ)$

This *Rule of Averaging* for expected values follows:

(2) $V(X) = \Sigma_Z C(Z/X)V(XZ)$

Thence we can get an alternative definition of expected value. For any number v, let [V=v] be the proposition that holds at just those worlds W for which $V(W)$ equals v. Call [V=v] a *value-level proposition*. Since the value-level propositions are a partition,

(3) $V(X) = \Sigma_v C([V=v]/X)v$

I have idealised and oversimplified in three ways, but I think the dodged complications make no difference to whether, and how, decision theory ought to be causal. First, it seems most unlikely that any real person could store and process anything so rich in information as the C and V functions envisaged. We must perforce make do with summaries. But it is plausible that someone who really did have these functions to guide him would not be so very different from us in his conduct, apart from his supernatural prowess at logic and mathematics and *a priori* knowledge generally. Second, my formulation makes

straightforward sense only under the fiction that the number of possible worlds is finite There are two remedies We could reformulate everything in the language of standard measure theory, or we could transfer our simpler formulations to the infinite case by invoking nonstandard summations of infinitesimal credences Either way the technicalities would distract us, and I see little risk that the fiction of finitude will mislead us Third, a credence function over possible worlds allows for partial beliefs about the way the world is, but not for partial beliefs about who and where and when in the world one is Beliefs of the second sort are distinct from those of the first sort, it is important that we have them, however, they are seldom very partial To make them partial we need either an agent strangely lacking in self-knowledge, or else one who gives credence to strange worlds in which he has close duplicates I here ignore the decision problems of such strange agents [3]

Let us next consider the agent's options Suppose we have a partition of propositions that distinguish worlds where the agent acts differently (he or his counterpart, as the case may be) Further, he can act at will so as to make any one of these propositions hold, but he cannot act at will so as to make any proposition hold that implies but is not implied by (is properly included in) a proposition in the partition The partition gives the most detailed specifications of his present action over which he has control Then this is the partition of the agents' alternative *options* [4] (Henceforth I reserve the variable *A* to range over these options) Say that the agent *realises* an option iff he acts in such a way as to make it hold Then the business of decision theory is to say which of the agent's alternative options it would be rational for him to realise

All this is neutral ground Credence, value, and options figure both in noncausal and in causal decision theory, though of course they are put to somewhat different uses

3 NONCAUSAL DECISION THEORY

Noncausal decision theory needs no further apparatus It prescribes the rule of V-maximising, according to which a rational choice is one

[3] I consider them in Attitudes *De Dicto* and *De Se* , *The Philosophical Review*, 88 (1979) pp 513–543, especially p 534 There, however, I ignore the causal aspects of decision theory I trust there are no further problems that would arise from merging the two topics

[4] They are his narrowest options Any proposition implied by one of them might be called an option for him in a broader sense, since he could act at will so as to make it hold But when I speak of options, I shall always mean the narrowest options

that has the greatest expected value An option A is V-*maximal* iff $V(A)$ is not exceeded by any $V(A')$, where A' is another option The theory says that to act rationally is to realise some V-maximal option

Here is the guiding intuition How would you like to find out that A holds? Your estimate of the value of the actual world would then be $V(A)$, if you learn by conditionalising on the news that A So you would like best to find out that the V-maximal one of the A's holds (or one of the V-maximal ones, in case of a tie) But it's in your power to find out that whichever one you like holds, by realising it So go ahead—find out whichever you'd like best to find out! You make the news, so make the news you like best

This seeking of good news may not seem so sensible, however, if it turns out to get in the way of seeking good results And it does

4 NEWCOMB PROBLEMS

Suppose you are offered some small good, take it or leave it Also you may suffer some great evil, but you are convinced that whether you suffer it or not is entirely outside your control In no way does it depend causally on what you do now No other significant payoffs are at stake Is it rational to take the small good? Of course, say I

I think enough has been said already to settle that question, but there is some more to say Suppose further that you think that some prior state, which may or may not obtain and which also is entirely outside your control, would be conducive both to your deciding to take the good and to your suffering the evil So if you take the good, that will be evidence that the prior state does obtain and hence that you stand more chance than you might have hoped of suffering the evil Bad news! But is that any reason not to take the good? I say not, since if the prior state obtains, there's nothing you can do about it now In particular, you cannot make it go away by declining the good, thus acting as you would have been more likely to act if the prior state had been absent All you accomplish is to shield yourself from the bad news That is useless (*Ex hypothesi*, dismay caused by the bad news is not a significant extra payoff in its own right Neither is the exhilaration or merit of boldly facing the worst) To decline the good lest taking it bring bad news is to play the ostrich

The trouble with noncausal decision theory is that it commends the ostrich as rational Let G and $-G$ respectively be the propositions that you take the small good and that you decline it, suppose for simplicity that just

these are your options Let E and $-E$ respectively be the propositions that you suffer the evil and that you do not Let the good contribute g to the value of a world and let the evil contribute $-e$, suppose the two to be additive, and set an arbitrary zero where both are absent Then by Averaging,

$$(4)\quad V(-G) = C(E/-G)V(E-G) + C(-E/-G)V(-E-G) = -eC(E/-G)$$
$$V(G) = C(E/G)V(EG) + C(-E/G)V(-EG) = -eC(E/G) + g$$

That means that $-G$, declining the good, is the V-maximal option iff the difference $(C(E/G) - C(E/-G))$, which may serve as a measure of the extent to which taking the good brings bad news, exceeds the fraction g/e And that may well be so under the circumstances considered If it is, noncausal decision theory endorses the ostrich's useless policy of managing the news It tells you to decline the good, though doing so does not at all tend to prevent the evil If a theory tells you that, it stands refuted

In Newcomb's original problem,[5] verisimilitude was sacrificed for extremity $C(E/G)$ was close to one and $C(E/-G)$ was close to zero, so that declining the good turned out to be V-maximal by an overwhelming margin To make it so, we have to imagine someone with the mind-boggling power to detect the entire vast combination of causal factors at some earlier time that would cause you to decline the good, in order to inflict the evil if any such combination is present Some philosophers have refused to learn anything from such a tall story

If our aim is to show the need for causal decision theory, however, a more moderate version of Newcomb's problem will serve as well Even if the difference of $C(E/G)$ and $C(E/-G)$ is quite small, provided that it exceeds g/e, we have a counterexample More moderate versions can also be more down-to-earth, as witness the medical Newcomb problems [6] Suppose you like eating eggs, or smoking, or loafing when

[5] Presented in Robert Nozick, Newcomb s Problem and Two Principles of Choice , in N Rescher *et al* eds , *Essays in Honor of Carl G Hempel* (Dordrecht, Holland D Reidel, 1970)

[6] Discussed in Skyrms, and Nozick, *opera cit* in Richard C Jeffrey, Choice, Chance, and Credence , in G H von Wright and G Fløistad, eds , *Philosophy of Logic* (Dordrecht, Holland M Nijhoff 1980), and in Richard C Jeffrey, How is it Reasonable to Base Preferences on Estimates of Chance? in D H Mellor, ed , *Science Belief and Behaviour Essays in Honour of R B Braithwaite* (Cambridge Cambridge University Press, 1980) I discuss another sort of moderate and down to earth Newcomb problem in Prisoners Dilemma is a Newcomb Problem *Philosophy and Public Affairs*, 8 (1979) pp 235–240

you might go out and run You are convinced, contrary to popular belief, that these pleasures will do you no harm at all (Whether you are right about this is irrelevant) But also you think you might have some dread medical condition a lesion of an artery, or nascent cancer, or a weak heart If you have it, there's nothing you can do about it now and it will probably do you a lot of harm eventually In its earlier stages, this condition is hard to detect But you are convinced that it has some tendency, perhaps slight, to cause you to eat eggs, smoke, or loaf So if you find yourself indulging, that is at least some evidence that you have the condition and are in for big trouble But is that any reason not to indulge in harmless pleasures? The V-maximising rule says yes, if the numbers are right I say no

So far, I have considered pure Newcomb problems There are also mixed problems You may think that taking the good has some tendency to produce (or prevent) the evil, but also is a manifestation of some prior state which tends to produce the evil Or you may be uncertain whether your situation is a Newcomb problem or not, dividing your credence between alternative hypotheses about the causal relations that prevail These mixed cases are still more realistic, yet even they can refute noncausal decision theory

However, no Newcomb problem, pure or mixed, can refute anything if it is not possible The Tickle Defence of noncausal decision theory[7] questions whether Newcomb problems really can arise It runs as follows "Supposedly the prior state that tends to cause the evil also tends to cause you to take the good The dangerous lesion causes you to choose to eat eggs, or whatever How can it do that? If you are fully rational your choices are governed entirely by your beliefs and desires so nothing can influence your choices except by influencing your beliefs and desires But if you are fully rational, you know your own mind If the lesion produces beliefs and desires favourable to eating eggs, you will be aware of those beliefs and desires at the outset of deliberation So you won't have to wait until you find yourself eating eggs to get the bad news You will have it already when you feel that tickle in the tastebuds—or whatever introspectible state it might be—

[7] Discussed in Skyrms, *op cit* and most fully presented in Ellery Eells Causality, Utility and Decision , *Synthese*, 48 (1981) 295–329 Eells argues that Newcomb problems are stopped by assumptions of rationality and self knowledge somewhat weaker than those of the simple Tickle Defence considered here but even those weaker assumptions seem to me unduly restrictive

that manifests your desire for eggs Your consequent choice tells you nothing more By the time you decide whether to eat eggs, your credence function already has been modified by the evidence of the tickle Then $C(E/G)$ does not exceed $C(E/-G)$, their difference is zero and so does not exceed g/e, $-G$ is not V-maximal, and noncausal decision theory does not make the mistake of telling you not to eat the eggs "

I reply that the Tickle Defence does establish that a Newcomb problem cannot arise for a fully rational agent, but that decision theory should not be limited to apply only to the fully rational agent [8] Not so, at least, if rationality is taken to include self-knowledge May we not ask what choice would be rational for the partly rational agent, and whether or not his partly rational methods of decision will steer him correctly? A partly rational agent may very well be in a moderate Newcomb problem, either because his choices are influenced by something besides his beliefs and desires or because he cannot quite tell the strengths of his beliefs and desires before he acts ("How can I tell what I think till I see what I say?"—E M Forster) For the dithery and the self-deceptive, no amount of *Gedankenexperimente* in decision can provide as much self-knowledge as the real thing So even if the Tickle Defence shows that noncausal decision theory gives the right answer under powerful assumptions of rationality (whether or not for the right reasons), Newcomb problems still show that a general decision theory must be causal

5 UTILITY AND DEPENDENCY HYPOTHESES

Suppose someone knows all there is to know about how the things he cares about do and do not depend causally on his present actions If something is beyond his control, so that it will obtain—or have a cer-

[8] In fact, it may not apply to the fully rational agent It is hard to see how such an agent can be uncertain what he is going to choose, hence hard to see how he can be in a position to deliberate See Richard C Jeffrey, A Note on the Kinematics of Preference , *Erkenntnis*, 11 (1977) 135–141 Further the fully rational agent required by the Tickle Defence is, in one way, not so very rational after all Self knowledge is an aspect of rationality but so is willingness to learn from experience If the agent s introspective data make him absolutely certain of his own credences and values, as they must if the Defence is to work then no amount of evidence that those data are untrustworthy will ever persuade him not to trust them

tain chance of obtaining—no matter what he does, then he knows that for certain And if something is within his control, he knows that for certain, further, he knows the extent of his influence over it and he knows what he must do to influence it one way or another Then there can be no Newcomb problems for him Whatever news his actions may bring, they cannot change his mind about the likely outcomes of his alternative actions He knew it all before

Let us call the sort of proposition that this agent knows—a maximally specific proposition about how the things he cares about do and do not depend causally on his present actions—a *dependency hypothesis* (for that agent at that time) Since there must be some truth or other on the subject, and since the dependency hypotheses are maximally specific and cannot differ without conflicting, they comprise a partition Exactly one of them holds at any world, and it specifies the relevant relations of causal dependence that prevail there

It would make no difference if our know-it-all didn't really know If he concentrates all his credence on a single dependency hypothesis, whether rightly or wrongly, then there can be no Newcomb problems for him His actions cannot bring him news about which dependency hypothesis holds if he already is quite certain which one it is

Within a single dependency hypothesis, so to speak, V-maximising is right It is rational to seek good news by doing that which, according to the dependency hypothesis you believe, most tends to produce good results That is the same as seeking good results Failures of V-maximising appear only if, first, you are sensible enough to spread your credence over several dependency hypotheses, and second, your actions might be evidence for some dependency hypotheses and against others That is what may enable the agent to seek good news not in the proper way, by seeking good results, but rather by doing what would be evidence for a good dependency hypothesis That is the recipe for Newcomb problems

What should you do if you spread your credence over several dependency hypotheses? You should consider the expected value of your options under the several hypotheses, you should weight these by the credences you attach to the hypotheses, and you should maximise the weighted average Henceforth I reserve the variable K to range over dependency hypotheses (or over members of partitions that play a parallel role in other versions of causal decision theory) Let us define the (*expected*) *utility* of an option A by

$$U(A) =^{df} \Sigma_K C(K)V(AK)$$

My version of causal decision theory prescribes the rule of U-*maximising* according to which a rational choice is one that has the greatest expected utility. Option A is U-maximal iff $U(A)$ is not exceeded by any $U(A')$, and to act rationally is to realise some U-maximal option.

In putting this forward as the rule of rational decision, of course I speak for myself, but I hope I have found a neutral formulation which fits not only my version of causal decision theory but also the versions proposed by Gibbard and Harper, Skyrms, and Sobel. There are certainly differences about the nature of dependency hypotheses, but if I am right, these are small matters compared to our common advocacy of utility maximising as just defined.

In distinguishing as I have between V and U—value and utility—I have followed the notation of Gibbard and Harper. But also I think I have followed the lead of ordinary language, in which "utility" means much the same as "usefulness". Certainly the latter term is causal. Which would you call the useful action: the one that tends to produce good results? Or the one that does no good at all (or even a little harm) and yet is equally welcome because it is a sign of something else that does produce good results? (Assume again that the news is not valued for its own sake.) Surely the first—and that is the one with greater utility in my terminology, though both may have equal value.

It is essential to define utility as we did using the unconditional credences $C(K)$ of dependency hypotheses, not their conditional credences $C(K/A)$. If the two differ, any difference expresses exactly that news-bearing aspect of the options that we meant to suppress. Had we used the conditional credences, we would have arrived at nothing different from V. For the Rule of Averaging applies to any partition, and hence to the partition of dependency hypotheses, giving

$$(5)\quad V(A) = \Sigma_K C(K/A)V(AK)$$

Let us give noncausal decision theory its due before we take leave of it. It works whenever the dependency hypotheses are probabilistically independent of the options, so that all the $C(K/A)$'s equal the corresponding $C(K)$'s. Then by (5) and the definition of U, the corresponding $V(A)$'s and $U(A)$'s also are equal. V-maximising gives the same right answers as U-maximising. The Tickle Defence seems to show that the K's must be independent of the A's for any fully rational agent. Even for partly rational agents, it seems plausible that they are at least close to independent in most realistic cases. Then indeed V-maximising works. But it works because the agent's beliefs about causal depen-

dence are such as to make it work It does not work for reasons which leave causal relations out of the story

I am suggesting that we ought to undo a seeming advance in the development of decision theory Everyone agrees that it would be ridiculous to maximise the "expected utility" defined by

$$\Sigma_Z C(Z)V(AZ)$$

where Z ranges over just any old partition It would lead to different answers for different partitions For the partition of value-level propositions, for instance, it would tell us fatalistically that all options are equally good! What to do? Savage suggested, in effect, that we make the calculation with unconditional credences, but make sure to use only the right sort of partition[9] But what sort is that? Jeffrey responded that we would do better to make the calculation with conditional credences, as in the right hand side of (2) Then we need not be selective about partitions, since we get the same answer, namely $V(A)$, for all of them In a way, Jeffrey himself was making decision theory causal But he did it by using probabilistic dependence as a mark of causal dependence, and unfortunately the two need not always go together So I have thought it better to return to unconditional credences and say what sort of partition is right

As I have formulated it, causal decision theory is causal in two different ways The dependency hypotheses are causal in their content they class worlds together on the basis of likenesses of causal dependence But also the dependency hypotheses themselves are causally independent of the agent's actions They specify his influence over other things, but over them he has no influence (Suppose he did Consider the dependency hypothesis which we get by taking account of the ways the agent can manipulate dependency hypotheses to enhance his control over other things This hypothesis seems to be right no matter what he does Then he has no influence over whether this hypothesis or another is right, contrary to our supposition that the dependency hypotheses are within his influence) Dependency hypotheses are "act-independent states" in a causal sense, though not necessarily in the probabilistic sense If we say that the right sort of partition for calculating expected utility is a causally act-independent one, then the parti-

[9] Leonard J Savage, *The Foundations of Statistics* (New York Wiley, 1954) p 15 The suggestion is discussed by Richard C Jeffrey in Savage s Omelet , in F Suppe and P D Asquith, eds , *PSA 1976*, Volume 2 (East Lansing, Michigan Philosophy of Science Association, 1977)

tion of dependency hypotheses qualifies But I think it is better to say just that the right partition is the partition of dependency hypotheses, in which case the emphasis is on their causal content rather than their act-independence

If any of the credences $C(AK)$ is zero, the rule of U-maximising falls silent For in that case $V(AK)$ becomes an undefined sum of quotients with denominator zero, so $U(A)$ in turn is undefined and A cannot be compared in utility with the other options Should that silence worry us? I think not, for the case ought never to arise It may seem that it arises in the most extreme sort of Newcomb problem suppose that taking the good is thought to make it absolutely certain that the prior state obtains and the evil will follow Then if A is the option of taking the good and K says that the agent stands a chance of escaping the evil, $C(AK)$ is indeed zero and $U(A)$ is indeed undefined What should you do in such an extreme Newcomb problem? V-maximise after all?

No, what you should do is not be in that problem in the first place Nothing should ever be held as certain as all that, with the possible exception of the testimony of the senses Absolute certainty is tantamount to firm resolve never to change your mind no matter what, and that is objectionable However much reason you may get to think that option A will not be realised if K holds, you will not if you are rational lower $C(AK)$ quite to zero Let it by all means get very, very small, but very, very small denominators do not make utilities go undefined

What of the partly rational agent, whom I have no wish to ignore? Might he not rashly lower some credence $C(AK)$ all the way to zero? I am inclined to think not What makes it so that someone has a certain credence is that its ascription to him is part of a systematic pattern of ascriptions, both to him and to others like him, both as they are and as they would have been had events gone a bit differently, that does the best job overall of rationalising behaviour [10] I find it hard to see how the ascription of rash zeros could be part of such a best pattern It seems that a pattern that ascribes very small positive values instead always could do just a bit better, rationalising the same behaviour without gratuitously ascribing the objectionable zeros If I am right

[10] See my Radical Interpretation *Synthese*, 23 (1974) pp 331–344 I now think that discussion is too individualistic, however, in that it neglects the possibility that one might have a belief or desire entirely because the ascription of it to him is part of a systematic pattern that best rationalises the behaviour of *other* people On this point, see my discussion of the madman in Mad Pain and Martian Pain , in Ned Block, ed , *Readings in Philosophy of Psychology* Volume 1 (Cambridge, Massachusetts Harvard University Press, 1980)

about this, rash zeros are one sort of irrationality that is downright impossible [11]

6 REFORMULATIONS

The causal decision theory proposed above can be reformulated in various equivalent ways These will give us some further understanding of the theory, and will help us in comparing it with other proposed versions of causal decision theory

Expansions We can apply the Rule of Averaging to expand the V(*AK*)'s that appear in our definition of expected utility Let *Z* range over any partition Then we have

(6) $U(A) = \Sigma_K \Sigma_Z C(K)C(Z/AK)V(AKZ)$

(If any C(*AKZ*) is zero we may take the term for *K* and *Z* as zero, despite the fact the V(*AKZ*) is undefined) This seems only to make a simple thing complicated, but if the partition is well chosen, (6) may serve to express the utility of an option in terms of quantities that we find it comparatively easy to judge

Let us call a partition *rich* iff, for every member *S* of that partition and for every option *A* and dependency hypothesis *K*, V(*AKS*) equals V(*AS*) That means that the *AS*'s describe outcomes of options so fully that the addition of a dependency hypothesis tells us no more about the features of the outcome that matter to the agent Henceforth I reserve the variable *S* to range over rich partitions Given richness of the partition, we can factor the value terms in (6) part way out, to obtain

(7) $U(A) = \Sigma_S(\Sigma_K C(K)C(S/AK))V(AS)$

Equation (7) for expected utility resembles equation (2) for expected value, except that the inner sum in (7) replaces the conditional credence C(*S*/*A*) in the corresponding instance of (2) As we shall see, the analogy can be pushed further Two examples of rich partitions to

[11] Those who think that credences can easily fall to zero often seem to have in mind cre dences conditional on some background theory of the world which is accepted, albeit tentatively, in an all–or–nothing fashion While I don t object to this notion, it is not what I mean by credence As I understand the term, what is open to reconsideration does not have a credence of zero or one these extremes are not to be embraced lightly

which (7) applies are the partition of possible worlds and the partition of value-level propositions [V=v]

Imaging Suppose we have a function that selects, for any pair of a world W and a suitable proposition X, a probability distribution W_X Suppose further that W_X assigns probability only to X-worlds, so that $W_X(X)$ equals one (Hence at least the empty proposition must not be "suitable") Call the function an *imaging function*, and call W_X the *image of W on X* The image might be sharp, if W_X puts all its probability on a single world, or it might be blurred, with the probability spread over more than one world

Given an imaging function, we can apply it to form images also of probability distributions We sum the superimposed images of all the worlds, weighting the images by the original probabilities of their source worlds For any pair of a probability distribution C and a suitable proposition X, we define C_X, the *image of C on X*, as follows First, for any world W',

$$C_X(W') =^{df} \Sigma_W C(W)W_X(W'),$$

think of $C(W)W_X(W')$ as the amount of probability that is moved from W to W' in making the image We sum as usual for any proposition Y,

$$C_X(Y) =^{df} \Sigma_{W \varepsilon Y} C_X(W)$$

It is easy to check that C_X also is a probability distribution, and that it assigns probability only to X-worlds, so that $C_X(X)$ equals one Imaging is one way—conditionalising is another—to revise a given probability distribution so that all the probability is concentrated on a given proposition [12]

For our present purposes, what we want are images of the agent's

[12] Sharp imaging by means of a Stalnaker selection function is discussed in my Probabilities of Conditionals and Conditional Probabilities , *The Philosophical Review*, 85 (1976) pp 297–315 especially pp 309–311 [Pages 146–148 in this volume] This generalisation to cover blurred imaging as well is due to Peter Gardenfors, Imaging and Conditionalization *Journal of Philosophy*, 79 (1982) 747–760 a similar treatment appears in Donald Nute, *Topics in Conditional Logic* (Dordrecht, Holland D Reidel, 1980) Chapter 6 What is technically the same idea, otherwise motivated and under other names, appears in my Counterfactuals and Comparative Possibility *Journal of Philosophical Logic* 2 (1973) pp 418–446, Section 8, in John L Pollock, *Subjunctive Reasoning* (Dordrecht, Holland D Reidel, 1976) pp 219–236, and in Sobel, *op cit* The possibility of deriving an imaging function from a partition was suggested by Brian Skyrms in discussion of a paper by Robert Stalnaker at the 1979 annual meeting of the American Philosophical Association Eastern Division

credence function on his various options The needed imaging function can be defined in terms of the partition of dependency hypotheses let

$$W_A(W') =^{df} C(W'/AK_W)$$

for any option A and worlds W and W', where K_W is the dependency hypothesis that holds at W In words move the credence of world W over to the A-worlds in the same dependency hypothesis, and distribute it among those worlds in proportion to their original credence (Here again we would be in trouble if any of the $C(AK)$'s were zero, but I think we needn't worry) It follows from the several definitions just given that for any option A and proposition Y,

(8) $C_A(Y) = \Sigma_K C(K)C(Y/AK)$

The inner sum in (7) therefore turns out to be the credence, imaged on A, of S So by (7) and (8) together,

(9) $U(A) = \Sigma_S C_A(S)V(AS)$

Now we have something like the Rule of Averaging for expected value, except that the partition must be rich and we must image rather than conditionalising For the rich partition of possible worlds we have

(10) $U(A) = \Sigma_W C_A(W)V(W)$

which resembles the definition of expected value For the rich partition of value-level propositions we have something resembling (3)

(11) $U(A) = \Sigma_v C_A([V = v])v$

7 PRIMITIVE IMAGING SOBEL

To reformulate causal decision theory in terms of imaging, I proceeded in two steps I began with the dependency hypotheses and used them to define an imaging function, then I redefined the expected utility of an option in terms of imaging We could omit the first step and leave the dependency hypotheses out of it We could take the imaging function as primitive, and go on as I did to define expected utility by means of it That is the decision theory of J Howard Sobel, *op cit*

Sobel starts with the images of worlds, which he calls *world-tendencies* (He considers images on all propositions possible relative to the given world, but for purposes of decision theory we can confine our attention to images on the agent's options) Just as we defined C_A

in terms of the W_A's, so Sobel goes on to define images of the agent's credence function He uses these in turn to define expected utility in the manner of (10), and he advocates maximising the utility so defined rather than expected value

Sobel unites his decision theory with a treatment of counterfactual conditionals in terms of closest antecedent-worlds [13] If $W_A(W')$ is positive, then we think of W' as one of the A-worlds that is in some sense closest to the world W What might be the case if it were the case that A, from the standpoint of W, is what holds at some such closest A-world, what would be the case if A, from the standpoint of W, is what holds at all of them Sobel's apparatus gives us quantitative counterfactuals intermediate between the mights and the woulds We can say that if it were that A, it would be with probability p that X, meaning that $W_A(X)$ equals p, or in Sobel's terminology that X holds on a subset of the closest A-worlds whose tendencies, at W and on the supposition A, sum to p

Though Sobel leaves the dependency hypotheses out of his decision theory, we can perhaps bring them back in Let us say that worlds *image alike* (on the agent's options) iff, for each option, their images on that option are exactly the same Imaging alike is an equivalence relation, so we have the partition of its equivalence classes If we start with the dependency hypotheses and define the imaging function as I did, it is immediate that worlds image alike iff they are worlds where the same dependency hypothesis holds, so the equivalence classes turn out to be just the dependency hypotheses

The question is whether dependency hypotheses could be brought into Sobel's theory by defining them as equivalence classes under the relation of imaging alike Each equivalence class could be described, in Sobel's terminology, as a maximally specific proposition about the tendencies of the world on all alternative suppositions about which option the agent realises That sounds like a dependency hypothesis to me Sobel tells me (personal communication, 1980) that he is inclined to agree, and does regard his decision theory as causal, though it is hard to tell that from his written presentation, in which causal language very seldom appears

If the proposal is to succeed technically, we need the following thesis if K_W is the equivalence class of W under the relation of imaging

[13] As in my *Counterfactuals* (Oxford Blackwell, 1973) without the complications raised by possible infinite sequences of closer and closer antecedent worlds

alike (of having the same tendencies on each option) then, for any option A and world W', $W_A(W')$ equals $C(W'/AK_W)$ If so, it follows that if we start as Sobel does with the imaging function, defining the dependency hypotheses as equivalence classes, and thence defining an imaging function as I did, we will get back the same imaging function that we started with It further follows, by our results in Section 6, that expected utility calculated in my way from the defined dependency hypotheses is the same as expected utility calculated in Sobel's way from the imaging function They must be the same, if the defined dependency hypotheses introduced into Sobel's theory are to play their proper role

Unfortunately, the required thesis is not a part of Sobel's theory, it would be an extra constraint on the imaging function It does seem a very plausible constraint, at least in ordinary cases Sobel suspends judgement about imposing a weaker version of the thesis (Connection Thesis 1, discussed in his Section 6 7) But his reservations, which would carry over to our version, entirely concern the extraordinary case of an agent who thinks he may somehow have foreknowledge of the outcomes of chance processes Sobel gives no reason, and I know of none, to doubt either version of the thesis except in extraordinary cases of that sort Then if we assume the thesis, it seems that we are only setting aside some very special cases—cases about which I, at least, have no firm views (I think them much more problematic for decision theory than the Newcomb problems) So far as the remaining cases are concerned, it is satisfactory to introduce defined dependency hypotheses into Sobel's theory and thereby render it equivalent to mine

8 FACTORS OUTSIDE OUR INFLUENCE SKYRMS

Moving on to the version of causal decision theory proposed by Brian Skyrms, *op cit* , we find a theory that is formally just like mine Skyrms' definition of *K-expectation*—his name for the sort of expected utility that should be maximised—is our equation (6) From that, with a trivial partition of Z's, we can immediately recover my first definition of expected utility Skyrms introduces a partition of hypotheses—the K's which give K-expectation its name—that play just the same role in his calculation of expected utility that the dependency hypotheses play in mine (Thus I have followed Skyrms in notation) So the only difference, if it is a difference, is in how the K's are characterised

Skyrms describes them at the outset as maximally specific specifications of the factors outside the agent's influence (at the time of decision) which are causally relevant to the outcome of the agent's action He gives another characterisation later, but let us take the first one first

I ask what Skyrms means to count as a "factor" Under a sufficiently broad construal, I have no objection to Skyrms' theory and I think it no different from mine On a narrower and more literal construal, I do not think Skyrms' theory is adequate as a general theory of rational decision, though I think that in practice it will often serve Insofar as Skyrms is serving up a general theory rather than practical rules of thumb, I think it is indeed the broad construal that he intends

(I also ask what Skyrms means by "relevant to the outcome" I can't see how any factor, broadly or narrowly construed, could fail to be relevant to some aspect of the outcome If the outcome is that I win a million dollars tomorrow, one aspect of this outcome may be that it takes place just one thousand years after some peasant felled an oak with ninety strokes of his axe So I suppose Skyrms' intent was to include only factors relevant to those features of the outcome that the agent cares about, as opposed to those that are matters of indifference to him That would parallel a like exclusion of matters of indifference in my definition of dependency hypotheses In neither case is the exclusion important Richer hypotheses, cluttered with matters of indifference, ought to give the same answers)

On the broad construal, a "factor" need not be the sort of localised particular occurrence that we commonly think of as causing or being caused It might be any matter of contingent fact whatever It might indeed be some particular occurrence It might be a vast dispersed pattern of occurrences throughout the universe It might be a law of nature It might be a dependency hypothesis On the broad construal, Skyrms is saying only that the *K*'s are maximally specific propositions about matters outside the agent's influence and relevant to features of the outcome that the agent cares about

A dependency hypothesis is outside the agent's influence It is relevant to features of the outcome that he cares about (*Causally* relevant?—Not clear, but if we're construing "factor" broadly, we can let that by as well) Any specification of something outside the agent's influence is included in a dependency hypothesis—recall that they cover what doesn't depend on the agent's actions as well as what does—unless it concerns something the agent doesn't care about I conclude that on the broad construal, Skyrms' *K*'s are nothing else

than the dependency hypotheses In that case his theory is the same as mine

On the narrow construal, a "factor" must be the sort of localised occurrence—event, state, omission, etc —that we normally think of as a cause In the medical Newcomb problems, for instance, the lesion or the nascent cancer or the weak heart is a causal factor narrowly and literally In motivating his theory, it is factors like these that Skyrms considers

Our topic is rational decision according to the agent's beliefs, be they right or wrong So it seems that we should take not the factors which really are outside his influence, but rather those he thinks are outside his influence But what if he divides his credence between several hypotheses as to which factors are outside his influence, as well he might? Skyrms responds to this challenge by redescribing his partition of hypotheses On his new description, each hypothesis consists of two parts (i) a preliminary hypothesis specifying which of the relevant causal factors are outside the agent's influence, and (ii) a full specification of those factors that are outside his influence according to part (i)

That is a welcome amendment, but I think it does not go far enough Influence is a matter of degree, so shouldn't the hypotheses say not just that the agent has some influence over a factor or none, but also how much? And if the hypothesis says that the agent has influence over a factor, shouldn't it also say which way the influence goes? Given that I can influence the temperature, do I make it cooler by turning the knob clockwise or counterclockwise? Make Skyrms' amendment and the other needed amendments, and you will have the dependency hypotheses back again

To illustrate my point, consider an agent with eccentric beliefs He thinks the influence of his actions ramifies but also fades, so that everything in the far future is within his influence but only a little bit Perhaps he thinks that his actions raise and lower the chances of future occurrences, but only very slightly Also he thinks that time is circular, so that the far future includes the present and the immediate past and indeed all of history Then he gives all his credence to a single one of Skyrms' two-part hypotheses the one saying that no occurrence whatever—no factor, on the narrow construal—is entirely outside his influence That means that on Skyrms' calculation his U(A)'s reduce to the corresponding V(A)'s, so V-maximising is right for him That's wrong Since he thinks he has very little influence over whether he has the dread lesion, his decision problem about eating eggs is very little

different from that of someone who thinks the lesion is entirely outside his influence V-maximising should come out wrong for very much the same reason in both cases

No such difficulty threatens Skyrms' proposal broadly construed The agent may well wonder which of the causal factors narrowly construed are within his influence, but he cannot rationally doubt that the dependency hypotheses are entirely outside it On the broad construal, Skyrms' second description of the partition of hypotheses is a gloss on the first, not an amendment The hypotheses already specify which of the (narrow) factors are outside the agent's influence, for that is itself a (broad) factor outside his influence Skyrms notes this, and that is why I think it must be the broad construal that he intends Likewise the degrees and directions of influence over (narrow) factors are themselves (broad) factors outside the agent's influence, hence already specified according to the broad construal of Skyrms' first description

Often, to be sure, the difference between the broad and narrow construals will not matter There may well be a correlation, holding throughout the worlds which enjoy significant credence, between dependency hypotheses and combinations of (narrow) factors outside the agent's influence The difference between good and bad dependency hypotheses may in practice amount to the difference between absence and presence of a lesion However, I find it rash to assume that there must always be some handy correlation to erase the difference between the broad and narrow construals Dependency hypotheses do indeed hold in virtue of lesions and the like, but they hold also in virtue of the laws of nature It would seem that uncertainty about dependency hypotheses might come at least partly from uncertainty about the laws

Skyrms is sympathetic, as am I,[14] to the neo-Humean thesis that every contingent truth about a world—law, dependency hypothesis, or what you will—holds somehow in virtue of that world's total history of manifest matters of particular fact Same history, same everything But that falls short of implying that dependency hypotheses hold just in virtue of casual factors, narrowly construed, they might hold partly in virtue of dispersed patterns of particular fact throughout history, including the future and the distant present Further, even if we are

[14] Although sympathetic, I have some doubts, see my A Subjectivist s Guide to Objective Chance in R C Jeffrey, ed *Studies in Inductive Logic and Probability*, Volume 2 (Berkeley and Los Angeles University of California Press, 1980) pp 290–292 [Pages 111–113 in this volume]

inclined to accept the neo-Humean thesis, it still seems safer not to make it a presupposition of our decision theory Whatever we think of the neo-Humean thesis, I conclude that Skyrms' decision theory is best taken under the broad construal of "factor" under which his K's are the dependency hypotheses and his calculation of utility is the same as mine [15]

9 COUNTERFACTUAL DEPENDENCE GIBBARD AND HARPER

If we want to express a dependency hypothesis in ordinary language, it is hard to avoid the use of counterfactual conditionals saying what would happen if the agent were to realise his various alternative options Suppose that on a certain occasion I'm interested in getting Bruce to purr I could try brushing, stroking, or leaving alone, pretend that these are my narrowest options Bruce might purr loudly, softly, or not at all, pretend that these alternatives are a rich partition (Those simplifying pretences are of course very far from the truth) Much of my credence goes to the dependency hypothesis given by these three counterfactuals

I brush Bruce $\Box\!\!\rightarrow$ he purrs loudly,
I stroke Bruce $\Box\!\!\rightarrow$ he purrs softly,
I leave Bruce alone $\Box\!\!\rightarrow$ he doesn't purr

($\Box\!\!\rightarrow$ is used here as a sentential connective, read "if it were that it would be that " I use it also as an operator which applies to two propositions to make a proposition, context will distinguish the uses)

[15] The decision theory of Nancy Cartwright, Causal Laws and Effective Strategies , *Nous*, 13 (1979) pp 419–437, is, as she remarks, structurally identical to Skyrms theory for the case where value is a matter of reaching some all or-nothing goal However, hers is not a theory of subjectively rational decision in the single case, like Skyrms theory and the others considered in this paper but instead is a theory of objectively effective generic strategies Since the subject matters are different, the structural identity is misleading Cartwright s theory might somehow imply a single case theory having more than structure in common with Skyrms theory, but that would take principles she does not provide, *inter alia* principles relating generic causal conduciveness to influence in the single case So it is not clear that Cartwright s decision theory, causal though it is, falls under my claim that we causal decision theorists share one common idea

This hypothesis says that loud and soft purring are within my influence—they depend on what I do It specifies the extent of my influence, namely full control And it specifies the direction of influence, what I must do to get what This is one dependency hypothesis I give some of my credence to others, for instance this (rather less satisfactory) one

I brush Bruce $\Box\!\!\rightarrow$ he doesn't purr,
I stroke Bruce $\Box\!\!\rightarrow$ he doesn't purr,
I leave Bruce alone $\Box\!\!\rightarrow$ he doesn't purr

That dependency hypothesis says that the lack of purring is outside my influence, it is causally independent of what I do Altogether there are twenty-seven dependency hypotheses expressible in this way, though some of them get very little credence

Note that it is the pattern of counterfactuals, not any single one of them, that expresses causal dependence or independence As we have seen, the same counterfactual

I leave Bruce alone $\Box\!\!\rightarrow$ he doesn't purr

figures in the first hypothesis as part of a pattern of dependence and in the second as part of a pattern of independence

It is clear that not just any counterfactual could be part of a pattern expressing causal dependence or independence The antecedent and consequent must specify occurrences capable of causing and being caused, and the occurrences must be entirely distinct Further, we must exclude "back-tracking counterfactuals" based on reasoning from different supposed effects back to different causes and forward again to differences in other effects Suppose I am convinced that stroking has no influence over purring, but that I wouldn't stroke Bruce unless I were in a mood that gets him to purr softly by emotional telepathy Then I give credence to

I stroke Bruce $\Box\!\!\rightarrow$ he purrs softly

taken in a back-tracking sense, but not taken in the sense that it must have if it is to be part of a pattern of causal dependence or independence

Let us define *causal counterfactuals* as those that can belong to patterns of causal dependence or independence Some will doubt that causal counterfactuals can be distinguished from others except in causal terms, I disagree, and think it possible to delimit the causal counterfactuals in other terms and thus provide noncircular counter-

factual analyses of causal dependence and causation itself But that is a question for other papers [16] For present purposes, it is enough that dependency hypotheses can be expressed (sometimes, at least) by patterns of causal counterfactuals I hope that much is adequately confirmed by examples like the one just considered And that much can be true regardless of whether the pattern of counterfactuals provides a noncircular analysis

Turning from language to propositions, what we want are causal counterfactuals $A \mathbin{\Box\!\!\rightarrow} S$, where A is one of the agent's options and S belongs to some rich partition The rich partition must be one whose members specify combinations of occurrences wholly distinct from the actions specified by the agent's options It seems a safe assumption that some such rich partition exists Suppose some definite one to be chosen (it should make no difference which one) Define a *full pattern* as a set consisting of exactly one such counterfactual proposition for each option I claim that the conjunction of the counterfactuals in any full pattern is a dependency hypothesis

Conjunctions of different full patterns are contraries, as any two dependency hypotheses should be For if S and S' are contraries, and A is possible (which any option is), then also $A \mathbin{\Box\!\!\rightarrow} S$ and $A \mathbin{\Box\!\!\rightarrow} S'$ are contraries,[17] and any two full patterns must differ by at least one such contrary pair

What is not so clear is that some full pattern or other holds at any world, leaving no room for any other dependency hypotheses besides the conjunctions of full patterns We shall consider this question soon But for now, let us answer it by fiat Assume that there is a full pattern for every world, so that the dependency hypotheses are all and only the conjunctions of full patterns

That assumption yields the causal decision theory proposed by Allan Gibbard and William Harper, *op cit* , following a suggestion of Robert Stalnaker My statement of it amounts to their Savage-style formulation with conjunctions of full patterns of counterfactuals as act-independent states, and their discussion of consequences in their Section 6 shows that they join me in regarding these conjunctions as expressing causal dependence or independence Although they do not

[16] In particular, my Causation , *Journal of Philosophy*, 70 (1973) pp 556–567, and Counterfactual Dependence and Time s Arrow , *Nous* 13 (1979) pp 455–476

[17] Here and henceforth I make free use of some fairly uncontroversial logical principles for counterfactuals namely, those given by the system CK+ID+MP of Brian F Chellas, Basic Conditional Logic *Journal of Philosophical Logic*, 4 (1975) pp 133–153

explicitly distinguish causal counterfactuals from others, their Section 2 sketches a theory of counterfactuals which plainly is built to exclude back-trackers in any ordinary situation This is essential to their purpose A theory which used counterfactuals in formally the same way, but which freely admitted back-trackers, would not be a causal decision theory Its conjunctions of full patterns including back-trackers would not be causal dependency hypotheses, and it would give just those wrong answers about Newcomb problems that we causal decision theorists are trying to avoid [18]

Consider some particular A and S If a dependency hypothesis K is the conjunction of a full pattern that includes $A \Box\!\!\rightarrow S$, then AK implies S and $C(S/AK)$ equals one If K is the conjunction of a full pattern that includes not $A \Box\!\!\rightarrow S$ but some contrary $A \Box\!\!\rightarrow S'$, then AK contradicts S and $C(S/AK)$ equals zero *Ex hypothesi*, every dependency hypothesis K is of one kind or the other Then the K's for which $C(S/AK)$ equals one comprise a partition of $A \Box\!\!\rightarrow S$, while $C(S/AK)$ equals zero for all other K's It follows by the Rule of Additivity for credence that

$$(12) \quad C(A \Box\!\!\rightarrow S) = \Sigma_K C(K)C(S/AK)$$

(Comparing (12) with (8), we find that our present assumptions equate $C(A \Box\!\!\rightarrow S)$ with $C_A(S)$, the credence of S imaged on the option A) Substituting (12) into (7) we have

$$(13) \quad U(A) = \Sigma_S C(A \Box\!\!\rightarrow S)V(AS),$$

which amounts to Gibbard and Harper's defining formula for the "genuine expected utility" they deem it rational to maximise [19]

We have come the long way around to (13), which is not only simple but also intuitive in its own right But (13) by itself does not display the causal character of Gibbard and Harper's theory, and that is what makes it worthwhile to come at it by way of dependency hypotheses No single $C(A \Box\!\!\rightarrow S)$ reveals the agent's causal views, since it sums the credences of hypotheses which set $A \Box\!\!\rightarrow S$ in a pattern of dependence and others which set $A \Box\!\!\rightarrow S$ in a pattern of independence Conse-

[18] Such a theory is defended in Terence Horgan Counterfactuals and Newcomb s Problem , *Journal of Philosophy*, 78 (1981) 331–356

[19] To get exactly their formula, take their outcomes as conjunctions AS with desirability given by $V(AS)$, and bear in mind (i) that $A \Box\!\!\rightarrow AS$ is the same as $A \Box\!\!\rightarrow S$, and (ii) that if A and A' are contraries, $A \Box\!\!\rightarrow A'S$ is the empty proposition with credence zero

quently the roundabout approach helps us to appreciate what the theory of Gibbard and Harper has in common with that of someone like Skyrms who is reluctant to use counterfactuals in expressing dependency hypotheses

10 COUNTERFACTUAL DEPENDENCE WITH CHANCY OUTCOMES

The assumption that there is a full pattern for each world is a consequence of Stalnaker's principle of Conditional Excluded Middle,[20] which says that either $X \mathbin{\square\!\!\rightarrow} Y$ or $X \mathbin{\square\!\!\rightarrow} -Y$ holds at any world (where $-Y$ is the negation of Y) It follows that if $Y, Y', \ldots$ are a partition and X is possible, then $X \mathbin{\square\!\!\rightarrow} Y, X \mathbin{\square\!\!\rightarrow} Y', \ldots$ also are a partition The conjunctions of full patterns are then a partition because, for any option A, the counterfactuals $A \mathbin{\square\!\!\rightarrow} S, A \mathbin{\square\!\!\rightarrow} S', \ldots$ are a partition

Conditional Excluded Middle is open to objection on two counts, one more serious than the other Hence so is the decision theory of Gibbard and Harper, insofar as it relies on Conditional Excluded Middle to support the assumption that there is a full pattern for each world Gibbard and Harper themselves are not to be faulted, for they tell us that their "reason for casting the rough theory in a form which gives these principles is that circumstances where these can fail involve complications which it would be best to ignore in preliminary work" (*Op cit* 128) Fair enough, still, we have unfinished business on the agenda

The first objection to Conditional Excluded Middle is that it makes arbitrary choices It says that the way things would be on a false but possible supposition X is no less specific than the way things actually are Some single, fully specific possible world is the one that would be actualised if it were that X Since the worlds $W, W', \ldots$ are a partition, so are the counterfactuals $X \mathbin{\square\!\!\rightarrow} W, X \mathbin{\square\!\!\rightarrow} W', \ldots$ saying exactly how things would be if X But surely some questions about how things would be if X have no nonarbitary answers if you had a sister, would she like blintzes?

The less specific the supposition, the less it settles, the more farfetched it is, the less can be settled by what carries over from actuality,

[20] Robert C Stalnaker, A Theory of Conditionals , in N Rescher, ed , *Studies in Logical Theory* (Oxford Blackwell, 1968), gives a semantical analysis in which Conditional Excluded Middle follows from ordinary Excluded Middle applied to the selected antecedent world

and the less is settled otherwise, the more must be settled arbitrarily or not at all But the supposition that an agent realises one of his narrowest options is neither unspecific nor far-fetched So the Arbitrariness Objection may be formidable against the general principle of Conditional Excluded Middle, yet not formidable against the special case of it that gives us a full pattern for each world

Further, Bas van Fraassen has taught us a general method for tolerating arbitrariness [21] When forced to concede that certain choices would be arbitrary, we leave those choices unmade and we ask what happens on all the alternative ways of making them What is constant over all the ways of making them is determinate, what varies is indeterminate If the provision of full patterns for certain worlds is partly arbitrary, so be it Then indeed some arbitrary variation may infect the C(*K*)'s, C(*S*/*AK*)'s, C(*A* □→*S*)'s, and even the U(*A*)'s It might even infect the set of U-maximal options Then indeed it would be (wholly or partly) indeterminate which options the Gibbard-Harper theory commends as rational All of that might happen, but it needn't The arbitrary variation might vanish part way through the calculation, leaving the rest determinate The less arbitrary variation there is at the start, of course, the less risk that there will be any at the end

I conclude that the Arbitrariness Objection by itself is no great threat to Gibbard and Harper's version of causal decision theory We can well afford to admit that the theory might fail occasionally to give a determinate answer Indeed, I admit that already, for any version, on other grounds I think there is sometimes an arbitrary element in the assignment of C and V functions to partly rational agents No worries, so long as we can reasonably hope that the answers are mostly determinate

Unfortunately there is a second, and worse, objection against Conditional Excluded Middle and the Gibbard-Harper theory In part it is an independent objection, in part an argument that van Fraassen's method of tolerating arbitrariness would be severely overloaded if we insisted on providing full patterns all around (and *a fortiori* if we insisted on saving Conditional Excluded Middle generally), and we could not reasonably hope that the answers are mostly determinate Suppose the agent thinks—as he should if he is well-educated—that the

[21] See Bas van Fraassen, Singular Terms, Truth-Value Gaps and Free Logic *Journal of Philosophy* 63 (1966) pp 481–495 Use of van Fraassen s method to concede and tolerate arbitrariness in counterfactuals was suggested to me by Stalnaker in 1968 (personal communication) and is discussed in my *Counterfactuals* pp 81–83

actual world may very well be an indeterministic one, where many things he cares about are settled by chance processes Then he may give little of his credence to worlds where full patterns hold In fact he may well give little credence to any of the $A \mathbin{\Box\!\!\rightarrow} S$ counterfactuals that make up these patterns

Consider again my problem of getting Bruce to purr I think that Bruce works by firing of neurons, I think neurons work by chemical reactions, and I think the making or breaking of a chemical bond is a chance event in the same way that the radioactive decay of a nucleus is Maybe I still give some small credence to the twenty-seven full patterns considered in Section 9—after all, I might be wrong to think that Bruce is chancy But mostly I give my credence to the denials of all the counterfactuals that appear in those patterns, and to such counterfactuals as

> I brush Bruce $\mathbin{\Box\!\!\rightarrow}$ a chance process goes on in him which has certain probabilities of eventuating in his purring loudly, softly, or not at all,

and likewise for the options of stroking and leaving alone A diehard supporter of the Gibbard-Harper theory (not Gibbard or Harper, I should think) might claim that I give my credence mostly to worlds where it is arbitrary which one of the twenty-seven full patterns holds, but determinate that some one of them holds If he is right, even this easy little decision problem comes out totally indeterminate, for the arbitrary variation he posits is surely enough to swing the answer any way at all Nor would it help if I believe that whichever I did, all the probabilities of Bruce's purring loudly, softly, or not at all would be close to zero or one Nor would a more realistic decision problem fare any better unless the agent is a fairly convinced determinist, the answers we want vanish into indeterminacy The diehard destroys the theory in order to save it

Anyway, the diehard is just wrong If the world is the chancy way I mostly think it is, there's nothing at all arbitrary or indeterminate about the counterfactuals in the full patterns They are flatly, determinately false So is their disjunction, the diehard agrees that it is determinate in truth value, but the trouble is that he thinks it is determinately true

Unlike the Arbitrariness Objection, the Chance Objection seems to me decisive both against Conditional Excluded Middle generally and against the assumption that there is a full pattern for each world Our conception of dependency hypotheses as conjunctions of full patterns is too narrow Fortunately, the needed correction is not far to seek

I shall have to assume that anyone who gives credence to indeterministic worlds without full patterns is someone who—implicitly and in practice, if not according to his official philosophy—distributes his credence over contingent propositions about single-case, objective chances Chance is a kind of probability that is neither frequency nor credence, though related to both I have no analysis to offer, but I am convinced that we do have this concept and we don't have any substitute for it [22]

Suppose some rich partition to be chosen which meets the requirement of distinct occurrences laid down in Section 9 Let the variable p range over candidate probability distributions for this rich partition functions assigning to each S in the partition a number $p(S)$ in the interval from zero to one, such that the $p(S)$'s sum to one Let [P=p] be the proposition that holds at just those worlds where the chances of the S's, as of the time when the agent realises his chosen option, are correctly given by the function p Call [P=p] a *chance proposition*, and note that the chance propositions are a partition Now consider the causal counterfactuals $A \mathbin{\square\!\!\rightarrow} [P=p]$ from the agent's options to the chance propositions Define a *probabilistic full pattern* as a set containing exactly one such counterfactual for each option I claim that the conjunction of the counterfactuals in any probabilistic full pattern is a causal dependency hypothesis It specifies plain causal dependence or independence of the chances of the S's on the A's, and thereby it specifies a probabilistic kind of causal dependence of the S's themselves on the A's

Here, for example, are verbal expressions of three chance propositions

$[P=p_1]$ The chance that Bruce purrs loudly is 50%, the chance that he purrs softly is 40%, and the chance that he purrs not at all is 10%

$[P=p_2]$ (similar, but with 30%, 50%, 20%)

$[P=p_3]$ (similar, but with 10%, 10%, 80%)

(The chance is to be at the time of my realising an option, the purring or not is to be at a certain time shortly after) And here is a dependency hypothesis that might get as much of my credence as any

[22] For a fuller discussion of chance and its relations to frequency and credence, see A Subjectivist s Guide to Objective Chance

I brush Bruce $\Box\!\!\rightarrow [P=p_1]$ holds,
I stroke Bruce $\Box\!\!\rightarrow [P=p_2]$ holds,
I leave Bruce alone $\Box\!\!\rightarrow [P=p_3]$ holds

Observe that this hypothesis addresses itself not only to the question of whether loud and soft purring are within my influence, but also to the question of the extent and the direction of my influence

If a chance proposition says that one of the *S*'s has a chance of one, it must say that the others all have chances of zero Call such a chance proposition *extreme* I shall not distinguish between an extreme proposition and the *S* that it favours If they differ, it is only on worlds where something with zero chance nevertheless happens I am inclined to think that they do not differ at all, since there are no worlds where anything with zero chance happens, the contrary opinion comes of mistaking infinitesimals for zero But even if there is a difference between extreme chance propositions and their favoured *S*'s, it will not matter to calculations of utility so let us neglect it Then our previous dependency hypotheses, the conjunctions of full patterns, are subsumed under the conjunctions of probabilistic full patterns So are the conjunctions of mixed full patterns that consist partly of $A \Box\!\!\rightarrow S$'s and partly of $A \Box\!\!\rightarrow [P=p]$'s

Dare we assume that there is a probabilistic full pattern for every world, so that on this second try we have succeeded in capturing all the dependency hypotheses by means of counterfactuals? I shall assume it, not without misgivings That means accepting a special case of Conditional Excluded Middle, but (i) the Chance Objection will not arise again,[23] (ii) there should not be too much need for arbitrary choice on other grounds, since the options are quite specific suppositions and not far-fetched, and (iii) limited arbitrary choice results in nothing worse than a limited risk of the answers going indeterminate

So my own causal decision theory consists of two theses My main thesis is that we should maximise expected utility calculated by means of dependency hypotheses It is this main thesis that I claim is implicitly accepted also by Gibbard and Harper, Skyrms, and Sobel My subsidiary thesis, which I put forward much more tentatively and which I won't try to foist on my allies, is that the dependency hypotheses are exactly the conjunctions of probabilistic full patterns

(The change I have made in the Gibbard-Harper version has been

[23] Chances aren t chancy, if [P=p] pertains to a certain time, its own chance at that time of holding must be zero or one, by the argument of A Subjectivist s Guide to Objec tive Chance pp 276–277 [Pages 96–98 in this volume]

simply to replace the rich partition of S's by the partition of chance propositions [P=p] pertaining to these S's. One might think that perhaps that was no change at all: perhaps the S's already were the chance propositions for some other rich partition. However, I think it at least doubtful that the chance propositions can be said to "specify combinations of occurrences" as the S's were required to do. This question would lead us back to the neo-Humean thesis discussed in Section 8.)

Consider some particular A and S. If a dependency hypothesis K is the conjunction of a probabilistic full pattern, then for some p, K implies $A \boxrightarrow [P=p]$. Then AK implies [P=p], and $C(S/AK)$ equals $p(S)$, at least in any ordinary case.[24] For any p, the K's that are conjunctions of probabilistic full patterns including $A \boxrightarrow [P=p]$ are a partition of $A \boxrightarrow [P=p]$. So we have

$$(14)\quad \Sigma_p C(A \boxrightarrow [P=p])p(S) = \Sigma_K C(K)C(S/AK)$$

Substituting (14) into (7) gives us a formula defining expected utility in terms of counterfactuals with chance propositions as consequents:

$$(15)\quad U(A) = \Sigma_S \Sigma_p C(A \boxrightarrow [P=p])p(S)V(AS)$$

For any S and any number q from zero to one, let $[P(S)=q]$ be the proposition that holds at just those worlds where the chance of S, at the time when the agent realises his option, is q. It is the disjunction of those [P=p]'s for which $p(S)$ equals q. We can lump together counterfactuals in (14) and (15) to obtain reformulations in which the consequents concern chances of single S's:

$$(16)\quad \Sigma_q C(A \boxrightarrow [P(S)=q])q = \Sigma_K C(K)C(S/AK),$$

$$(17)\quad U(A) = \Sigma_S \Sigma_q C(A \boxrightarrow [P(S)=q])qV(AS)$$

There are various ways to mix probabilities and counterfactuals. I have argued that when things are chancy, it isn't good enough to take credences of plain $A \boxrightarrow S$ counterfactuals. The counterfactuals themselves must be made probabilistic. I have made them so by giving them chance propositions as consequents. Sobel makes them so in a different way: as we noted in Section 7, he puts the probability in the connec-

[24] That follows by what I call the Principal Principle connecting chance and credence, on the assumption that (i) AK holds or fails to hold at any world entirely in virtue of the history of that world up to action time together with the complete theory of chance for that world, and (ii) the agent gives no credence to worlds where the usual asymmetries of time break down. Part (ii) fails in the case which we have already noted in Section 7 as troublesome, in which the agent thinks he may have foreknowledge of the outcomes of chance processes. See "A Subjectivist's Guide to Objective Chance," pp. 266–276 [Pages 86–96 in this volume.]

tive Under our present assumptions (and setting aside extraordinary worlds where the common asymmetries of time break down), the two approaches are equivalent Sobel's quantitative counterfactual with a plain consequent

If it were that A, it would be with probability q that S

holds at W iff $W_A(S)$ equals q Given my derivation of the imaging function from the dependency hypotheses, that is so iff $C(S/AK_W)$ equals q That is so (setting aside the extraordinary worlds) iff K_W implies $A \boxrightarrow [P(S)=q]$ Given that there is a probabilistic full pattern for each world, that is so iff $A \boxrightarrow [P(S)=q]$ holds at W Hence the Sobel quantitative counterfactual with a plain consequent is the same proposition as the corresponding plain counterfactual with a chance consequent If ever we must retract the assumption that there is a probabilistic full pattern for each world (or if we want to take the extraordinary worlds into account), the two approaches will separate and we may need to choose, but let us cross that bridge if we come to it

11 THE HUNTER-RICHTER PROBLEM

That concludes an exposition and survey of causal decision theory In this final section, I wish to defend it against an objection raised by Daniel Hunter and Reed Richter [25] Their target is the Gibbard-Harper version, but it depends on nothing that is special to that version, so I shall restate it as an objection against causal decision theory generally

Suppose you are one player in a two-person game Each player can play red, play white, play blue, or not play If both play the same colour, each gets a thousand dollars, if they play different colours, each loses a thousand dollars, if one or both don't play, the game is off and no money changes hands Value goes by money, the game is played only once, there is no communication or prearrangement between the players, and there is nothing to give a hint in favour of one colour or another—no "Whites rule OK!" sign placed where both can see that both can see it, or the like So far, this game seems not worthwhile But you have been persuaded that you and the other player are very much alike psychologically and hence very likely to choose alike, so that you are much more likely to play and win than to play and lose Is it rational for you to play?

[25] Counterfactuals and Newcomb s Paradox *Synthese* 39 (1978) pp 249–261, especially pp 257–259

Yes So say I, so say Hunter and Richter, and so (for what it is worth) says noncausal decision theory But causal decision theory seems to say that it is not rational to play If it says that, it is wrong and stands refuted It seems that you have four dependency hypotheses to consider, corresponding to the four ways your partner might play

K_1 Whatever you do, he would play red,
K_2 Whatever you do, he would play white,
K_3 Whatever you do, he would play blue,
K_4 Whatever you do, he would not play

By the symmetry of the situation, K_1 and K_2 and K_3 should get equal credence Then the expected utility of not playing is zero, whereas the expected utilities of playing the three colours are equal and negative So we seem to reach the unwelcome conclusion that not playing is your U-maximal option

I reply that Hunter and Richter have gone wrong by misrepresenting your partition of options Imagine that you have a servant You can play red, white, or blue, you can not play, or you can tell your servant to play for you The fifth option, delegating the choice, might be the one that beats not playing and makes it rational to play Given the servant, each of our previous dependency hypotheses splits in three For instance K_1 splits into

$K_{1\,1}$ Whatever you do, your partner would play red, and your servant would play red if you delegated the choice,
$K_{1\,2}$ Whatever you do, your partner would play red, and your servant would play white if you delegated the choice,
$K_{1\,3}$ Whatever you do, your partner would play red, and your servant would play blue if you delegated the choice

(If you and your partner are much alike, he too has a servant, so we can split further by dividing the case in which he plays red, for instance, into the case in which he plays red for himself and the case in which he delegates his choice and his servant plays red for him However, that difference doesn't matter to you and is outside your influence, so let us disregard it) The information that you and your partner (and your respective servants) are much alike might persuade you to give little credence to the dependency hypotheses $K_{1\,2}$ and $K_{1\,3}$ but to give more to $K_{1\,1}$, and likewise for the subdivisions of K_2 and K_3 Then you give your credence mostly to dependency hypotheses according to which you would either win or break even by delegating your choice Then

causal decision theory does not tell you, wrongly, that it is rational not to play Playing by delegating your choice is your U-maximal option

But you don't have a servant What of it? You must have a tie-breaking procedure There must be something or other that you do after deliberation that ends in a tie Delegating your choice to your tie-breaking procedure is a fifth option for you, just as delegating it to your servant would be if you had one If you are persuaded that you will probably win if you play because you and your partner are alike psychologically, it must be because you are persuaded that your tie-breaking procedures are alike You could scarcely think that the two of you are likely to coordinate *without* resorting to your tie-breaking procedures, since *ex hypothesi* the situation plainly *is* a tie! So you have a fifth option, and as the story is told, it has greater expected utility than not playing This is not the option of playing red, or white, or blue, straightway at the end of deliberation, although if you choose it you will indeed end up playing red or white or blue What makes it a different option is that it interposes something extra—something other than deliberation—after you are done deliberating and before you play [26]

Postscript to

"Causal Decision Theory"

REPLY TO RABINOWICZ

In a recent article, Włodzimierz Rabinowicz carries the comparison between my theory and Sobel's farther than I had done [1] He also

[26] This paper is based on a talk given at a conference on Conditional Expected Utility at the University of Pittsburgh in November 1978 It has benefited from discussions and correspondence with Nancy Cartwright, Allan Gibbard, William Harper, Daniel Hunter, Frank Jackson, Richard Jeffrey, Gregory Kavka, Reed Richter Brian Skyrms, J Howard Sobel and Robert Stalnaker

[1] Two Causal Decision Theories Lewis vs Sobel in Tom Pauli et al , eds *320311 Philosophical Essays Dedicated to Lennart Åqvist on his Fiftieth Birthday* (Uppsala Filosofiska Studier, 1982)

advances two criticisms against my discussion One uncovers a clear mistake on my part, but the other rests on a misunderstanding

First the mistake Suppose we start, as Sobel does, with the imaging function—in Sobel's terminology, the tendencies of worlds—and we take equivalence classes under the relation of imaging alike Call these classes *tendency propositions* I suggested that these should turn out to be the same as my dependency hypotheses Rabinowicz rightly objects (p 311) *Distinguo* let a *practical* dependency hypothesis be a maximally specific proposition about how the things the agent cares about do and do not depend causally on his present actions, let a *full* dependency hypothesis be a maximally specific proposition about how all things whatever do and do not depend causally on the agent's present actions By my definition, a "dependency hypothesis" is a practical dependency hypothesis, whereas a tendency propostition is, if anything, not a practical but a full dependency hypothesis Luckily my mistake does not damage my discussion, since it would have made no difference if I had worked in terms of full rather than practical dependency hypotheses throughout

Next the misunderstanding I had presupposed (1) that any option would be compatible with any dependency hypothesis, I had also supposed (2) that at least sometimes, an image of a world on a proposition would be "blurred", dividing its probability over several worlds My discussion of counterfactuals elsewhere indicated that I also accept (3) an assumption of "centering " But Rabinowicz shows that (1), (2), and (3) are inconsistent (Theorem 1, p 313) This looks like trouble for me Not so—*distinguo* again

(3A) *Centering of the imaging function* is the thesis that whenever a proposition A holds at a world W, the image of W on A is the distribution that puts all its probability on world W

(3B) *Centering of counterfactuals* is the thesis that whenever a proposition A holds at a world W, a "would" counterfactual with antecedent A holds at W iff its consequent does

What Rabinowicz shows is that (1), (2) and (3A) are inconsistent What my discussion of counterfactuals indicates is that I accept (3B) I do indeed But I reject (3A), therefore, Rabinowicz's difficulties for decision theory with a centered imaging function are no threat to me

Sobel discusses counterfactuals and decision theory together, using the same apparatus of imaging (or "tendency") functions His theory of counterfactuals says that a "would" counterfactual with antecedent

A holds at a world W iff its consequent holds at every world to which the image of W on A assigns positive probability That means that for Sobel, centering of the imaging function and of counterfactuals are equivalent Not so for me I might have done well to warn the reader that I disagree with Sobel on this point, though strictly speaking a disagreement about counterfactuals is irrelevant to the comparison of our decision theories

Example A coin is about to be tossed (proposition A) The coin will be tossed fairly, with equal chance of heads and tails It will in fact fall heads Then I say that the image of our world on A is blurred, not centered it distributes probability equally between heads-worlds (among them ours) and tails-worlds But I also say, by centering of counterfactuals, that if it were that A—as is in fact the case—then the coin would fall heads *Contra* Sobel's theory, this counterfactual holds although its consequent does not hold at all worlds to which the image assigns positive probability

(I don't deny that if it were that A, then there would be some chance that the coin would fall tails For this too follows from centering of counterfactuals There would be some chance of it, but it would not happen I say that the counterfactuals about outcomes and the counterfactuals about chances are compatible For further discussion, see Postscript D to "Counterfactual Dependence and Time's Arrow," in this volume)

TWENTY-EIGHT

Utilitarianism and Truthfulness[1]

A demon has seized two highly rational act-utilitarians—call them "You" and "I"—and put them in separate rooms In each room there are two buttons, a red one and a green one The demon has arranged that by both pushing our red buttons or by both pushing our green buttons we bring about the Good, but by pushing one red button and one green button (or by pushing both buttons or neither button in one of the rooms) we bring about the Bad The demon has made sure that we both know all the facts I have listed so far, that we both know that we both know them, and so on

You manage to send me a message, and the message is "I pushed red " But, strange to say, that does not help For I reason as follows "You are a highly rational utilitarian You act in whatever way you think will have the best consequences, with no regard to any other consideration This goes for sending messages you send whatever message you think will have the best consequences, caring not at all about truthfulness for its own sake So I have not the slightest reason to believe your message unless I have reason to believe that you think truthfulness will have the best consequences In this case, you must know that truthfulness has the best consequences only if I have some

[1] This research was supported by a fellowship from the American Council of Learned Societies

reason to believe you and to act accordingly If not, there is nothing to choose between the expected consequences of truth and untruth, so you have no reason whatever to choose truth rather than untruth I have not the slightest reason to believe you unless I have reason to believe that you think that I have reason to believe you But I know that you—knowledgeable and rational creature that you are—will not think that I have reason to believe you unless I really do have Do I? *I cannot show that I have reason to believe you without first assuming what is to be shown that I have reason to believe you* So I cannot, without committing the fallacy of *petitio principii*, show that I have reason to believe you Therefore I do not Your message gives me not the slightest reason to believe that you pushed red, and not the slightest reason to push red myself" Arguing thus, I push at random By chance I push green

Such is the disutility of utilitarianism, according to D H Hodgson [2]

We might better say such is the disutility of *expecting* utilitarianism, and it is not sufficiently compensated by the efforts to maximize utility that fulfil the expectation Hodgson says that knowledgeable and rational act-utilitarians would have no reason to expect one another to be truthful, not even when the combination of truthfulness with expectation of truthfulness would have good consequences, so they would forfeit the benefits of communication Similarly they would forfeit the benefits of promising, for an example of this, just change the message in my example to "I will push red" More generally, it seems that Hodgson's utilitarians would forfeit the benefits of all the conventions whereby we coordinate our actions to serve our common interests The conventions of truthfulness and of promise-keeping are but two of these

But to talk myself into ignoring your message "I pushed red" is absurd My example has no special features, it is just a simple and stark instance of the general situation Hodgson says would prevail among knowledgeable and rational act-utilitarians I conclude that Hodgson is wrong in general Where, then, is the flaw in my Hodgsonian argument that I ought to ignore your message? Every step up to the italicised one seems true, and every step beyond that seems false

I think the argument went wrong when I tacitly assumed that I could not have reason to believe you unless I could show, using nothing but the facts set forth in the first paragraph—our situation, our utilitarianism and rationality, our knowledge of these, our knowledge

[2] *Consequences of Utilitarianism* (Oxford University Press Oxford, 1967), pp 38–46

of one another's knowledge of these, and so on—that I did have reason to believe you But why must my premises be limited to these? I should not use any premise that is inconsistent with the facts of the first paragraph, but there is nothing wrong with using a premise that is independent of these facts, if such a premise is available

The premise that you will be truthful (whenever it is best to instill in me true beliefs about matters you have knowledge of, as in this case) is just such a premise It *is* available to me At least, common sense suggests that it would be, and our only reason to suppose that it would not is the Hodgsonian argument we are now disputing It is independent of the facts listed in the first paragraph On the one hand, it is *consistent* with our rationality and utilitarianism, our knowledge thereof, and so on For if you are truthful (except when it is best that I should have false beliefs), and if I expect you to be, and if you expect me to expect you to be, and so on, then you will have a good utilitarian reason to be truthful You will be truthful without compromising your utilitarianism and without adding to your utilitarianism an independent maxim of truthfulness On the other hand, it is not *implied* by our rationality and utilitarianism, our knowledge thereof, and so on For if you are systematically untruthful (expect when it is best that I should have false beliefs), and if I expect you to be, and if you expect me to expect you to be, and so on, then you will have a good utilitarian reason to be untruthful I am speaking, of course, of truthfulness and untruthfulness *in English*, I should mention that systematic untruthfulness in English is the same thing as systematic truthfulness in a different language *anti-English*, exactly like English in syntax but exactly opposite in truth conditions

Therefore I should have decided that I did have reason to believe your message and to push red myself This reason is admittedly not premised merely on our situation, our rationality and utilitarianism, our knowledge of these, and so on But it is premised on further knowledge that I do in fact possess, and that is perfectly consistent with these facts

Bibliography of the Writings of David Lewis

1966

"An Argument for the Identity Theory," *Journal of Philosophy* 63 (1966) 17–25, reprinted with additions in David Rosenthal, ed., *Materialism and the Mind-Body Problem* (Prentice-Hall, 1971), German translation by Andreas Kemmerling published as "Eine Argumentation fur die Identitatstheorie," in Ansgar Beckermann, ed., *Analytische Handlungstheorie*, Volume II (Suhrkamp Verlag, 1977), Spanish translation by Enrique Villaneuva published as "Un argumento en favor de la teoria de la identidad," *Cuadernos de Critica* No 30 (Instituto de Investigaciones Filosoficas, 1984), reprinted with additions in David Lewis, *Philosphical Papers*, Volume I

"Percepts and Color Mosaics in Visual Experience," *Philosophical Review* 75 (1966) 357–68

Abstract Roderick Firth, in his Sense-Data and the Percept Theory," opposes his own theory that visual experience consists of *percepts* of ostensible external objects or facts to the old-fashioned theory that it consists of a mosaic of color-spots I claim to resolve the opposition thus Call two visual experiences *modification-equivalent* if they are connected by a chain of actual or possible visual experiences such that, given any two adjacent members of the chain, one can pass from one to the other with change of percept but no percept of change We often ignore differences between modification-equivalent visual experiences Some visual experience, at least, consists of a percept of a color-mosaic, it is plausible that any visual experience is modification-equivalent to experiences of some

definite color-mosaic If so, then any visual experience is indistinguishable from color mosaic experience for some practical purposes

'Scriven on Human Unpredictability," *Philosophical Studies* 17 (1966) 69–74 (Jane S Richardson, co-author)

Abstract Michael Scriven has argued that we are unpredictable if I want to foil your attempts to predict me, I can in principle replicate your prediction and do the opposite But Scriven assumes that it is possible both that you have time to finish your prediction (else your failure is of no significance) and that I have time to finish my replication (else you might not fail) This assumption is suspect, since the times consumed by our two tasks are increasing functions of each other

1968

'Counterpart Theory and Quantified Modal Logic," *Journal of Philosophy* 65 (1968) 113–26, reprinted in Michael J Loux, ed , *The Possible and the Actual* (Cornell University Press, 1979), reprinted in David Lewis, *Philosophical Papers*, Volume I

Finitude and Infinitude in the Atomic Calculus of Individuals," *Nous* 2 (1968) 405–10 (Wilfrid Hodges, co-author)

Abstract Nelson Goodman has raised the question whether there is any sentence in the language of his calculus of individuals which is true in every finite intended model no matter how large, but false in every infinite atomic intended model We prove that there is no such sentence This negative answer to Goodman's question is obtained as a corollary to the following theorem any sentence in the language is equivalent—under certain axioms which hold in every atomic intended model—to a truth-functional compound of sentences setting lower limits on the number of atomic individuals

1969

"Policing the Aufbau," *Philosophical Studies* 20 (1969) 13–17

Abstract The method of construction employed in Carnap's 'Logische Aufbau der Welt" is not certain to work properly in every case, as Carnap, Goodman, and others have shown, but certain error-detecting procedures can be added to the construction which ought to increase the rate of success

Review of Capitan and Merrill, eds , *Art, Mind, and Religion, Journal of Philosophy* 66 (1969) 22–27, excerpt reprinted in Ned Block, ed , *Readings in the Philosophy of Psychology*, Volume I (Harvard University Press, 1980)

Abstract Concerning 'Psychological Predicates," by Hilary Putnam I

respond to Putnam's principal objection against the identity of mental and brain states by suggesting that mental terms may denote different brain states in the case of different species, or even different individuals

I briefly discuss other papers in the collection

Convention A Philosophical Study (Harvard University Press, 1969, reprinted, Blackwell and Harvard University Press, 1986), preliminary version titled *Conventions of Language* submitted as a doctoral dissertation (Harvard, 1966), Italian translation by Gabriele Usberti published as *La Convenzione* (Bompiani, 1974), excerpt reprinted in Gilbert Harman, ed , *On Noam Chomsky Critical Essays* (Anchor, 1974), German translation by Roland Posner and Detlef Wenzel published as *Konvention Eine Sprachphilosophische Abhandlung* (Walter de Gruyter, 1975)

Abstract Social conventions are analyzed, roughly, as regularities in the solution of recurrent coordination problems—situations of interdependent decision in which common interest predominates An example is our regularity of driving on the right each does so to coordinate with his fellow drivers, but we would have been just as well off to coordinate by all driving on the left Other examples are discussed, conventions are contrasted with other sorts of regularities, conventions governing systems of communication are singled out for special attention It is shown that the latter can be described as conventions to be truthful with respect to a particular assignment of truth conditions to sentences or other units of communication

"Lucas against Mechanism," *Philosophy* 44 (1969) 231–33

Abstract J R Lucas's contention (in ' Minds, Machines, and Godel") that his potential output of truths of arithmetic cannot be duplicated by a machine is true to this extent a certain infinitary inference rule, which Lucas can recognize to be truth-preserving, will yield a set of truths of arithmetic which cannot be the potential output of a machine However, there is no reason to believe that Lucas can verify, in every case, that a sentence of arithmetic is one of the ones that the rule yields Therefore it may yet be that Lucas's output could be duplicated by a machine

1970

"Anselm and Actuality, ' *Nous* 4 (1970) 175–88, reprinted in Baruch A Brody, ed , *Readings in the Philosophy of Religion* (Prentice-Hall, 1974), reprinted in David Lewis, *Philosophical Papers*, Volume I

How to Define Theoretical Terms," *Journal of Philosophy* 67 (1970) 427–46, reprinted in David Lewis, *Philosophical Papers*, Volume I

"Holes, *Australasian Journal of Philosophy* 48 (1970) 206–12 (Stephanie R Lewis, co-author), reprinted in David Lewis, *Philosophical Papers*, Volume I

"Nominalistic Set Theory," *Nous* 4 (1970) 225–40

Abstract Taking as primitive the calculus of individuals and a relation of nextness between atoms, it is possible to define several pseudo-membership relations between individuals These relations have many of the properties of the membership relation, and can be employed to provide nominalistic counterparts of various standard set-theoretic constructions

General Semantics", *Synthese* 22 (1970) 18–67, reprinted in Donald Davidson and Gilbert Harman, eds , *Semantics of Natural Language* (Reidel, 1972), Italian translation by Ugo Volli of an excerpt published as "Semantica Generale in Andrea Bonomi, ed , *La Struttura Logica del Linguaggio* (Bompiani, 1973), reprinted in Barbara Partee, ed , *Montague Grammar* (Academic Press, 1976), Spanish translation by Alejandro Herrera Ibañez published as "Semantica general,' *Cuadernos de Critica* No 29 (Instituto de Investigaciones Filosoficas, 1984), reprinted in David Lewis, *Philosophical Papers*, Volume I

1971

Immodest Inductive Methods,' *Philosophy of Science* 38 (1971) 54–63

Abstract Inductive methods can be used to estimate the accuracies of inductive methods Call a method *immodest* if it estimates that it is at least as accurate as any of its rivals It would be unreasonable to adopt any but an immodest method Under certain assumptions, exactly one of Carnap's lambda-methods is immodest This may seem to solve the problem of choosing among the lambda-methods, but sometimes the immodest lambda-method is lambda = 0, which it would not be reasonable to adopt We should therefore reconsider the assumptions that led to this conclusion for instance, the measure of accuracy

'Counterparts of Persons and Their Bodies," *Journal of Philosophy* 68 (1971) 203–11, reprinted in David Lewis, *Philosophical Papers*, Volume I

'Analog and Digital," *Nous* 5 (1971) 321–27

Abstract Counterexamples are offered to Nelson Goodman s proposal, in *Languages of Art*, that the difference between analog and digital systems of representation is the difference between dense systems and differentiated ones Alternative definitions of analog and digital representations are proposed

"Completeness and Decidability of Three Logics of Counterfactual Conditionals, ' *Theoria* 37 (1971) 74–85

Abstract Three axiomatic systems for the counterfactual conditional connective are given one equivalent to the system C2 of Stalnaker and Thomason, and two weaker systems The systems are shown to be sound and complete (under various combinations of conditions) if the counterfactual is taken to be true at a world if and only if, roughly, the consequent is true in those of the worlds where the antecedent is true that are closest in similarity to the given world Further, the systems are decidable

1972

"Utilitarianism and Truthfulness," *Australasian Journal of Philosophy* 50 (1972) 17–19, reprinted in this volume

"Psychophysical and Theoretical Identifications, *Australasian Journal of Philosophy* 50 (1972) 249–58, reprinted in Chung-ying Cheng, ed , *Philosophical Aspects of the Mind-Body Problem* (University Press of Hawaii, 1975), reprinted in Ned Block, ed , *Readings in the Philosophy of Psychology*, Volume I (Harvard University Press, 1980)

Abstract The psychophysical identity theory may be subsumed under a general account of the meaning of theoretical terms and the nature of theoretical identifications, as follows Theoretical terms, by their meaning, denote whichever entities uniquely realize the theory that introduced them, by learning which entities do so, we can establish an identification In particular, the names of mental states denote whichever entities uniquely realize common-sense psychology, if certain neural states do so, they must be identical to the mental states

1973

Counterfactuals (Blackwell and Harvard University Press, 1973, revised printing, 1986), excerpt reprinted as 'Possible Worlds," in Michael J Loux, ed , *The Possible and the Actual* (Cornell University Press, 1979)

Abstract A counterfactual conditional has the form if it were that A, then it would be that B (where A is usually assumed false) What does this mean? Roughly in certain possible worlds where A holds, B holds also But which A-worlds should we consider? Not all, those that differ gratuitously from our actual world should be ignored Not those that differ from our world only in that A holds, for no two worlds can differ in one respect only Rather, we should consider the A-worlds most similar, overall, to our world If there are no most similar A-worlds, then we should consider whether some A-world where B holds is more similar to ours than any where B does not hold

An analysis of counterfactuals is given along these lines It is shown to admit of various formulations It is compared with other theories of counterfactuals Its foundations, in comparative similarity of possible worlds, are defended Analogies are drawn between counterfactuals, thus analyzed, and other concepts An axiomatic logic of counterfactuals is given

' Causation, ' *Journal of Philosophy* 70 (1973) 556–67, reprinted in Ernest Sosa, ed , *Causation and Conditionals* (Oxford University Press, 1975), German translation by Gunter Posch (with additions) published as Kausalitat," in Gunter Posch, ed , *Kausalitat—Neue Texte* (Philip Reclam, 1981), reprinted in this volume

'Counterfactuals and Comparative Possibility," *Journal of Philosophical Logic* 2 (1973) 418–46, reprinted in Donald Hockney et al , eds , *Contemporary Research in Philosophical Logic and Linguistic Semantics* (Reidel, 1975), Italian translation by Claudio Pizzi published as "Controfattuali e possibilita comparativa," in Claudio Pizzi, ed , *Leggi di natura, modalita, ipotesi* (Feltrinelli, 1978), reprinted in W L Harper et al , eds , *Ifs* (Reidel, 1981), reprinted in this volume

1974

"Semantic Analyses for Dyadic Deontic Logic," in Soren Stenlund, ed , *Logical Theory and Semantic Analysis Essays Dedicated to Stig Kanger on His Fiftieth Birthday* (Reidel, 1974)

Abstract According to one conception of deontic conditionals, "Ought *A* given *B*" means roughly that *A* holds at the best of the worlds where *B* holds I compare different ways of developing this approach to the semantics of dyadic deontic logic, seeking to distinguish deep from superficial differences

' Spielman and Lewis on Inductive Immodesty," *Philosophy of Science* 41 (1974) 84–85

Abstract Recent theorems on inductive immodesty due to S Spielman and D Lewis appear to be contradictory when applied to the case of null evidence Spielman's theorem implies that every method in Carnap s continuum is immodest in this case, whereas Lewis's theorem implies that the straight rule alone is The contradiction is resolved by observing that Spielman and Lewis are speaking of immodesty under slightly different measures of inductive accuracy

"Intensional Logics Without Iterative Axioms, ' *Journal of Philosophical Logic* 3 (1974) 457–66

Abstract Any classical intensional propositional logic that can be axiomatized in such a way that no intensional operator appears within the scope of another in any axiom is complete, in the sense that it is determined by the class of all classical frames with unrestricted valuations that validate it

'Radical Interpretation," *Synthese* 23 (1974) 331–44, reprinted in David Lewis, *Philosophical Papers*, Volume I

' Tensions," in Milton K Munitz and Peter K Unger, eds, *Semantics and Philosophy* (New York University Press, 1974), reprinted in David Lewis, *Philosophical Papers*, Volume I

1975

"Languages and Language," in Keith Gunderson, ed , *Minnesota Studies in the Philosophy of Science*, Volume VII (University of Minnesota Press, 1975), Italian translation by Ugo Volli of a preliminary version published as

"Lingue e Lingua," *Versus* 4 (1973) 2–21, excerpt preprinted in Gilbert Harman, ed., *On Noam Chomsky: Critical Essays* (Anchor, 1974), German translation by Georg Meggle published as 'Die Sprachen und die Sprache," in Georg Meggle, ed., *Handlung, Kommunikation, Bedeutung* (Suhrkamp Verlag, 1979), reprinted in A. P. Martinich, ed., *The Philosophy of Language* (Oxford University Press, 1985), reprinted in David Lewis, *Philosophical Papers*, Volume I

"Adverbs of Quantification," in Edward L. Keenan, ed., *Formal Semantics of Natural Language* (Cambridge University Press, 1975)

Abstract Such adverbs as 'always," 'sometimes," "never," "usually," "often," and 'seldom" are quantifiers, but often, contrary to first impression, they do not quantify over moments of time. They can be seen as unselective quantifiers, binding all variables in their scopes. Various sentences, including some notorious puzzlers, can be seen as transformed versions of sentences formed using adverbs of quantification with restrictive "if"-clauses.

Review of Olson and Paul, *Contemporary Philosophy in Scandinavia*, *Theoria* 41 (1975) 39–60 (Stephanie R. Lewis, co-author)

Abstract Concerning "Rights and Parliamentarianism," by Stig and Helle Kanger. We question whether their taxonomy of rights covers rights versus the world at large.

Concerning "On the Analysis and Logic of Questions," by Lennart Åqvist. We argue that Åqvist's imperative-epistemic analysis should give way to an imperative-assertoric analysis. Where Åqvist says 'Let it be that I know ," we suggest 'Let it be that you tell me ." Either way, we note that questions may join other imperatives as deontic sentences made true by their appropriate utterance.

Concerning 'Decision-theoretic Approaches to Rules of Acceptance," by Risto Hilpinen. We pose a dilemma: Does the agent have a full system of quantitative degrees of belief? If so, why does he also need non-quantitative acceptance? If not, how can he govern his acceptances by decision-theoretic rules?

We briefly discuss some other papers in the collection.

1976

Convention: Reply to Jamieson," *Canadian Journal of Philosophy* 6 (1976) 113–20

Abstract Several proposed counterexamples against my analysis of social convention are considered. Some exemplify derivative usages of the term "convention." The rest fail either (1) through disregarding unconscious preferences and expectations, or (2) through disregarding the relativity of conventions to populations, or (3) through confusing conditional preferences with conditionals about preferences.

The Paradoxes of Time Travel," *American Philosophical Quarterly* 13 (1976) 145–52, reprinted in Fred D Miller, Jr , and Nicholas D Smith, eds , *Thought Probes* (Prentice-Hall, 1981), reprinted in this volume

'Probabilities of Conditionals and Conditional Probabilities," *Philosophical Review* 85 (1976) 297–315, reprinted in W L Harper et al , eds , *Ifs* (Reidel, 1981), reprinted in this volume

'Survival and Identity," in Amelie O Rorty, ed , *The Identities of Persons* (University of California Press, 1976), German translation by Thomas Nenon published as ' Uberleben und Identitat,' in Ludwig Siep, ed , *Identitat der Person* (Schwabe, 1983), Spanish translation by Mercedes Garcia Oteyza published as ' Supervivencia e identidad," *Cuadernos de Critica* No 27 (Instituto de Investigaciones Filosoficas, 1984), reprinted in David Lewis, *Philosophical Papers*, Volume I

1977

Possible-World Semantics for Counterfactual Logics A Rejoinder," *Journal of Philosophical Logic* 6 (1977) 359–63

Abstract Ellis, Jackson, and Pargetter have claimed that a certain feature of the logic of counterfactual conditionals—namely, the apparent validity of the inference from "if *A* or *B*, then *C*" to "if *A* then *C*"—cannot be accounted for by any sort of possible-world semantics However, no less than three solutions to their problem have already been proposed by Fine and others, and they have given no reason to reject any of the three

1978

'Truth in Fiction," *American Philosophical Quarterly* 15 (1978) 37–46, reprinted in David Lewis, *Philosophical Papers*, Volume I

Reply to McMichael, *Analysis* 38 (1978) 85–86

Abstract McMichael showed that my semantics for deontic conditionals, when applied to a ranking of worlds on radically utilitarian principles, yields counterintuitive results I concur, but suggest that it is the radical utilitarianism—not the semantics—that goes against our common opinions

1979

A Problem about Permission, in E Saarinen et al , eds , *Essays in Honour of Jaakko Hintikka* (Reidel, 1975)

Abstract Brian Chellas has produced a semantic analysis for imperative and permissive sentences, modeled on the standard possible world semantics for deontic modalities I show how to incorporate this semantic analysis into the description of a language game of commanding and

permitting The crucial rule of the game stipulates that the permissibility of possible worlds shall change when commands and permissions are given, in such a way that any imperative or permissive sentence uttered by someone in authority shall be true under Chellas's semantics A precise formulation of this rule is easy in the case of commands, the case of permissions, however, is problematic

'Prisoners' Dilemma is a Newcomb Problem," *Philosophy and Public Affairs* 8 (1979) 235–40, reprinted in R Campbell and L Sowden, eds , *Paradoxes of Rationality and Cooperation Prisoner's Dilemma and Newcomb's Problem* (University of British Columbia Press, 1985), reprinted in this volume

"Counterfactual Dependence and Time s Arrow,' *Nous* 13 (1979) 455–76, reprinted in this volume

"Scorekeeping in a Language Game," *Journal of Philosophical Logic* 8 (1979) 339–59, reprinted in R Bauerle et al , eds , *Semantics from Different Points of View* (Springer-Verlag, 1970), reprinted in David Lewis, *Philosophical Papers*, Volume I

"Attitudes *De Dicto* and *De Se*," *Philosophical Review* 88 (1979) 513–43, reprinted in D L Boyer et al , eds , *The Philosopher's Annual*, Volume III (Ridgeview, 1981), reprinted in David Lewis, *Philosophical Papers*, Volume I

'Lucas Against Mechanism II," *Canadian Journal of Philosophy* 9 (1979) 373–76

Abstract Lucas insists on the dialectical character of his Godelian refutation of mechanism This means that his arithmetical output depends on the mechanistic accusation he is out to refute Then if he is a machine, he is one that responds to input But if he is such a machine, there is no reason to think that his arithmetical output (when responding to a mechanistic accusation in the way he intends to) is true or consistent, hence no reason to doubt that it might contain a false sentence expressing its own consistency The refutation therefore fails

1980

'A Subjectivist's Guide to Objective Chance," in Richard C Jeffrey, ed , *Studies in Inductive Logic and Probability*, Volume II (University of California Press, 1980), reprinted in W L Harper et al , eds , *Ifs* (Reidel, 1981), reprinted in this volume

' Mad Pain and Martian Pain," in Ned Block, ed , *Readings in Philosophy of Psychology*, Volume I (Harvard University Press, 1980), reprinted in David Lewis, *Philosophical Papers*, Volume I

' Index, Context, and Content," in Stig Kanger and Sven Ohman, eds , *Philosophy and Grammar* (Reidel, 1980)

Abstract A context is a location—time, place, world—in which a sen-

tence may be said, an index is an n-tuple of features of context that can vary independently I argue that semantics of natural language must involve both context-dependence and index-dependence, neither can replace the other I also argue that two different strategies for combining the two dependences differ only superficially

Veridical Hallucination and Prosthetic Vision,' *Australasian Journal of Philosophy* 58 (1980) 239–49, reprinted in this volume

1981

"Causal Decision Theory,' *Australasian Journal of Philosophy* 59 (1981) 5–30, reprinted in this volume

Ordering Semantics and Premise Semantics for Counterfactuals, *Journal of Philosophical Logic* 10 (1981) 217–34

Abstract The analysis of counterfactual conditionals requires some device for taking account of factual background Orderings of worlds, perhaps partial, may be used, as in the theories of Stalnaker, Lewis, and Pollock, or premise sets, as in the theory of Kratzer The two approaches are shown to be equivalent

What Puzzling Pierre Does Not Believe,' *Australasian Journal of Philosophy* 59 (1981) 283–89

Abstract Kripke s puzzle about belief refutes a certain simple analysis of belief sentences The analysis fails for another reason as well, since it requires believers to have a knowledge of essences which they do not in fact possess

'Why Ain'cha Rich?'" *Nous* 15 (1981) 377–80

Abstract Under the conception of rationality favored by two-boxers, Newcomb s problem is an arrangement in which predicted irrationality is rewarded One-boxers favor a different conception of rationality Could we devise a problem in which predicted irrationality according to this different conception is rewarded? It turns out that we could not

Are We Free To Break the Laws?" *Theoria* 47 (1981) 113–21, reprinted in this volume

1982

' Logic for Equivocators," *Nous* 16 (1982) 431–41

Abstract It has been argued that relevance must be respected in logic because irrelevant implication may not preserve truth when we are dealing with sentences that are both true and false I suggest that the best way to understand how a sentence may be both true and false is that it may be both true on some disambiguations and false on some disambiguations, and accordingly I commend a form of (partly) relevant logic to those who

fear they cannot fully disambiguate the sentences that figure in their reasoning

"Whether Report," in Tom Pauli et al., eds., *320311 Philosophical Essays Dedicated to Lennart Åqvist on his Fiftieth Birthday* (Filosofiska Studier, 1982)

Abstract By exploiting double indexing, it is possible to treat a "whether"-clause as a sentence expressing whichever is the true one of the alternative propositions presented in it, hence as a suitable argument for epistemic or assertoric modalities It is further possible to treat the 'or' s that punctuate these clauses as ordinary disjunctions

1983

"Individuation by Acquaintance and by Stipulation," *Philosophical Review* 92 (1983) 3–32, reprinted in Fred Landman and Frank Veltman, eds., *Varieties of Formal Semantics Groningen-Amsterdam Studies in Semantics III* (Foris Publications, 1984)

Abstract Hintikka has demonstrated the importance of cross-identification by acquaintance, in which individuals that figure in alternative possibilities are united by likeness in their relations to a subject of attitudes This requires prior cross-identification of the subject, which cannot be either by acquaintance or by description The problem is solved if we take the alternatives not as possible worlds but as possible individuals situated in worlds

Philosophical Papers, Volume I (Oxford University Press, 1983)

"Extrinsic Properties," *Philosophical Studies* 44 (1983) 197–200

Abstract Kim has suggested, roughly, that an extrinsic property is a property that implies accompaniment a property that could not belong to a thing unless some other, distinct thing coexisted with it I offer counter-examples to Kim's proposal and to certain near relatives of it

"New Work For a Theory of Universals," *Australasian Journal of Philosophy* 61 (1983) 343–77, reprinted in P Athay et al., eds., *The Philosopher's Annual*, Volume VI (Ridgeview, 1985)

Abstract D M Armstrong puts forward his theory of universals as a solution to the problem of one over many But this problem, depending on how we understand it, either admits of nominalistic solutions or else admits of no solution of any kind Nevertheless, Armstrong s theory meets other urgent needs in systematic philosophy its very sparing admission of genuine universals offers us a means to make sense of several otherwise elusive distinctions

"Levi Against U-Maximization," *Journal of Philosophy* 80 (1983) 531–34

Abstract Isaac Levi claims that Gibbard and Harper's theory of U-maximizing, unless clarified by the addition of further principles, yields contradictory recommendations But the only "further principles '

needed are (1) a prohibition against fallacies of equivocation, and (2) a stipulation, already made explicitly by Gibbard and Harper, that outcomes are completely specific with respect to the agent's concerns

1984

"Devil s Bargains and the Real World," in Douglas MacLean, ed , *The Security Gamble Deterrence in the Nuclear Age* (Rowman and Allenheld, 1984)

Abstract I agree with Kavka, against Kenny and Gauthier, that in some hypothetical cases it is not wrong to form an effective conditional intention to retaliate, even though it would be wrong to fulfill that intention by retaliating I compare such cases with a Devil's bargain in which a hero volunteers for damnation to buy salvation for seven others But the most important thing to say about the Devil's bargain is that the case is bogus, and likewise for cases of paradoxical" deterrence Fascinating though they may be, they have no place in serious discussions of public policy

Putnam's Paradox," *Australasian Journal of Philosophy* 62 (1984) 221–36

Abstract Putnam's "model-theoretic argument against metaphysical realism" is a correct refutation of a global description theory of reference It demonstrates that if, as we usually suppose, we achieve more-or-less determinate reference, that must be so in virtue of constraints not established by our own stipulation—perhaps, as Merrill has suggested, constraints based on an objective discrimination between things and classes which are more and less eligible to serve as referents

1986

On the Plurality of Worlds (Blackwell, 1986)

Abstract We ought to believe in other possible worlds and individuals because systematic philosophy goes more smoothly in many ways if we do, the reason parallels the mathematicians' reason for believing in the set-theoretical universe By "other worlds" I mean other things of a kind with the world we are part of concrete particulars, unified by spatiotemporal unification or something analogous, sufficient in number and variety to satisfy a principle to the effect, roughly, that anything can coexist with anything I answer objections claiming that such modal realism is trivially inconsistent, or leads to paradoxes akin to those of naive set theory, or undermines the possibility of modal knowledge, or leads to scepticism or indifference or a loss of the seeming arbitrariness of things But I concede that its extreme disagreement with common opinion is a high price to pay for its advantages I therefore consider various versions of ersatz model realism, in which abstract representations are supposed to replace the other worlds, different versions suffer from different objections, and none is satisfactory Finally, I consider the so-called problem of trans-world

identity I stress a distinction between the uncontroversial thesis that things exist *according to* many worlds and the very problematic thesis that things exist *as part of* many worlds

"Against Structural Universals," *Australasian Journal of Philosophy* 64 (1986) 25–46

Abstract A structural universal is one such that, necessarily, any instance of it consists of proper parts that instantiate certain simpler universals in a certain pattern Forrest has suggested that structural universals could serve as ersatz possible worlds, Armstrong has offered several reasons why a theory of universals must accept them I distinguish three conceptions of what a structural universal is, and I raise objections against structural universals under all three conceptions I then consider whether uninstantiated structural universals, which are required by Forrest s proposal, are more problematic than instantiated ones

"A Comment on Armstrong and Forrest," *Australasian Journal of Philosophy* 64 (1986) 92–93

Abstract Armstrong and Forrest observe that my case against structural universals has equal force against ' structures composed of universals plus particular instances thereof To this I say that the friend of universals might get by without the structures Whether he can depends on what work he wants his theory to do, in particular on whether he wants it to provide truthmaking entities for all truths

Philosophical Papers, Volume II (Oxford University Press, 1986)

'Probabilities of Conditionals and Conditional Probabilities II,' *Philosophical Review* 95 (1986) 581–89

Abstract In the paper to which this is a sequel, I had shown that no uniform interpretation of $\rightarrow$ guarantees the equality $P(A \rightarrow C) = P(C/A)$ throughout a class of non-trivial probability functions closed under conditionalizing Here I extend that result to classes satisfying weaker closure conditions

Index