Clarendon Library of Logic and Philosophy

Graeme Forbes

The Metaphysics of Modality

Web Edition

Preface to the First (1985) Edition

I HAVE written this book with two purposes in mind. First, I wanted to produce something which a reader could use as a means of entry into the area of analytic metaphysics concerned with modality. Secondly, I wanted to make a contribution to the literature in this area which would be of interest to those working in it. A reader of the latter sort may recognize that much of the content of this book has appeared in rather disconnected fashion in various journal papers over the last few years, but it is my hope that even those familiar with those papers will find something new here in what I say about topics covered in them, and something worthwhile in what I say about topics on which I have not previously written. I also hope that the theory of individual essences with which this book is mainly concerned benefits from being presented as a whole in a single place.

I have tried to make as few demands for prerequisites as possible on the reader who would like a way into the general area of the metaphysics of modality, so the only real requirement for reading this book is familiarity with modern logic at the level at which most elementary symbolic logic courses are conducted. Specifically, I have assumed that the reader understands the distinction between valid and invalid arguments (argument-forms, sequents) in propositional and predicate calculus, and knows something of how this distinction can be characterized. But I have not assumed any previous acquaintance with modal logic, and the first two chapters of the book are therefore given over to introducing it to the reader. I have done this on what secretive bureaucrats call a "need to know" basis, and so I have made no attempt at completeness or even a very high degree of rigour in my presentation (for example, I have included nothing about axiomatic formulations of modal systems, since the topic is irrelevant to my philosophical purposes). Sometimes, I have tried to repair the effects of this

casualness in the longer footnotes, most of which have been written for enthusiasts, and are therefore less accessible than the main text; but this is the way of the world with footnotes.

I should also say something about my use of English quotation marks. When these surround a complex expression, they may either be performing their usual function of forming a name of that expression, or they may be functioning as quasi-quotes, depending on whether, in the context, it is more natural to think of the complex expression as a formula of some imagined formal language, or as a schema for such formulae; in the Appendix, I use corner quotes for quasi-quotation, but they would have given a cluttered look to the main text. Concomitantly, atomic sentence letters '*P*', '*Q*', etc., are quoted or unquoted depending on whether it is more natural to think of them as belonging to the lexical primitives of some formal language or as names of particular, say English, sentences.

Many people saw parts of this book in typescript form and gave me helpful comments, but I will not give a long list of names here, although I would like to thank an anonymous reader for the Press. However, a number of people played perhaps a more crucial role through the influence they had on me at a time when my ideas were being formed, and in this connection I ought to mention Martin Davies, Michael Dummett, Kit Fine, David Kaplan and David Wiggins. Most especially, I must acknowledge my long-standing and continuing debt to Christopher Peacocke, who was successively my dissertation supervisor and colleague; I do not much like to dwell upon the thought of what this work would have been like without his influence, encouragement, and help.

G.F.

New Orleans, October 1983

Preface to the Second Edition

Though it did not fall deadborn from the press, *The Metaphysics of Modality* went out of print some years ago. But I have received enough queries to make me think that more copies should be available than the ones already on personal or institutional library shelves. I have therefore prepared this new edition with the intention that it be distributed electronically and free of charge.

Though a few passages have been rewritten to make their meaning clearer and I have corrected as many errors in the first

edition as I know about, the main text of this new version of *The Metaphysics of Modality* is close to the original. Of course, my views have changed about certain issues, and in many areas much interesting work has appeared since 1983. But rather than rewrite the main text of the book to accommodate these developments, I have confined new material to postscripts following various chapters.

There are certain other changes from the first edition. I have abandoned consistent use of quotation marks for mentioning formal expressions, and tried instead to avoid quote-marks clutter. Spelling and punctuation conventions were previously standard British ones, but now they have become a mishmash of British and American – readers of both nationalities will all find much to horrify them. And there are other improvements, too numerous to mention.

G.F.
New Orleans, January 1998

Addendum(Boulder, 2016): there is a Chinese proverb about a man who decided to make a voyage to the moon. He planned to begin by climbing a tall tree, then proceed from there. According to the proverb, the initial stages of the journey went well. And so it proved with the postscripts I started to write in 1998. After writing two, for Chapters 3 and 4, I was unable to find the time to do any more. The Postscript for Chapter 3, mainly on counterpart theory, is now seriously outdated, so for this web edition of the book I decided just to drop both postscripts. The text uploaded here is therefore essentially the same as that of the physical book. – GF

Acknowledgements for the First Edition

Certain passages of this book have been drawn from my previously published papers. For their co-operation in this respect, I wish to thank the publishers of the following journals: *Philosophical Studies*, for material from "Origin and Identity", 37 (1980) 353–62, copyright © 1980 by D. Reidel Publishing Company, Dordrecht, Holland; *The Journal of Philosophical Logic*, for material from "On The Philosophical Basis of Essentialist Theories", 10 (1981) 73–99 copyright © 1981 by D. Reidel Publishing Company, Dordrecht, Holland, and "Physicalism, Instrumentalism and the Semantics of Modal Logic" 12 (1983), copyright © 1983 by D. Reidel Publishing Company, Dordrecht, Holland; *Synthese*, for material from "Thisness and Vagueness", 54 (1983) 235–59, copyright © 1983 by D.

Reidel Publishing Company, Dordrecht Holland; and *Mind*, for material from "Wiggins on Sets and Essence", 92 (1983) 114–19.

The later stages of the production of the first complete draft of this book were assisted by a grant-in-aid awarded to me by the American Council of Learned Societies, for which I would like to express my sincere appreciation.

Colophon

To prepare the web edition, the first edition of *The Metaphysics of Modality* was scanned into a Power Computing Corporation PowerTower Pro using Omnipage Pro running under Mac OS 7.x and later OS 8. Editing and page layout for this edition was done in Adobe FrameMaker, and pdf files were prepared in Adobe Acrobat Distiller. Pages were typeset by the author in 12 on 16 Monotype Plantin, with logic symbols drawn from Richard Jones's Zed font. The page format itself follows a suggestion of Beatrice Beaujoin in her article "Designing for Letter Paper" in *Serif 4*. I thank Charles Poynton for advice on the best way of implementing her suggestion in FrameMaker.

Copyright conditions

Table of Contents

Chapter 1
Propositional Modal Logic

The sentential operators 'possibly' and 'necessarily'

In many respects, things could be different from the way they actually are. We often have 'if only' thoughts – if only Jones had taken his broker's advice, he would be a millionaire today – thoughts which would lose some poignancy if the way things are, including Jones's impecunious state, is the only way it is possible for them to be. So consider the assertion

(1) Jones could be a millionaire today.

Intuitively, we can say much the same thing with

(2) It could be that Jones is a millionaire today

which, although more cumbersome, suggests that in very simple cases like (1), we can regard a sentence containing a verb modified by 'could' or 'could have' as a contraction of a sentence with a non-subjunctive verb which itself occurs in a subsentence governed by a *modal operator*. In (2), the modal operator is the phrase 'it could be that', governing the subsentence 'Jones is a millionaire today' that contains the non-subjunctive 'is'. We symbolize the modal operator as '◇' and often read it as 'it is possible that' or 'possibly'.

However, there is ambiguity here of which one should beware. On one perfectly natural way of hearing

(3) It is possible that Jones is a millionaire today

an utterer of (3) is saying that *nothing he knows* is *inconsistent with* Jones's being a millionaire today. This is the *epistemic* sense of 'it is possible that', in which it means something like 'for all that is known'. In this sense, typical readers of this book cannot truly say

that it is possible that they are millionaires today, since they know very well that today, like other days, they are overdrawn at the bank (see DeRose [1991] for a good discussion). But we mention the epistemic sense of 'possibly' only to distinguish it from the sense of 'possibly' with which we shall be concerned. This second sense involves the kind of possibility relevant to 'if only' thoughts, and is sometimes called the *broadly logical* sense of 'possibly' (as in Plantinga [1974]). As a rough elucidatory guide, 'it is possible that *P*' in the broadly logical sense means that there are ways things might have gone, no matter how improbable they may be, as a result of which it would have come about that *P*. So in this sense it is true, for typical readers, that it is possible they are millionaires today, just as Jones would have been if he had taken his broker's advice.

Not everyone agrees that things could be different from the way they actually are. A *fatalist*, for example, holds that the way things actually are is the only way it is possible for them to be; so a kind of necessity is imputed to things being as they actually are. In fact, fatalism is usually held with respect to the future: if it is going to rain tomorrow, then it must come to pass that it rains tomorrow.[1] We shall return briefly to fatalism in the next section of this chapter. For the moment, let us just note how we can use either the notion of possibility or the notion of necessity to express the fatalist doctrine. We can say that how things are is how they *must* be, or that it is *not possible* for them *not* to be that way. Thus, a necessary state of affairs is one whose failing to obtain is impossible, which means we can define necessity in terms of possibility:

(4) 'It is necessary that *P*' means that it is not possible that not-*P*

or in symbols, using '□' for 'it is necessary that',

(5) $\Box P =_{df} \sim\Diamond\sim P$.

As the reader may suspect, we can do the same thing the other way round, since a state of affairs is possible if and only if ('iff') it is not necessary that it does not obtain. Again in symbols,

1. The philosophical *locus classicus* is Chapter 9 of Aristotle's *De Interpretatione;* see, for instance, Aristotle [1928] and the papers and bibliography in Moravcsik [1968]. A well-known modern discussion is Chapter 6 of Taylor [1992].

(6) $\Diamond P =_{df} \sim\Box\sim P$.

In these definitions, we are introducing □ and ◇ as operators on sentences, but in ordinary speech many different kinds of thing are said to be possible or necessary, including facts, states of affairs, and propositions, and no harm will come of exploiting this variety of means of expression.

Invalid arguments; semantics for S5

□ and ◇ are syntactically just like negation; where, but only where, it is grammatically permissible to have ~, it is permissible to have □ or ◇; thus, for instance, just as $P\sim Q$ is nonsense, so are $P \Box Q$ and $P \Diamond Q$. What *logical* principles do □ and ◇ obey? In classical sentential logic, the logical principles which the operators, or *connectives*, obey, are fixed by truth-tables, which are said to give the meanings of the connectives. They do this by stating how the truth-values of longer sentences are fixed by the truth-values of the shorter sentences from which they are composed by linking with or prefixing the connectives. We can then discover whether or not a particular argument is valid; for example, we accept the principle to infer Q from P and $P \rightarrow Q$ because the truth-table for → tells us there are no ways for P and $P \rightarrow Q$ to be true while Q is false. What we would like, therefore, is something analogous for □ and ◇, so that we can answer such questions as whether it is valid to infer, say, $\Diamond Q$ from $\Box P$ and $\Diamond(P \rightarrow Q)$.

But we cannot give a truth-table for ◇, since this operator lacks a property which must be possessed by any word whose meaning can be encapsulated in a truth-table. Consider the table for negation displayed in the margin. This tells us all we need to know about negation for the purposes of logic: the negation of a sentence takes the opposite value from the sentence negated. Suppose we try something similar for ◇. We can certainly fill in the first row of the table. For if P is true, 'it is possible that P' must also be true: there are ways things could have gone as a result of which P would be true, for P is true, and therefore the way things have actually gone is one. But we cannot fill in the second row of the table, because the mere fact that P is false does not determine whether or not it might have been true. Suppose P is only contingently false; for instance, suppose P is the sentence 'Jones is a millionaire'. If we are not fatalists, we agree that Jones could have been a millionaire, so we put $\Diamond P$ true. But if P is the sentence 'Jones is a married bachelor' then P is not only false, it is necessarily false, and so in this

P	$\sim P$
T	F
F	T

P	$\Diamond P$
T	T
F	?

case we must put $\Diamond P$ false. Incidentally, it would be a mistake to dispute this example on the grounds that 'bachelor' might not have meant 'unmarried man': when we evaluate any sentence, whether or not it contains modal operators, we take the words in the sentence to mean what they do mean, not something else.[2]

The difference which has emerged between $\sim$ and $\Diamond$ is as follows. In order to compute the truth-value of $\sim\varphi$, it suffices to know the truth-value of φ. But to compute the truth-value of $\Diamond\varphi$, if the notion of computation is at all applicable, we may need more than φ's truth-value, for if φ is false, it matters whether it is *necessarily* or only *contingently* false. When a connective forms a longer sentence out of one or more shorter sentences, and to compute the truth-value of the longer sentence it suffices just to be given the value(s) of the shorter sentence(s), the connective is said to be *truth-functional*, and it has a truth-table which gives the truth-value of any longer sentence in terms of the values of the shorter constituent sentence(s). But when information just about the truth-values of shorter sentences does not suffice to determine the values of the longer sentences the connective forms from the shorter ones, the connective is said to be *non-truth-functional*, and there will be at least one row of an attempted truth-table for it where we are in a quandary over what entry to make, as we were at the bottom row of the attempted table for $\Diamond$. Readers should confirm for themselves that $\Box$ suffers from a similar problem, this time on the upper row of the table; see Ch. 3.8 of Forbes [1994] for other examples.[3]

We can make some headway with the task of providing a semantics for $\Box$ and $\Diamond$ by considering sample incorrect inferences and asking why, at the intuitive level, we reject them. Consider inference (A), displayed in the margin. The premises tell us that P is possible and Q is possible, the conclusion asserts that P and Q are compossible (possibly true together). But this does not follow; it is

(A)

$$\frac{\Diamond P \quad \Diamond Q}{\Diamond(P \,\&\, Q)}$$

2. Another way of making this point is to say that, strictly, it is not the sentence 'Jones is a married bachelor' which is necessarily false, but rather the proposition it expresses. To imagine a situation in which the sentence expresses a different proposition is irrelevant to the question of whether or not the proposition it actually expresses is necessarily false.

3. Readers may wonder whether the impossibility of giving a truth-table for $\Box$ or $\Diamond$ is somehow related to the fact that we are working with only two truth-values. That this is not so is shown in Dugundji [1940], where it is established that no finitary truth-tables can be given for the modal operators. However, an interesting partial characterization of the sense of the operators is possible with four values; see Kearns [1981].

possible that it now be raining everywhere and possible that it now be dry everywhere, but it is evidently not possible to have both these states of affairs obtaining together. A very natural way of explaining what is going on here – we shall return to the question of exactly how it is an explanation – involves yet another way of reading the operator ◇, in which a sentence of the form ◇φ is read as 'there are some possible circumstances in which φ'. On this reading, ◇ has become a kind of existential quantifier ranging over objects which we are calling possible circumstances. Now it is well known that inference (B) is incorrect in ordinary first-order logic: if one object satisfies *F* and a *different* object satisfies *G*, then the two premises are true, but there need be no object which satisfies both *F* and *G*. So the assimilation of ◇ to an existential quantifier, one restricted to ranging over possible circumstances, enables us to explain the incorrectness of (A) by analogy with the usual explanation of the incorrectness of (B).

(B)

$$\frac{(\exists x)Fx \quad (\exists x)Gx}{(\exists x)(Fx \,\&\, Gx)}$$

In more detail, let us use *w*, *u*, and *v* as variables ranging over possible circumstances, and for any such variable *x*, let us abbreviate '*P* holds in *x*' by *Px* and '*P* & *Q* holds in *x*' by *Px* & *Qx*.[4] Then with the new reading of ◇ as an existential quantifier, we can rewrite (A) as (C). The simplest formal counterexample to (B) involves a domain of two objects, one satisfying *F* but not *G*, the other *G* but not *F*. The counterexample to (C) and therefore to (A) is just the same, except that we use different terminology to describe it. We have a domain *W* of two possible circumstances *u* and *v*. We let *P* hold at *u* but not at *v* and *Q* hold at *v* but not at *u*; so the premises of (C) come out true while its conclusion is false; and since we are saying that $(\exists w)Pw$ means the same as ◇*P*, $(\exists w)Qw$ means the same as ◇*Q*, and $(\exists w)(Pw \,\&\, Qw)$ means the same as ◇(*P* & *Q*), it follows that the counterexample to (C) is also a counterexample to (A).

(C)

$$\frac{(\exists w)Pw \quad (\exists w)Qw}{(\exists w)(Pw \,\&\, Qw)}$$

If ◇ is to be read as an existential quantifier over possible circumstances, how should □ be read? From the equivalence of ∀ with ~∃~ and the definition of □ as ~◇~, we have little choice but to read □ as a universal quantifier over possible circumstances. Furthermore, this is intuitively correct: what else could be involved in asserting that a proposition is necessary than that it holds in all

4. The reader will find a complete account of the method of translating formulae of modal propositional logic into possible worlds language in the Appendix.

possible circumstances? Now consider inference (D). It is easy to think of an informal counterexample. Let P be 'Jones is a bachelor' and Q be 'Jones is unmarried'. Suppose P is true; $\Box(P \rightarrow Q)$ is also true, of course, but the conclusion $\Box Q$ is false, for it is not necessary that Jones is unmarried – there are many ways things could go or could have gone in which Jones gets married. However, we can give a more formal demonstration of the incorrectness of (D) in possible circumstance terminology. $\Box(P \rightarrow Q)$ becomes 'in every possible circumstance w, if P holds in w then Q holds in w', or in symbols, $(\forall w)(Pw \rightarrow Qw)$. What becomes of P? All other sentences we have considered up to now have a modal operator as main connective; in the possible circumstance translations, the operator becomes a quantifier and the sentential letters become predicates attached to the circumstance variables; if you like, the states of affairs for which P, Q, etc., stand, become properties of the circumstances. When a sentential letter occurs on its own, or in a truth-functional combination with another formula, it is interpreted as making an assertion about the *actual* circumstances, which we denote conventionally by '$w^\star$'. That is, a sentential letter on its own still becomes a predicate, but a predicate of the actual circumstances $w^\star$: we translate P as $Pw^\star$, a simple subject-predicate sentence of first-order logic. (D) translated into possible circumstance terminology becomes (E). This is a straightforwardly invalid sequent of first-order logic: consider a domain of two objects, one of which satisfies both P and Q (this is $w^\star$) and the other neither; then the premises of (E) are true but the conclusion false. The terminologically appropriate way of describing this counterexample is as follows. Choose a set W of two possible circumstances u and v; let each of P and Q be true at u and false at v, and let '$w^\star$' denote u. Then P and $\Box(P \rightarrow Q)$ are both true, since true at $w^\star$, while $\Box Q$ is false at $w^\star$, since Q is false at v. Hence (D) is incorrect. Readers should test their understanding of this argument by establishing in a similar way that ($D^\#$) in the margin is also incorrect.

(D)

$$\frac{P \quad \Box(P \rightarrow Q)}{\Box Q}$$

(E)

$$\frac{Pw^\star \quad (\forall w)(Pw \rightarrow Qw)}{(\forall w)Qw}$$

($D^\#$)

$$\frac{\Diamond P \quad \Diamond(P \rightarrow Q)}{\Box Q}$$

At this point, we will digress a little to return to our formulation of fatalism on page 2. The fatalist was represented as having the thought 'if it is going to rain tomorrow then it must be that it is going to rain tomorrow'. If we put P for 'it is going to rain tomorrow', this thought appears to have the form

(7) $P \rightarrow \Box P$

and the whole fatalist argument can be written as (F), which is obviously valid. But why should we believe (7)? There may be some justice in the suspicion that the superficial attractiveness of fatalism is rooted in a failure to distinguish (7) from the triviality

(8) $\Box(P \rightarrow P)$

since by failing to distinguish (7) and (8), the incontestable correctness of (8) carries over in the mind to lend (7) a degree of plausibility it does not deserve. But if the fatalist is confusing (7) and (8) then there is no argument for the conclusion $\Box P$, since instead of (F), the fatalist argument becomes (G), which we know is incorrect, since it is just an instance of (E). (This example shows that modal logic has the same kind of clarificatory power as standard logic in the exposing of fallacies.)

(F)

$$\frac{\begin{array}{c} Pw^{\star} \\ Pw^{\star} \rightarrow (\forall w)Pw \end{array}}{(\forall w)Pw}$$

(G)

$$\frac{\begin{array}{c} Pw^{\star} \\ (\forall w)(Pw \rightarrow Pw) \end{array}}{(\forall w)Pw}$$

In establishing the incorrectness of these arguments, our procedure has been to translate them into first-order logic, then to show the invalidity of the translated argument using standard semantics for that logic, and finally to redescribe the resulting counterexample in possible circumstance terminology. It ought to be clear that the two intermediate steps in this process are unnecessary: we should be able to go straight from the modal logical argument to the structure with possible circumstances which shows it to be incorrect. We can describe how to do this by analogy with first-order logic. There, an *interpretation* or *model* is a domain D of objects together with a stipulation of which predicates apply to which objects in D, and of which names stand for which objects in D. Formulae are evaluated in models by simple rules which tell us, for example, that the conjunction of two predicates applies to an object iff both conjuncts do, or that an existential quantification of a predicate with one free variable, for instance, $(\exists x)Fx$, is true in a model iff the quantified predicate F applies to, or is satisfied by, at least one object in the domain D of the model. More or less the same ideas, although clothed in different terminology, give us a model theory for sentential modal logic.

In stating this model theory, we will change our earlier practice in one respect: instead of speaking of possible circumstances, we will speak of possible *worlds*. A possible world is a *complete* way things might have been – a total alternative history. The possible circumstance in which Jones takes his broker's advice and makes a million is really just a component of a world, indeed a component

of many worlds, since there are many distinct total histories alternative to that of the actual world of which this circumstance is a part. In terms of our model theory, the requirement that worlds be complete is reflected in the constraint that every sentence-letter occurring in the argument in question be assigned one or other truth-value at each world. We shall see in the last section of this chapter that we can get by without this sort of completeness, but that we pay a price in terms of simplicity.

We now give the following definition of what we shall rather mysteriously call an 'S5' model.

An S5 model **M** for a set of sentences Σ of sentential modal logic is

(i) a non-empty set W of possible worlds;
(ii) for each world w in W and each sentence-letter π which occurs in any sentence in Σ, a specification of which truth-value π takes at w;
(iii) a selection of a particular world w in W to play the role of the actual world; we denote this world by '$w^\star$'.

In the terms of our analogy with first-order structures, W is like the domain D of objects in such a structure, and the specification (ii) of letter-values at worlds is like the assignment of an extension from D to a one-place predicate letter. So the sentence-letters are treated as if they expressed properties of worlds. (iii) has no first-order parallel, for '$w^\star$' is not a name in the language of the sentences in Σ, while, in first-order logic, only names from the sentences in Σ are assigned referents.

We will say that an argument of sentential modal logic is *valid in S5* iff there is no S5 model in which its premises are all true and its conclusion false. This is just the standard notion of validity common to all systems of logic in which sentences are true or false. However, in order to apply it, we evidently have to know how to evaluate sentences of sentential modal logic in S5 models. So next we must spell out the evaluation rules which tell us how to do this. In general, a sentence σ is true in a model **M** iff σ is true at the actual world $w^\star$ of **M**. 'true at $w^\star$' is a special case of the more general notion of being true at a world, so our rules tell us when a sentence of a particular form is true at an arbitrary world w. The rules are exhaustive because they cover all the forms it is possible for a sentence in sentential modal logic to have.

(iv) a sentence-letter π (e.g., P, Q, etc.) is true at a world w in a model **M** iff the specification for **M** described in clause (ii) above says that π has the value true at w;

(v) a negation $\sim\varphi$ is true at a world w (in a model **M** – henceforth we omit this) iff φ is not true (i.e., φ is false) at w;

(vi) a disjunction $\varphi \vee \psi$ is true at a world w iff φ is true at w or ψ is true at w;

(vii) a conjunction φ & ψ is true at a world w iff φ is true at w and ψ is true at w;

(viii) a conditional $\varphi \rightarrow \psi$ is true at w iff either φ is false at w or ψ is true at w;

(ix) a possibilitate $\Diamond\varphi$ is true at w iff there is some world u in W such that φ is true at u;

(x) a necessitate $\Box\varphi$ is true at w iff for all worlds u in W, φ is true at u.

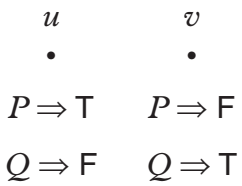
Instead of rules for translating modal formulae into first order formulae, these are rules for evaluating modal formulae directly in S5 models of the kind defined. Any formula may now be evaluated in a model, applying these rules one after another as the structure of the formula dictates. Note that in (ix) and (x), the quantifier readings of the modal operators are preserved, in virtue of the occurrence of 'there is' in (ix) and 'for all' in (x).

Now let us apply this apparatus to (A) and (D) for the purposes of illustration. To show that (A) is invalid, we need an S5 model for the set of sentences $\Sigma = \{\Diamond P, \Diamond Q, \Diamond(P \,\&\, Q)\}$ in which the first two members of this set are both true at $w^\star$ but $\Diamond(P \,\&\, Q)$ is false there. So let $W = \{u, v\}$. The sentence-letters which occur in the sentences in Σ are P and Q, so we have to say what truth-values these have at u and at v. We can do this by defining a function f for our model which assigns to each sentence-letter exactly the set of worlds of W at which that letter is true. So we should say: $f(P) = \{u\}, f(Q) = \{v\}$.

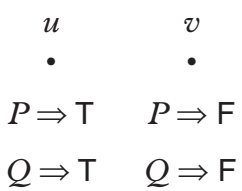
[Sometimes the set of worlds at which a sentence-letter is true is called the "proposition" assigned to that sentence-letter; according to this account, propositions are identified with sets of worlds, and the proposition expressed by any sentence, simple or complex, is the set of worlds at which that sentence is true (as determined by such rules as (iv)–(x) above). However, this is a rather special use of 'proposition' – for instance, it implies that all necessarily false sentences express the same proposition, the empty set of worlds –

and we shall not employ it. See Stalnaker [1984] for an extended discussion and defense.]

As an alternative to defining the function f, we can just draw a picture which represents the same information, and which is perhaps easier to work with. We write '$P \Rightarrow \mathsf{T}$' for '$f(P) = \mathsf{T}$', etc. Let us say that the actual world $w^\star$ is u; then both $\Diamond P$ and $\Diamond Q$ are true at $w^\star$, according to clause (ix), but according to (ix) and (vi), $\Diamond(P \,\&\, Q)$ is false at $w^\star$. Hence (A)'s premises are true but its conclusion false in this model.



For (D), Σ is the set of sentences $\{P, \Box(P \rightarrow Q), \Box Q\}$, and since we want P to be true in the model, it has to be true at $w^\star$. So let $W = \{u, v\}$ and put $f(P) = \{u\}$ and $f(Q) = \{u\}$, as displayed. $w^\star$ must be u if P is to be true at $w^\star$, and since $P \rightarrow Q$ is true at both u and v, $\Box(P \rightarrow Q)$ is true at $w^\star$ as well. But by (x), $\Box Q$ is false at $w^\star$, and so we have a counterexample to (D).



Finally, what about a valid argument in S5? (H) is an example. If (H)'s premises are true in a model, then P and $P \rightarrow Q$ are true at each world in the model, and so by *modus ponens* Q must be true at each world too. The system S5 is the totality of all arguments like (H) to which there is no S5 counterexample.

(H)

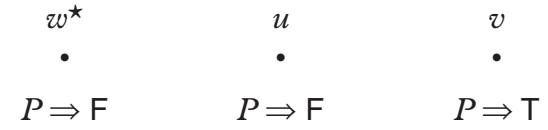
$$\frac{\Box P \quad \Box(P \rightarrow Q)}{\Box Q}$$

Other systems

Readers will have surmised from the use of the label 'S5' that there are other systems of sentential modal logic. These arise out of divergent treatments of arguments involving sentences with iterated modalities, that is, blocks of modal operators. (I) is an example. It is not unnatural to hear 'it is possible that it is possible that P' as making a weaker assertion than 'it is possible that P'. In support of this, one might say that there are ways things could have been such that if things had been that way, it would have been possible that P, but as things actually are, it is not possible that P. In this case, (I) is invalid. However, (I) is valid in S5. If $\Diamond\Diamond P$ is true at $w^\star$ in some model, then $\Diamond P$ is true at some world u in that model (put $\Diamond P$ for φ in clause (ix) on page 9) and therefore P is true at some world v in the model, as in the following three-world model:

(I)

$$\frac{\Diamond\Diamond P}{\Diamond P}$$



But then from clause (ix) it follows that $\Diamond P$ is true at $w^\star$; indeed, by repeated applications of clause (ix), we can see that if $\Diamond\ldots\Diamond P$ is

true at any world in a model, for any number of ◇'s, then ◇*P* is true at every world in the model. In terms of our intuitive argument for the invalidity of (I), the problem with S5 is that there is no way of expressing the thought that the truth of *P* at *v* does not guarantee the truth of ◇*P* at *w*★ if *v* is reached 'through' an intermediate world *u*, the idea that although *P* would be possible were things as in *u*, this does not mean that *P* is possible as things actually are (as they are in *w*★). For a counterexample to (I), we need some way of representing the idea that while what obtains at *v* is possible at *u*, it is not possible at *w*★.

To represent this idea we can put "barriers" between worlds, so to speak: if two worlds *w* and *w′* are separated by a one-way barrier, then what obtains in *w* need not be possible at *w′*, or conversely, depending on the direction through which the barrier cannot be passed. Of course, a two-way barrier is also possible. When a world *w′* is barred from a world *w*, so that what obtains in *w′* need not be possible at *w*, we say that *w′* is not *accessible* from *w*.[5] However, in the three-world model above, there is only one world, *v*, at which *P* is true, and so we can prevent ◇*P* from being true at *w*★ simply by stipulating that *v* is not accessible from *w*★. If we also say that *v* is accessible from *u*, then ◇*P* will be true at *u*; therefore, if we say that *u* is accessible from *w*★, so that what obtains at *u* is possible at *w*★, then ◇◇*P* is true at *w*★. So this gives us our counterexample to the inference (I).

We have to change some of the definitions on pages 8 and 9 to accommodate this new notion of accessibility. Suppose we symbolize '*y* is accessible from *x*', as 'Acc(*x*, *y*)', which we may also read as '*x* has access to *y*', or more colloquially, '*x* can see *y*'. Then the stipulations needed to obtain our counterexample to (I) are as follows: (a) Acc(*w*★, *u*); (b) Acc(*u*, *v*); (c) Not-Acc(*w*★, *v*). It is immediately obvious from (a)–(c) that what we have done to obtain the counterexample is to allow the accessibility relation to be non-transitive. That is, we have not insisted that:

(9) $(\forall x)(\forall y)(\forall z)(\text{Acc}(x, y) \,\&\, \text{Acc}(y, z) \rightarrow \text{Acc}(x, z))$.

So the example suggests that which arguments are counted as valid or invalid will turn on which structural conditions (transitivity,

5. Note that what obtains in *w′* may be possible at *w* even if *w′* is not accessible from *w*, since there may be other worlds which are accessible from *w* in which the same things obtain as obtain in *w′*.

symmetry, etc.) we impose or refuse to impose on the accessibility relation. Let us now reformulate some earlier definitions with all this in mind. We want to make room for the accessibility relation in our definition of model and we want to alter (ix) and (x) on page 9 so that the truth-value of a modal sentence at a world w depends only on what obtains at the worlds accessible from w.

A *general* model for a set of sentences Σ in sentential modal logic is:

(i) a non-empty set W of possible worlds;
(ii) for each sentence-letter π which occurs in any sentence in Σ, and for each world w in W, a specification of the truth-value of π at w;
(iii) for each world w in W, a specification of which worlds in W are accessible from w;
(iv) a selection of a particular world w from W as $w^\star$.

Here clause (iii) is the novelty; note that (iii) does not impose any structural constraints at all on the accessibility relation; (iii) even admits a model in which no world is accessible from any world, not even itself; to rule out such a model we would have to stipulate that accessibility is reflexive, i.e., that each world can see itself.

All the rules on page 9 for evaluating formulae at worlds in S5 models carry over to general models, except the rules for the modal operators. In place of (ix) and (x), we have the following:

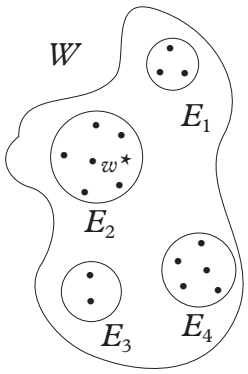
($\Diamond$) a possibilitate $\Diamond\varphi$ is true at a world w in a general model **M** iff there is some world w' in the set W of worlds of **M** such that φ is true at w' and w' is accessible from w according to the specification of accessibility described in (iii) above for **M** (for short: $\Diamond\varphi$ is true at w iff φ is true at some world w can see);
($\Box$) a necessitate $\Box\varphi$ is true at a world w iff for all worlds w' in W, if w' is accessible from w then φ is true at w' (for short: $\Box\varphi$ is true at w iff φ is true at every world w can see).

Different systems of modal logic arise as we impose different structural constraints upon the accessibility relation; moreover, relationships between these constraints will be reflected in relationships among the systems they generate. Let us start with the constraints on accessibility which define the system S5. That is, we want to define a subclass of general models which validates exactly

the arguments validated by our earlier account of an S5 model. Truth in a model, as always, is truth at the actual world of the model. In our original S5 models we were in effect allowing that any world can see any other world, and also itself, because we did not even mention accessibility in the definition of a model or in the clauses for the modal operators. That is, the accessibility relation is *universal*, hence invisible, in those models: $(\forall u)(\forall v)\mathrm{Acc}(u,v)$. Thus, if we take the subclass of general models in which the accessibility relation is universal, we obtain exactly the S5 models.

However, something a little more general is possible. A universal relation is a special case of an *equivalence* relation, a relation that is reflexive, symmetric and transitive. An equivalence relation is so-called because it partitions any domain on which it is defined into *mutually exclusive* and *jointly exhaustive equivalence classes*. In English, most equivalence relations are expressed by phrases of the form 'is the same…as' or 'has the same…as'. For example, 'has the same date of birth as' partitions the domain of people into classes of people who are equivalent with respect to date of birth, i.e., who have the same date of birth. The classes are mutually exclusive because no-one has two dates of birth, and jointly exhaustive because everyone has a date of birth.[6]

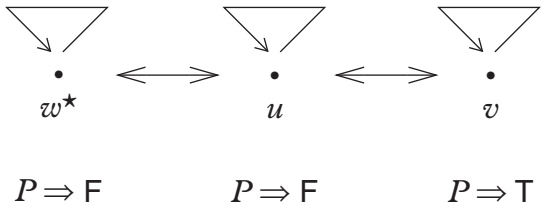
In a general model in which accessibility is an equivalence relation, the set of worlds W may be partitioned into a number of different equivalence classes $E_1, E_2,\ldots$ Such a general model is not identical with any S5 model of our original sort, since in these there is only one equivalence class of worlds, the whole set W. However, one may think of one of our original S5 models as sitting 'inside' various general models in which accessibility is an equivalence relation (E_2 in the figure) and from which the original S5 model could be obtained by first deleting all the equivalence classes of worlds other than the class of $w^\star$, and then dropping all mention of accessibility. The same sentences hold in all models



6. It is a useful exercise for readers to check that if (i) R is an equivalence relation (reflexive, symmetric and transitive), (ii) D is a non-empty domain of discourse, and (iii) for every $x \in D$ the class $E_x = \{y \in D\colon Rxy\}$, then the sets E_x, $x \in D$, constitute a mutually exclusive and jointly exhaustive partitioning of D. (For exhaustiveness, show that $\forall y \in D\ \exists x \in D\colon y \in E_x$; for mutual exclusiveness, show that if $E_x \neq E_y$, then $E_x \cap E_y = \varnothing$.) Conversely, show that if the sets E_i, $i \in I$, constitute a mutually exclusive and jointly exhaustive partitioning of a non-empty domain D, then the relation R on D defined by Rxy iff $\exists k \in I\colon x \in E_k$ & $y \in E_k$, is an equivalence relation. (I is an arbitrary "index" set.)

thus related, since no matter how many modal operators there are in φ, evaluation of φ at $w^\star$ will not introduce into consideration any world outside the equivalence class of $w^\star$. So it follows that an argument has an S5 counterexample of the original sort if and only if it has a counterexample involving a general model in which accessibility is an equivalence relation. For if the counterexample is of the original sort, it remains a counterexample if we turn it into a general model in which accessibility is universal. And if the counterexample is a general model, it remains a counterexample after deletion of all the equivalence classes of worlds other than the class of $w^\star$ and removal of the accessibility relation. So the system S5 is the system whose models are exactly those general models in which accessibility is reflexive, symmetric and transitive.

The three structural properties of accessibility characteristic of S5 suggest a natural way of obtaining other systems, by ringing changes on the three properties. We have already seen that to obtain a counterexample to (I) we have to permit accessibility to be non-transitive. So if we make the three stipulations (a), (b), and (c) on page 11 about the counterexample to (I) we pictured, and then add the five further stipulations {Acc($w^\star$,$w^\star$), Acc(u,u), Acc(v,v), Acc(u,$w^\star$) Acc(v,u)}, we obtain a general model in which accessibility is non-transitive, symmetric and reflexive:



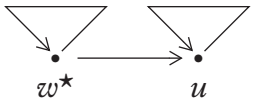
In such diagrams as these, an arrow running from x to y represents the accessibility of y from x, so in this particular diagram, v is not accessible from $w^\star$, since there is no arrow running from $w^\star$ to v. The class of all models in which the accessibility relation is reflexive and symmetric, but not necessarily transitive (it will be transitive in some models but not in others), determines a set of valid inferences, usually known as the system B, because of a tenuous connection with the intuitionistic logic of Brouwer (for detailed historical information about modal logic see Hughes and Cresswell [1968] and Lemmon and Scott [1977]). So the general models which are B models consist in all S5 models together with those

other general models in which accessibility is non-transitive (though still reflexive and symmetric). Thus any S5 counterexample to an inference will also be a B counterexample, since S5 models are B models, but some B counterexamples to inferences will not be S5 counterexamples, since no B model which refutes the inference is an S5 model (these are the inferences, such as (I), refuted only by B models with non-transitive accessibility relations). So all B validities are valid in S5, while some S5 validities are invalid in B, for instance, (I). It follows that B is a *subsystem* of S5, in the precise sense that its validities are a subset of those of S5.

Another subclass of general models is the class of all general models which have reflexive and transitive accessibility relations (in some of them accessibility is symmetric and in others not). Symmetry is the crucial property for the inference (J). For $\Box\Diamond P$ to hold at $w^\star$, every world which $w^\star$ can see must itself be able to see a world where P is true. If accessibility is not symmetric, then this requirement can fail even if P holds at $w^\star$. The simplest general model meeting these requirements is pictured in the margin. Although P is true at $w^\star$, $\Box\Diamond P$ is false there, since there is a world $w^\star$ can see, namely, u, at which $\Diamond P$ is false. $\Diamond P$ is false at u because P does not hold at any world which u can see – in particular, u cannot see $w^\star$, because although there is an arrow going from $w^\star$ to u there is none in the other direction. Hence symmetry fails.

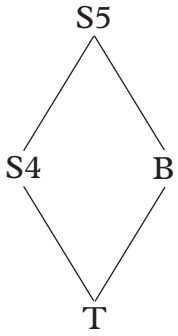
(J)

$$\frac{P}{\Box\Diamond P}$$



The system whose models are all and only the general models in which accessibility is reflexive and transitive is called S4, and so the model we have just described is an S4 model, since, as the arrows are drawn, it is only symmetry which fails. Thus (J) is not valid in S4. Neither S4 nor B is a subsystem of the other, since the S4 model just described is not a B model, and the B model described on page 14 is not an S4 model and, in fact, (J) is valid in B and (I) is valid in S4. However, the S5 models are a subclass of S4 models, just as they were of B models, and so by the same reasoning as showed that B was a subsystem of S5, we can infer that S4 is also a subsystem of S5.

A fourth system results if we relax the restrictions on accessibility even further, and require only that it be reflexive. The system whose models are exactly the general models in which accessibility is reflexive is called T, and it should be clear that T is a subsystem of B, S4, and S5, since every model for any of these three systems is also a T model, but some T models, those in which accessibility



is non-transitive and non-symmetric, are not models for any of these other systems. So we can picture the inclusion relations among the systems in the figure in the margin, with S5 at the top, since it has the most inclusive set of valid arguments. Inverting the diagram yields a representation of the inclusion relations among the classes of models for the systems.

A fifth system which we can generate by these methods is the system whose models are *all* the general models, the system defined by the complete absence of restrictions on accessibility. But this system, known as K, does not appear to capture any natural notion of possibility or necessity. For instance, it would be incorrect in this system to infer $\Diamond P$ from P: consider a model with one world, $w^\star$, at which P is true, in which $w^\star$ cannot see itself (accessibility is non-reflexive). Then $\Diamond P$ is false at $w^\star$, so we have a counterexample to "whatever is the case is possible"! We can also give a counterexample to the inference of P from $\Box P$. Such a system is of little interest in the present context.[7]

Clearly, we have not exhausted the different possible stipulations which we can impose on accessibility: we could stipulate that any world be seen by only a finite number of worlds, that of any two worlds at least one see the other, and so on. Such stipulations define different subclasses of general models and so determine different systems; it is hardly possible to give a complete account here (see Chellas [1980]). But with such a plurality of systems, it is natural to ask which is the "right" one. For our purposes, the right one is the one that captures our concepts of broadly logical possibility and necessity. We have said enough about these concepts to see that they are not captured by a system which permits models with non-reflexive accessibility, but we have not said enough to make further discriminations. However, it will turn out that S5 is an appropriate framework for the philosophical issues we will explore, which motivates restricting further development of modal semantics to S5 alone.[8] We may therefore drop mention of accessibility and revert to the definition of S5 model on page 8.

7. But systems with non-reflexive accessibility are of considerable interest in a context in which $\Box$ stands for provability in formal arithmetic, rather than broadly logical necessity. See Boolos [1979].

8. So I am not endorsing the motivation for introducing accessibility that I suggested, namely, that the premise of (I) can be heard as making a weaker assertion than its conclusion. The question of which modal logic is the right one is pursued in Salmon [1989].

Incomplete circumstances: possibility semantics

In motivating the quantifier treatment of □ and ◇, we used the phrase 'possible circumstance', but when it came to defining the notion of an S5 model, we spoke instead of possible worlds, with the explanation that worlds are complete circumstances. Completeness amounts to the condition that in each world in a model every sentence-letter in the language of the relevant argument is assigned a truth-value, the language of the argument being the set of all sentence-letters that occur in any sentence in the argument. It is not evident in the resulting semantics what role this completeness condition plays, and the simplest way of highlighting the role is to see what we have to do to get by without completeness. So our question is: if we allow worlds to be incomplete, how can we obtain the system S5?[9]

Let us say that each sentence of the form ◇φ, where φ contains no modal operators, introduces a *possibility:* obviously enough, the possibility that φ. One can think of the possibility that φ as a possibility which can be realized in a number of different ways, corresponding to the different assignments of truth-values to the sentence-letters in φ on which φ is true. By analogy with worlds, then, one can think of a possibility as an object at which certain sentence-letters are assigned truth-values, and on some of these assignments, one can say that φ is true at that possibility. Possibilities will be incomplete if they do not assign truth-values to every sentence-letter in the language, and φ could be true at a possibility even although there are some sentence-letters in the language which are not assigned any truth-value at that possibility. The idea, then, is that possibilities are a broader class of entities than possible worlds: they include worlds, which are complete possibilities, but also include genuinely incomplete entities.

According to this rough account, completeness is a relative notion: an entity is complete with respect to a sentential language *L* iff each sentence-letter in *L* is assigned one or other truth-value at that entity. So in the discussion to follow, the relativity to language will be explicit. We now want to define the idea of a possibility model and give evaluation clauses for the connectives in such a

9. What follows is based on Humberstone [1981] and is of a more advanced nature than the preceding material. Readers may prefer to proceed directly to Chapter 2, and read through to the end of Chapter 3, omitting the last part of Chapter 2, which extends Humberstone's semantics to first-order logic. It is not until the end of Chapter 4 that the results of these sections are appealed to.

way that possibility models validate exactly the arguments validated by S5 models (a more general treatment which allows for an accessibility relation and yields something analogous to the notion of a general model is also possible; see Humberstone [1981]).

The initial steps are straightforward. Part of the definition should say that a possibility model for L ('p-model for L') is

(i) a set Ω of possibilities;
(ii) for each possibility ρ in Ω, and for some subset of the sentence-letters of L, a specification of which truth-value each sentence-letter in the subset has at ρ;
(iii) a selection of a particular possibility ρ as the actual possibility $\rho^\star$; since the possibility which is actual is, at least arguably, complete, $\rho^\star$ should be a possibility at which every sentence-letter in L is assigned a truth-value.

However, we need a little more apparatus if we want to end up with an S5 semantics. Consider two possibilities ρ and τ in a p-model for L, and suppose that the sentence-letters which are assigned a truth-value at ρ are a subset of those assigned a truth-value at τ. Then if every sentence-letter assigned a truth-value at ρ retains that value at τ, we say that τ is a *refinement* of ρ, or that τ *refines* ρ. We will symbolize this relationship with '$\gg$', so if τ refines ρ we write '$\tau \gg \rho$'. Let us also say that the sentence-letters which are assigned some value at a possibility ρ are the sentence-letters for which ρ is *defined*. Then '$\tau \gg \rho$' means that τ assigns the same truth-values as ρ does to the sentence-letters for which ρ is defined. We can now state the extra clause we need in the definition of p-model:

(iv) *Refinability:* for any sentence-letter π in L and any possibility ρ in Ω, if π is undefined (neither true nor false) at ρ then there are possibilities τ and ζ in Ω such that $\tau \gg \rho$ and $\zeta \gg \rho$ and π is true at τ and π is false at ζ.

The technical justification for clause (iv) is that without it we do not even obtain classical propositional logic, much less S5. Intuitively, the clause says that every possibility may be further refined, or extended, in each of two ways, for any atomic proposition undefined at that possibility. We might say that this is guaranteed by the mutual logical independence of the atomic propositions, and if it is possible that a possibility be refined in such-and-such a way, then there is a possibility which is the refinement of the original

one in that way. So Refinability is a rather natural closure condition on the domain of a possibility model.

The next step is to give evaluation clauses for each of the connectives, and it is here that there is a considerable loss of simplicity by comparison with possible worlds semantics. Two clauses are rather obvious, corresponding to clauses (iv) and (v) on page 9.

(v) a sentence-letter π is true at a possibility ρ in a *p*-model **M** iff π is true at ρ according to the truth-value specification described in clause (ii) (page 18);

(vi) a conjunction $\varphi \,\&\, \psi$ is true at a possibility ρ iff φ is true at ρ and ψ is true at ρ.

All we need to do now is to give a clause for negation, and then we can treat all the other connectives as being introduced by the usual definitions in terms of $\sim$ and &. But under what condition should a negated sentence $\sim\varphi$ be said to be true at a possibility ρ? Suppose that φ is the sentence-letter Q. If Q is false at ρ then $\sim Q$ should certainly be true there, and vice versa. But if Q is undefined at ρ, then it verges on inconsistency with Refinability that $\sim Q$ should be true. According to Refinability, some possibility τ refines ρ and Q is true at τ; if $\sim Q$ is true at ρ, it follows that although the refinement relation preserves the truth-values of atomic sentences, it does not preserve the values of complex sentences: otherwise, Q and $\sim Q$ would both be true at τ. But this clashes with our conception of refinement as elaboration or expansion of a possibility rather than revision of it. For a non-atomic sentence φ, however, the problem will not arise if φ is undefined at ρ and there is no refinement of ρ at which φ is true. This prompts the thought that for *any* sentence φ, the condition under which $\sim\varphi$ should be true at ρ is just if φ is not true at any refinement of ρ (by the definition of Refinement, each possibility refines itself). The clause is:

(vii) $\sim\varphi$ is true at ρ iff $\nexists\tau$ such that $\tau \gg \rho$ and φ is true at τ.

It is now possible to introduce the other propositional connectives by the standard definitions, and it is worth investigating what evaluation conditions these definitions bestow on the connectives. Using (vi) and (vii) to unpack $\sim(\sim\varphi \,\&\, \sim\psi)$ we find that

(viii) $\varphi \vee \psi$ is true at ρ iff $\forall\tau$: if τ refines ρ then $\exists\zeta$ such that ζ refines τ and φ is true at ζ or ψ is true at ζ.

The inconvenience of working with incomplete worlds is apparent from this clause. However, the clauses for the modal operators are simple. For $\Diamond$, we just say

(ix) $\Diamond\varphi$ is true at ρ iff there is some τ such that φ is true at τ.

The connective $\Box$ can now be introduced by the definition $\sim\Diamond\sim$, and, again by unpacking, it can be established that the evaluation condition for $\Box\varphi$ thus bestowed is the standard one, requiring the truth of φ at every possibility.

Next, we have to show that the semantics just outlined validates the same arguments as the S5 possible worlds semantics on pages 8–9. For the details, readers must refer to Humberstone [1981], but the main idea is not difficult. To show that the two semantics validate the same arguments, it suffices to show that for any set Σ of *L*-sentences, Σ has an S5 possible worlds model iff Σ has a possibility model. One half of this biconditional is immediate, for if Σ has an S5 possible worlds model, it has a possibility model *ipso facto*, one in which the possibilities are all complete and the only refinement any possibility has is itself (it was to permit this move that we said that possibilities may be, but do not have to be, incomplete, but the same development can be worked out even if complete possibilities are disallowed). For the other half of the biconditional, it is necessary to take some set of axioms and/or rules of inference from which exactly the valid S5 arguments can be proved, and show that each axiom is true in any possibility model and that each rule of inference preserves truth in any possibility model. Then if Σ has no S5 possible-worlds model, so that Σ is S5-inconsistent and we can infer an explicit contradiction from the sentences in Σ using the axioms and rules for S5, the soundness of these axioms and rules with respect to possibility models means that Σ has no possibility model. Hence, by contraposition, if Σ has a possibility model, it must have an S5 model. Thus the original S5 semantics and the possibility semantics validate the same arguments.

Despite this equivalence, we shall not make much use of possibility semantics: the clause for negation makes the system too complicated to work with, especially when we move to first-order modal logic. However, for certain philosophical purposes, it will be useful to know that it is *in principle* possible to employ incomplete circumstances; we will return to this topic in Chapter 5.

Chapter 2
First-Order Modal Logic

Operators, quantifiers, and invalid inferences

JUST as sentential modal logic was obtained by adding □ and ◊ to ordinary sentential logic and allowing these operators to function syntactically like ~, so first-order modal logic is obtained by adding □ and ◊ to ordinary first-order logic with identity, where they are again governed by the same syntactic formation rules as govern ~. So we can have □*Fa* and ◊*Fa*, but not *F*□*a* and *F*◊*a*, since we cannot have *F*~*a*.

Intuitions about the nature of a semantics for first-order modal logic are best engaged by consideration of how modal operators interact with the characteristic connectives of first-order logic, the existential and universal quantifiers. I presume readers are familiar with the distinction between (∃*x*)~*Fx* and ~(∃*x*)*Fx*, and between (∀*x*)~*Fx* and ~(∀*x*)*Fx*. We begin, therefore, by investigating the question of whether there are corresponding differences between such pairs of formulae as (∃*x*)◊*Fx* ('something is possibly F') and ◊(∃*x*)*Fx* ('possibly something is *F*'), and (∀*x*)□*Fx* ('everything is necessarily F') and □(∀*x*)*Fx* ('necessarily, everything is *F*').

An atheist is someone who holds that there is no such being as God, but he may think of this as either a contingent or a necessary truth. Consider the atheist who believes that God's non-existence is contingent, perhaps because he accepts that the mere idea of a being like the God of traditional Christianity is not in itself incoherent. As there are certain difficulties with proper names for entities which do not exist – indeed, the phrase 'entity which does not exist' is regarded by some as self-contradictory – this atheist's view is more properly expressed as the claim that there could have been such a being as God, where the phrase 'such a being as God' is elliptical for a list of predicates which are supposed to characterize

this being: 'omniscient', 'omnipotent', and so on. Let us summarize this list in the predicate 'God-like'. Then the first part of our atheist's view, that there is no such being as God, may be formalized in this way:

(1) $\sim(\exists x)(x \text{ is God-like})$.

But how should we formulate the second part of his view, that there could have been such a being? There are two candidates:

(2) $\Diamond(\exists x)(x \text{ is God-like})$

and

(3) $(\exists x)\Diamond(x \text{ is God-like})$.

English renderings of (2) and (3) which respect the relative orders of the modal operator and quantifier in them are, respectively,

(4) It could have been that there is a God-like being

and

(5) There is a being which could have been God-like.

It is obvious that (4) and (5) do not mean the same, but unclear exactly what the difference between them is. We can bring the difference out by applying to (4) and (5) the quantifier treatment of the modal operators we developed in the previous chapter. As a first approximation, (4) and (5) become respectively:

(6) There is some world w in which there is a God-like being;

and

(7) There is some object x such that there is some world w in which x is God-like.

(6) and (7) state different requirements on the actual world $w^\star$. (7) requires that *there is* some object x which is God-like in some world, and if it is true at $w^\star$ that *there is* an x meeting this condition, then that object must exist at $w^\star$; otherwise, it cannot be true at $w^\star$ that *there is* such an object. This means that (7), and therefore (5) and (3), express the view that some entity which actually exists could have been God-like, even though that entity actually is not

God-like, if atheism is true. It is unlikely, then, that (3) is what our atheist means when he says that God's non-existence is contingent, for it is unlikely that he holds that some actual object could have been just as God is supposed to be (perhaps he believes that every actual object is made of matter, and it is not very easy to imagine a way things could have gone in which something which is actually material turns up as a 'spiritual' being).

So it looks as if (2), (4), and (6) are the formulations which we seek. (6) requires of the actual world that there be some world in which some being is God-like. Let us suppose that there is such a world, say, *u*. Then 'there is some God-like being' is true at *u*, which seems to require that in *u*, a God-like being exists (otherwise, it would not be true at *u* that *there is* such a being). It may be wondered if it is consistent to hold both that no actual object is God-like at any world, and that at some world, some being is God-like. But there is really no difficulty here. All we have to assume is that *there could have been things which do not actually exist*, the God-like thing in *u* being one of them. This is presumably what our atheist has in mind.

In sum, we can now see that the difference between (6) and (7) amounts to a difference in the 'domain' ranged over by the existentially quantified variable '*x*': in (7) it ranges over the actual world, but in (6) it ranges over whatever world is introduced by an evaluation of the existential quantifier over worlds. We will make this more precise below.

We can bring out the difference between $\Box(\forall x)Fx$ and $(\forall x)\Box Fx$ with a similar example. The atheist we are imagining believes that everything is made of matter, a belief which is true at the actual world iff all actually existing things are made of matter; therefore his belief is consistent with the view that there could have been things which are not made of matter. But let us imagine a different atheist who believes that it is a necessary truth that everything is made of matter; this atheist denies that there could have been a God-like being, since immateriality is one of the supposed attributes of God. What is the correct modal formulation of his view about matter? Again there are two candidates, (8) and (9):

(8) $\Box(\forall x)$(*x* is made of matter)

(9) $(\forall x)\Box$(*x* is made of matter).

(8) is what is wanted. Reading □ as a universal quantifier over worlds, (8) says that at every world *w*, '$(\forall x)$(*x* is made of matter)' is true, and this demands of each *w* that everything in *w*, that is, everything which exists at *w*, be made of matter in *w* (concomitantly with our treatment of the objectual quantifier in (6), the quantifier in (8) ranges over only the existents of a world). So (8) says that there are no worlds in which there exist non-material things.

(9) appears to be doubly inappropriate as a formulation of metaphysical materialism. If (9) is true at the actual world, then for every actual object *x*, it is necessary that *x* is made of matter, that is, every such *x* is made of matter at every world *w*. So if (9) is true at the actual world, then each actual object is made of matter in every world. But presumably an object cannot be made of matter at a world unless it exists at that world. So (9) implies that every actual object exists at every world, a statement that conflicts with the common sense view that material things exist contingently.

We can alter (9) so as to eliminate this unwanted implication:

(10) $(\forall x)\Box$(*x* exists $\rightarrow$ *x* is made of matter).

(10) says that for each object *x*, for each world *w*, if *x* exists at *w* then *x* is made of matter at *w*, that is, *x* is made of matter at every world at which *x* exists. But though this is consistent with *x*'s failing to exist in any number of worlds, the truth of (10) at the actual world is still not sufficient to express our second atheist's view; for (10)'s being true at the actual world is also consistent with there being some world in which there exists something which is not made of matter, something which, *a fortiori*, does not actually exist. This is inconsistent with the *a priori* materialism expressed by (8).

With these examples before us, we can reiterate the main points more formally. Let *Exw* be a predicate which means '*x* exists in *w*', and let 'Trans[φ]' stand for the result of translating a formula φ of first-order modal logic into possible worlds language according to the principles just developed. Then for any monadic predicate *F:*

(11) $\text{Trans}[\Diamond(\exists x)Fx] = (\exists w)(\exists x)(Exw \,\&\, Fxw)$;
(12) $\text{Trans}\,[(\exists x)\Diamond Fx] = (\exists x)(Exw^{\star} \,\&\, (\exists w)(Fxw))$;
(13) $\text{Trans}[\Box(\forall x)Fx] = (\forall w)(\forall x)(Exw \rightarrow Fxw)$;
(14) $\text{Trans}[(\forall x)\Box Fx] = (\forall x)(Exw^{\star} \rightarrow (\forall w)Fxw)$.

(11) should be compared with (2), (4), and (6), and (12) with (3), (5), and (7). Note that these translations use the existence predi-

cate to make explicit our intuitive judgements about the meanings of (6) and (7), according to which the variable 'x' ranges over different domains (worlds) in the two formulae. (13) and (14) differ similarly; according to (13), 'Necessarily, everything is F' means that in every world w, everything which exists at w is F at w, while according to (14), 'everything is necessarily F' means that everything which actually exists is F at every world. By the same principles, if we translate a formula with the form of (10), we obtain

(15) $(\forall x)(Exw^{\star} \rightarrow (\forall w)(Exw \rightarrow Fxw))$,

that is, 'every actual object is F at every world at which it exists'.[1]

By inspecting their translations, we can see that (instances of) the following schemata are not logically true.

(16) $\Diamond(\exists x)Fx \rightarrow (\exists x)\Diamond Fx$

is not logically true, because worlds may contain objects which do not actually exist, so that for particular instances of F, e.g. 'God-like' (on the first atheist's view) the conditional has a true antecedent and a false consequent. Similarly, the conditional

(17) $\Box(\forall x)Fx \rightarrow (\forall x)\Box Fx$

is not logically true: the antecedent might be true at the actual world since everything is F in every world, but the consequent false at the actual world, since not every actual object exists at every world (a clearly false substitution instance is obtained if 'exists' is put for F). Furthermore, elementary manipulations of (16) and (17), exploiting the interdefinabilities of $\Box$ and $\Diamond$, reveal that each instance of (16) is logically equivalent to some instance of

(18) $(\forall x)\Box Fx \rightarrow \Box(\forall x)Fx$.

1. As we see from these examples, monadic predicates of the modal language are translated as dyadic predicates in the possible-worlds language. Strictly, we should distinguish the monadic and the dyadic F typographically, but we allow the number of places displayed to resolve in context whether or not a predicate belongs to the modal language. Also, the translation scheme as exhibited treats the possible-worlds language as two-sorted, with one type of variable for individuals and another for worlds. Strictly, we regard the two-sorted formulae as abbreviations for single-sorted with special predicates. So, for example, '$(\forall w)(\forall x)(Exw \rightarrow Fxw)$' abbreviates '$(\forall y)(y$ is a world $\rightarrow (\forall x)(Exy \rightarrow Fxy)$'. 'is an individual' may be symbolized as the negation of 'is a world' so long as we are sharply distinguishing between worlds and individuals, as opposed to treating the former as a special kind of the latter.

Conversely, *mutatis mutandis* for (17) and

(19) $(\exists x)\Diamond Fx \rightarrow \Diamond(\exists x)Fx$

which therefore also fail to be valid schemata. To see directly that (19) has an instance that is not logically true, put 'does not exist' for *F*. Then the antecedent of the resulting instance is true at the actual world if some actual thing might not have existed, but the consequent is a *contradictio in adjecto*, since it requires that there be some world at which there exists something which does not exist. In the literature on modal logic, (16) and (18) are sometimes referred to as 'Barcan formulae', and (17) and (19) as 'converse Barcan formulae'.[2]

Semantics for quantified S5

The foregoing discussion imparts a general picture of what model theory for quantified S5 is going to look like. As in the sentential case, there will be a set of possible worlds, but in addition, each world will be assigned a set of objects, the things which exist at that world. The sum total of all objects which exist at the various worlds forms the set of all possible objects, and the atomic predicates and relation symbols of the language will be assigned extensions at each world drawn from this universal set, just as sentence letters were assigned truth values at each world. The details were given a classic formulation by Saul Kripke, and the presentation below is essentially a version of Kripke's semantics due to Kit Fine.[3]

An S5 model **M** for a set Σ of sentences of quantified modal logic consists of the following six components:

(i) a non-empty set W of possible worlds;
(ii) a non-empty set D of possible objects;
(iii) a function d which assigns to each world w in W a subset $d(w)$ of D; $d(w)$ is the set of objects which exist at w, sometimes called the *domain* or *inner domain* of w; the function d satisfies the condition that for every x in D there is some w in W such that x is in $d(w)$ (every possible object exists in at least one world);

2. After Ruth Barcan Marcus; see Marcus [1962].
3. See Kripke [1963] and Fine [1978].

(iv) for every n-place atomic predicate π occurring in some sentence in Σ and for each world w, a specification of which n-tuples of objects drawn from D (not: 'drawn from $d(w)$') are in the extension of π at w;

(v) for each individual constant occurring in some sentence in Σ, an assignment to it of a referent, an object in D (so the reference of a constant is the same at every world);

(vi) a selection of a particular w in W as $w^\star$.

As before, a sentence is true in a model **M** iff it is true at $w^\star$ in **M**. Quantified S5 is the system whose validities are exactly those arguments validated by the class of models conforming to (i)–(vi); that is, those arguments for which there does not exist a model conforming to (i)–(vi) that makes their premises true and conclusion false. (i)–(vi) define a notion of S5 model which is a natural extension of the notion defined for propositional logic in Chapter 1 (page 8). In that case, each possible world was associated with a model for non-modal propositional logic, that is, an assignment of truth values to sentence letters. Here we are associating each world with something like a model for non-modal first-order logic, which consists in a domain of discourse D and an assignment from D of extensions to predicate letters and referents to names. But because of clauses (iv) and (v), the parallel is not quite exact, and some further comments on these two clauses are in order.

Clause (v), according to which the referent of a constant at w is fixed independently of the nature of w and need not occur in $d(w)$, is now orthodox; for justification, note that if the sentence 'Socrates does not exist' is to be true at a world w where Socrates does not exist, then if this sentence is to be treated on a par with any other sentence of the form '$\sim F$(Socrates)', its truth at w must consist in the referent of 'Socrates' at w not being a member of the extension of 'exists' at w (the extension of 'exists' at w is $d(w)$); and because the sentence has to be true at w in virtue of its being *Socrates* who does not exist at w, it follows that the referent of 'Socrates' at w must be Socrates, even though he does not exist at w. In sum, the natural treatment of 'Socrates does not exist' demands that the name 'Socrates' denote Socrates at all worlds, even those at which Socrates does not exist. A constant which designates the same object at every possible world is sometimes called a (strongly) *rigid designator*. It is a somewhat controversial thesis that the proper names of natural language are rigid designators,

though it seems to be borne out by our example, but the controversy is irrelevant to the concerns of this book.[4]

Clause (iv) allows an atomic predicate to hold at a world *w* of an object which does not exist at *w*, and this seems strange, because the predicates of English which one thinks of as the natural candidates for regimentation by the atomic predicate letters of a formal language are usually predicates which can be satisfied only by existents: they are *existence-entailing* ('made of matter' or 'material' is a case in point). When we have an expression such as 'does not exist' which can be true at a world of non-existents at that world, we are inclined to think of the property it picks out as being in some way complex, and analysable into simpler components, so we would be reluctant to regiment it as an unstructured expression. Following Fine [1981], we may distinguish two versions of this attitude towards primitive predicates and non-existents. On one, primitive predicates neither apply nor fail to apply to non-existents at a world (see Davies [1978] for a version of this approach), while on the other, they apply, but are false of them. Since the second of these views is technically easier to implement, we shall work with it, that is, we shall work with what Fine calls the *Falsehood Principle*. In full generality, this principle says that at any world, no *n*-tuple of possible objects satisfies any *n*-place primitive predicate at that world unless every member of the *n*-tuple exists at that world.

Even if we are willing to assert the Falsehood Principle outright, however, clause (iv) of the model theory would still be justifiable, since it is evident that the considerations which recommend the Principle are highly philosophical in nature, and it would be bad practice to build such a philosophical view into modal logic itself: a dispute, for example, about the content of the notion of structure employed above should not be misrepresented as a dispute about logic. Moreover, there are even stronger grounds for retaining (iv), since not every predicate or relation which holds among non-exis-

4. The standard treatment is Kripke [1972], pp. 278–303. For objections, see Schiffer [1978]. Readers familiar with the literature may be surprised at the claim that the controversy about names is irrelevant to the metaphysics of modality, since arguments about a thesis known as the *necessity of identity* appear to turn on how names are treated. But this is a misconception, as I argue in Chapter 3; the necessity of identity asserts that one thing could not have been many nor many things one, a thesis with no implications for the semantics of proper names. I should add that Kripke is officially neutral between the exact treatment of names embodied in (v) and certain slight variants; see Kaplan [1989a], pp. 569–571, for details.

tents or between non-existents and existents, stands for a complex with genuine components in the rather obvious way in which 'does not exist' does (we give an example in footnote 22 to Chapter 7). So one may formalize some predicates as atomic and then, as a further step, ask whether or not the Falsehood Principle applies to each of them. From this point of view, it is better to regard the Principle as a schema, with some true instances and (perhaps) some false ones. The schematic version of the Principle is:

(F) $\Box(\forall v_1)\Box\ldots\Box(\forall v_n)\Box(\pi(v_1,\ldots,v_n) \rightarrow (Ev_1 \,\&\ldots\&\, Ev_n)$

where π takes only atomic substituends (readers who have difficulty construing (F) at this point should return to it at the end of this section). So if we decide, in any particular context, to assert (F) of some atomic predicate, we are asserting a non-logical axiom governing that predicate; where 'Mx' abbreviates 'x is made of matter', we have already adopted this instance of (F):

$$\Box(\forall x)\Box(Mx \rightarrow Ex).$$

The effect of the initial $\Box(\forall x)$ is to quantify over all possible objects, i.e., all elements of the domains of all worlds. The inner $\Box$ allows us to make the strong statement, of each possible object, that it satisfies $Mx \rightarrow Ex$ at all worlds, not just those where it exists.

Finally, in remarking on clauses (i)–(vi), we should note some consequences for the set of validities of the conditions on the function d which assigns domains to worlds. Because D is non-empty and d must assign each x in D to at least one world, the formula

$$\Diamond(\exists x)Ex$$

which means 'possibly something exists', is a logical truth. But because nothing prevents d from assigning the empty set to some world as its domain (recall that the empty set is a subset of every set), a valid schema of first-order logic,

$$(\forall x)\varphi x \rightarrow (\exists x)\varphi x$$

will fail at some worlds. Additionally, our treatment of quantifiers, on which the quantified variable ranges over just the domain of the world at which the quantified formula is being evaluated, has the consequence that $(\forall x)Ex$ is trivially true at every world, so that it, and also $\Box(\forall x)Ex$ are logically true. Another consequence of this

treatment, combined with clauses (iv) and (v), is that the schema

$$\varphi t \rightarrow (\exists v)\varphi v$$

is invalid: consider a world w such that no object in $d(w)$ satisfies φ, but a, which does not exist at w, does satisfy φ at w (simplest case: let φ be $\sim E$). So the usual rule of existential generalization fails in our system; rules for the quantifiers have to be the rules of free logic (for the empty domain).[5] Note that if we had decided to build the Falsehood Principle into our logic, changing clause (iv) accordingly, the above schema would in general still be invalid, though its instances with atomic F would be valid.

The next step in the development of the semantics is to give evaluation clauses for the logical operators, and here it pays to be rigorous about how the new operators, the objectual quantifiers, are to be handled. Let us step back from modal logic for a moment and consider evaluation clauses for the quantifiers in ordinary first-order logic. Suppose **M** is a model for ordinary first-order logic with domain $D_{\mathbf{M}}$ (henceforth we suppress the subscript). Under what circumstances is an existentially quantified sentence, say $(\exists x)(Fx \,\&\, Gx)$, true in **M**? One standard answer is that there must be some object a in D such that $Fx \,\&\, Gx$ is *true of* a. Those familiar with this approach will be aware of the complications involved in basing a fully general theory on the notion of 'true of' (those unfamiliar with the approach will be spared the details here). So instead we will develop the idea that $(\exists x)(Fx \,\&\, Gx)$ is true in **M** iff there is some object a in D such that the sentence obtained from $(\exists x)(Fx \,\&\, Gx)$ by deleting the string of symbols $(\exists x)$ and substituting some name of a for the free occurrences of the variable x is itself true in D. We use the underscore '_' to signify a function that takes an object to some fixed name of it. Then if we write $\varphi[\underline{a}/x]$ for the result of substituting the name $\underline{a}$ for each free occurrence of the variable x in the expression φ, we can say

> $(\exists x)(Fx \,\&\, Gx)$ is true in **M** iff for some object a in D, $Fx \,\&\, Gx\ [\underline{a}/x]$ is true in **M**.

5. For a comprehensive account of free logics, see Schock [1968]. To obtain a free logic for the empty domain it suffices to change the classical natural deduction rules to require $Et \rightarrow \varphi t$ as the premise for $\forall$-Intro and as the inferred formula for $\forall$-Elim, with corresponding changes in the $\exists$-rules so that the standard interdefinability is preserved.

The expression Fx & Gx [$\underline{a}/x$] stands for the sentence $F\underline{a}$ & $G\underline{a}$, so we obtain the correct result that $(\exists x)(Fx$ & $Gx)$ is true in **M** iff for some a in D, $F\underline{a}$ & $G\underline{a}$ is true in **M**. Thus the truth of an existentially quantified sentence consists in its having some true singular instance. (If Smith, Jones, and Robinson, are all the people in a domain D, then we can say that 'Someone is a spy' is true in D iff for some a in D, '$\underline{a}$ is a spy' is true D. In this example, this reduces to the condition that at least one of 'Smith is a spy', 'Jones is a spy', and 'Robinson is a spy', is true in D.)

However, there is a minor complication to deal with. Consider:

Some planet in the Milky Way is larger than the Sun.

Though an existentially quantified sentence of English, this example's truth-value cannot be settled by whether or not it has a true English instance of the form 't is a planet in the Milky Way & t is larger than the Sun'. It may well be that there are planets larger than the Sun in other solar systems, but that none of these have names in English, while none of the planets which do have English names is larger than the Sun. The example shows that the suggested approach to the existential quantifier works only if we are evaluating sentences of a language which has a name for every object within the range of the existentially quantified variable. So we shall just assume that we are always working with such a language. A model **M** for a set of sentences Σ will therefore be understood from now on to be a model for the language L that contains the symbols in the sentences in Σ and also an individual constant $\underline{a}$ for each object a in the domain of the model. With this explanation, we can now generalize our example to give the following clauses (∃) and (∀) for ordinary first-order logic:

(∃) A sentence of the form $(\exists v)\varphi v$ is true in a model **M** with domain D iff for some object a in D, $\varphi v[\underline{a}/v]$ is true in **M**.

(∀) A sentence of the form $(\forall v)\varphi v$ is true in a model **M** with domain D iff for every a in D, $\varphi v[\underline{a}/v]$ is true in **M**.

With one more piece of notation, we can proceed with the evaluation clauses for quantified modal logic. Returning to the definition of S5 model just given, let us embody the specification described in clause (iv), page 27,which gives the extensions of predicates and relation symbols at worlds, in a two-place function

Ext. Thus to say that $Ext(H, w) = \{\langle a,a\rangle, \langle b,b\rangle\}$ is to say that the extension of the two-place relation symbol H at world w are the two pairs of objects $\langle a,a\rangle$ and $\langle b,b\rangle$: that is, the only atomic sentences involving H which are true at w are Haa and Hbb. And let us embody the assignment of a referent to each constant in the language of the model, as described in clause (v), page 27, in a function *Ref;* so *Ref*('Plato') = Plato says that the referent of 'Plato' is Plato. The evaluation clauses are then as follows:

(vii) an atomic sentence of the form $\pi(t_1,\ldots,t_n)$ is true at a world w iff $\langle Ref(t_1),\ldots,Ref(t_n)\rangle$ is a member of $Ext(\pi, w)$;
(viii) an identity sentence $t = t'$ is true at a world w iff $Ref(t) = Ref(t')$;
(ix) Et is true at w iff $Ref(t)$ is in $d(w)$;
(x) $\sim\varphi$ is true at w iff φ is not true at w;
(xi) φ & ψ is true at w iff φ is true at w and ψ is true at w;
(xii) $\Diamond\varphi$ is true at w iff for some u in W, φ is true at u;
(xiii) $\Box\varphi$ is true at w iff for all u in W, φ is true at u;
(xiv) $(\exists v)\varphi v$ is true at w iff $\exists a \in d(w)$: $\varphi[\underline{a}/v]$ is true at w;
(xv) $(\forall v)\varphi v$ is true at w iff $\forall a \in d(w)$, $\varphi[\underline{a}/v]$ is true at w.

Two points about these evaluation clauses should be noted, the first concerning the treatment of objectual quantifiers in (xiv) and (xv), the second the treatment of $\Box$ in (xiii). In (xiv) and (xv) we have embodied our idea that the truth of a quantified sentence at a world should turn only on how things are with the objects which exist at that world (unless the sentence contains names of things which do not exist there). Instead of 'exist at that world' we might have used the locution 'are actual at that world', involving a relativized notion of actuality. Because of this terminology, the treatment of the quantifiers in (xiv) and (xv) is sometimes called the *actualist* interpretation. One justification for this treatment is that it straightforwardly accommodates non-modal quantified sentences; the non-modal assertion that everything is made of matter is true *simpliciter* iff all actual (i.e. existing) things are made of matter, so in this non-modal use the universal quantifier ranges over only the existents of the actual world. Certainly, we do not want to say that that assertion is false because things which might have existed, but do not, are not actually made of matter. So outside the scope of a modal operator, objectual quantifiers are evidently actualist, and (xiv) and (xv) generalize this. Of course, it is easy to add

extra objectual quantifiers to modal language, the so-called 'possibilist' quantifiers Π and Σ, where $(\Pi v)\varphi v$ is true at a world w iff for every a in D (not just $d(w)$), $\varphi[\underline{a}/v]$ is true at w, but we will in fact have no use for such quantifiers.[6]

The second point to notice is that, according to (xiii), a sentence such as $\Box Fa$ is true at any world iff Fa is true at every world. This is sometimes known as *strong* necessity, in contrast to an interpretation of '$\Box$' on which the truth of $\Box F\underline{a}$ at w would require only the truth of $F\underline{a}$ at all worlds where a exists, an interpretation known as weak necessity. One good reason for choosing the strong rather than the weak interpretation of necessity is that use of the latter deprives us of any natural way of regimenting certain modal propositions. The simplest real example is once again from theology, where certain writers have wanted to say that God's existence is non-contingent, or that He exists of necessity, i.e., in every possible world. With strong necessity, this proposition is expressed by '$\Box$(God exists)', but when $\Box$ is read weakly, that formula only requires that God exist at every world where He exists, a condition which every possible object satisfies trivially. This is an example of an expressive weakness in modal language with weak $\Box$, relative to modal language with strong $\Box$, and many other examples can be given. So (xiii) is to be preferred.[7]

Let us end by illustrating the semantics, first with counterexamples to the formulae (16) and (17) from earlier in this chapter:

(16) $\Diamond(\exists x)Fx \rightarrow (\exists x)\Diamond Fx$

(17) $\Box(\forall x)Fx \rightarrow (\forall x)\Box Fx$.

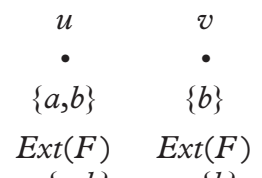
To defeat (16), let $W = \{u,v\}$, $D = \{a,b\}$, $d(u) = \{a\}$, $d(v) = D$, and suppose that the extension of F is empty ($= \varnothing$) at u and is $\{b\}$ at v. We may picture this as exhibited in the margin (the second place of *Ext* is suppressed). Choose u for $w^\star$. Then $(\exists x)\Diamond Fx$ is false at $w^\star$ by clause (iv), page 27, since there is no a in $d(u)$ such that $\Diamond F\underline{a}$ is true at u; in particular, $F\underline{a}$ is false at both u and v; on the other hand, $\Diamond(\exists x)Fx$ is true at u, since $(\exists x)Fx$ is true at v (because Fb is true at v). So (16) is false in this model.

u	v
•	•
$\{a\}$	$\{a,b\}$
$Ext(F)$ $= \varnothing$	$Ext(F)$ $= \{a,b\}$

6. See Fine [1981a], pp. 192–193, and the Appendix to this book, for more on possibilist quantifiers.

7. The terminology is from Kripke [1971], p. 137. The S5-system of Davies [1978] employs weak necessity. See Hazen [1976] for further examples of expressive weakness.

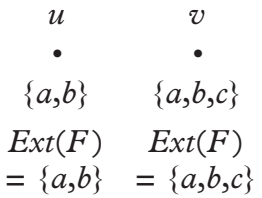
To obtain a counterexample to (17), let $W = \{u,v\}$, $D = \{a,b\}$, $d(u) = D$, $d(v) = \{b\}$, and for each w, let the extension of F be $d(w)$, as in the picture. Choosing u for $w^\star$, we have $\Box(\forall x)Fx$ true at $w^\star$ since $(\forall x)Fx$ is true at u and at v, but $(\forall x)\Box Fx$, is false at $w^\star$ because $\Box F\underline{a}$ is false at v in virtue of a's not being in $Ext(F, v)$.



These examples involve rather simple formulae with just one modal operator and one quantifier. To impart a proper "feel" for the semantics, we should consider more complicated arrangements of operators and quantifiers. For instance, we have seen that $\Box(\forall x)Fx$ and $(\forall x)\Box Fx$, are not equivalent, but how does the formula $\Box(\forall x)\Box Fx$ stand to these two? This formula implies both of the others, so let us consider the converses, (20) and (21):

(20) $\Box(\forall x)Fx \rightarrow \Box(\forall x)\Box Fx$
(21) $(\forall x)\Box Fx \rightarrow \Box(\forall x)\Box Fx$.

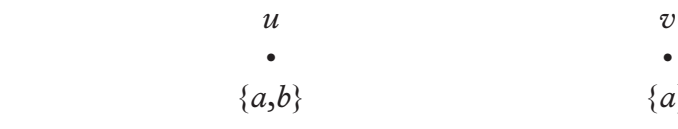
The counterexample to (17) displayed above is already a counterexample to (20), since, in that model, $\Box(\forall x)Fx$ is true at $w^\star$ while $(\forall x)\Box Fx$ is false there, and if $(\forall x)\Box Fx$ is false at some world, then $\Box(\forall x)\Box Fx$ is false at every world. A more interesting question is whether we can have (17) true but (20) still false in a model. The answer is in the affirmative. Let $W = \{u, v\}$, $D = \{a,b,c\}$, $d(u) = \{a,b\}$, $d(v) = D$, and for each w, $Ext(F, w) = d(w)$. Here $\Box(\forall x)Fx$ is true at $w^\star$ $(= u)$ since $(\forall x)Fx$ is true at u and v; also, $(\forall x)\Box Fx$ is true at $w^\star$ since a and b are in the extension of F at both worlds, so (17) is true in this model, as is the antecedent of (20). But the consequent of (20), $\Box(\forall x)\Box Fx$, is false at $w^\star$, since $(\forall x)\Box Fx$ is false at v, because $\Box Fc$ is false at v, since Fc is false at u. Note that in the course of this argument (21) was also refuted.



Consider next the formula

(22) $\Box(\forall x)(\forall y)Fxy \rightarrow \Box(\forall x)\Box(\forall y)Fxy$

and the model where $W = \{u, v\}$, $D = \{a,b\}$, $d(u) = D$, $d(v) = \{a\}$, $Ext(F,u) = \{\langle a,a\rangle, \langle b,b\rangle, \langle a,b\rangle, \langle b,a\rangle\}$ and $Ext(F,v) = \{\langle a,a\rangle\}$. The picture for this model is as follows:



$(\forall x)(\forall y)Fxy$ is true at $w^\star$ $(= u)$ and at v, so $\Box(\forall x)(\forall y)Fxy$, is true

at $w^\star$. Let us now evaluate the consequent of (22) step by step. If $\Box(\forall x)\Box(\forall y)Fxy$ is true at $w^\star$, then $(\forall x)\Box(\forall y)Fxy$ must be true at u and at v; and if this latter formula is true at u, then $\Box(\forall y)F\underline{a}y$, and $\Box(\forall y)F\underline{b}y$ must both be true at u (by clause (xiii) on page 32). But $\Box(\forall y)Fby$ is false at u since $(\forall y)Fby$ is false at v, in turn because $Fb\underline{a}$ is false at v. It follows that the consequent of (22) is false at $w^\star$, and since we already saw that its antecedent is true there, (22) is refuted by our model. Readers may like to confirm their understanding of this reasoning by constructing a counterexample to:

(23) $\Box(\forall x)\Box(\forall y)Fxy \rightarrow \Box(\forall x)\Box(\forall y)\Box Fxy$.

Lastly, consider the formula

(24) $\Box(\forall x)\Box Ex$

which reads 'necessarily everything necessarily exists'. The import of this claim is easier to grasp in possible worlds discourse, where we may render it 'in any world, anything in that world exists at every world'. This formula therefore has the effect of ruling out the kind of situation we exploited to construct the various models above, where an element in the domain of one world failed to appear in the domain of another: (24) says that any object appearing in the domain of one world also appears in the domain of every other. It follows from this that in any model **M** in which (24) is true, the domains $d(w)$ of all the worlds w in W must be the same. In fact, since each object in D is assigned to the domain of at least one world, it follows that the domain of every world is D itself, the set of all possible objects. One variant of the system we have presented adds to the definition of model the condition that the domain of every world is D; we might call this variant 'S5 with constant domains', or, for short, 'S5B'. Here 'B' is for 'Barcan', since, as is easy to check, (24) holds in a model iff the Barcan and converse Barcan schemata are valid in the model. But we will regard the 'unchanging domains' condition as an extra stipulation, like the Falsehood Principle (though without a similar philosophical rationale), and what we have just seen is that this condition can be formulated as a single sentence, (24), of our modal language. Of course, (24) is invalid in our system as it stands – it fails in every model we have presented so far – so when it is necessary to contrast our system with S5B, we shall call it 'S5C', where 'C' connotes 'with contingent existence'.

First-order tense logic

Since we have already laid the groundwork in the sentential case, we might at this point proceed with the introduction of first-order versions of the systems T, B (here 'B' is for the propositional system B, not for 'Barcan' – see page 14) and S4. However, as we will not want to use any of these systems in what is to come, developing them would involve too great a digression. Instead, we will pursue the not wholly unrelated question of what happens when we replace our modal operators with operators corresponding to the tenses of natural language; for it will turn out that model theory for tense logic (the logic of the tense operators) bears a striking resemblance to model theory for modal logic with an accessibility relation, a resemblance with some philosophical significance.

First, we extract tense operators from the tenses of sentences just as we extracted modal operators from their moods. Consider the sentence

(25) Nixon resigned.

A more cumbersome cognate of (25) is

(26) It was the case that Nixon resigns

which suggests that simple subject-predicate tensed sentences such as (25) may be thought of as contractions of longer sentences which contain an untensed sentence as a part, modified by a tense operator, in this instance, 'it was the case that'. This (loosely) parallels the idea that subjunctive English sentences are contractions of indicative sentences governed by a modal operator (for 'loosely', see Evans [1985]). We will symbolize 'it was the case that' by the operator $\mathbb{P}$, and we will also have a future-tensed operator $\mathbb{F}$, meaning 'it will be the case that'.[8] It is now possible to formulate a definition of a general model **M** for a language L which is just like a first-order modal language, except that it has the operators $\mathbb{P}$ and $\mathbb{F}$ instead of $\Box$ and $\Diamond$.

A general model **M** for a first-order language L with tense operators has the following seven components:

8. If tense operators modify untensed sentences, then strictly, there ought to be a present-tense operator for forming present-tense sentences from untensed sentences. In tense logic, the effect of the present-tenseis obtained by the absence of tense operators, but this is philosophically dubious. See Evans [1985].

(i) a non-empty set T of times;
(ii) a non-empty set D of objects;
(iii) a function d such that for each time t in T, $d(t)$ is a subset of D, intuitively, the objects which exist at t;
(iv) for each n-place predicate F of the language L, and each time t in T, a specification of which n-tuples of objects in D (not: 'in $d(t)$') are in the extension of F at t;
(v) for each individual constant in the language L, a specification of which object in D is the referent of the constant;
(vi) a designation of a particular t in T as $t^\star$, the present moment;
(vii) for each time t in T, a specification of which other times t' in T are preceded by t; we write '$\textsc{Prec}(t, t')$', which may be read as 't precedes t'', 't is before t'', 't' is after t', etc.

Some of the features of this model theory for tense logic help to illuminate corresponding features in the S5 semantics. In place of worlds, we have moments of time, and the changes in the sets $d(t)$ as t changes represent things coming into and going out of existence as time passes. By (v), individual constants are temporally rigid: they have the same reference at each time regardless of what exists at that time. It seems clear that proper names in natural language are temporally rigid: at all times, 'Julius Caesar' denotes the man who happened to be the Roman conqueror of Gaul. And it is quite reasonable to allow the extension of a predicate or relation symbol at a time to include n-tuples of objects containing members which do not exist at t. For instance, the two-place relation *is a descendant of* holds now between the author and his father, although the latter no longer exists; and this relation will continue to hold of these two persons at later times when neither exists.

To complete the semantics, we have to give evaluation clauses for the logical constants of L, which are just those from the modal case except for the replacement of modal operators by tense operators. So the evaluation clauses for the non-modal operators carry over to the tense logical case, with minor terminological changes: for instance, we are evaluating formulae at times, not worlds. But from the formal point of view this is not a genuine difference. To these clauses we have to add two clauses for the tense operators. These should reflect the intuitive meaning of $\mathbb{F}$ and $\mathbb{P}$ and should also refer to the relation $\textsc{Prec}$, which fixes the range of a tense operator in any particular evaluation. So we have

(𝔽) $\mathbb{F}\varphi$ is true at a time t in T iff there is some time t' in T such that $\textsc{Prec}(t, t')$ and φ is true at t';

(ℙ) $\mathbb{P}\varphi$ is true at a time t in T iff there is some time t' in T such that $\textsc{Prec}(t', t)$ and φ is true at t'.

Given these operators, it is easy to introduce more into our language by definition. For instance, we can have an operator meaning 'henceforth', since 'henceforth φ' may be defined as $\sim\mathbb{F}\sim\varphi$; we can have an operator 'heretofore', since 'heretofore, φ' may be defined as $\sim\mathbb{P}\sim\varphi$; and we can even keep tense analogues of our modal operators $\Box$ and $\Diamond$, which in the context of tense logic have the senses 'always' and 'sometimes'; these operators may either be defined – $\Box\varphi$ means φ & $\sim\mathbb{F}\sim\varphi$ & $\sim\mathbb{P}\sim\varphi$ – or their modal evaluation clauses ($\Box$) and ($\Diamond$) from page 12 can be transcribed into tense-logical terminology.

A consequence of our retaining the evaluation clauses for the objectual quantifiers $\forall$ and $\exists$ is that when a quantified sentence is evaluated at a time, the range of the quantifier is just the domain of things which exist at that time. So validities and invalidities parallel to those we investigated in the previous section also arise in tense logic. To illustrate briefly, consider the formula

(27) $\mathbb{F}(\exists x)Cx \rightarrow (\exists x)\mathbb{F}Cx$.

It is left to the reader to construct a formal counterexample to (27), combining the methods of involved in constructing counterexamples to invalid first-order modal arguments and invalid sentential modal arguments using accessibility, but it is much easier to see the invalidity of (27) at the intuitive level than it was to see the invalidity of comparable modal formulae. For the predicate C substitute 'travels to Jupiter'. Then we can suppose that in the future someone does travel to Jupiter, so the antecedent of (27) is true at the present moment. But the consequent may well be false, for there may be no-one who exists at the present moment who travels to Jupiter in the future: perhaps the first expedition from Earth will not leave until the twenty-second century.

Of course, alternative tense-logical clauses to the ones we have given are also possible. The main candidates are motivated by the idea that the present and past are "real" in a way that the future is not, and that this should be reflected in the semantics. Thus it might be objected, first, that the rigidity of a proper name should

consist only in its having a certain denotation x at and after the time at which x has come into existence; secondly, that quantifiers should be allowed to range over not just the existents of the time of evaluation, but also the existents of all previous times; and thirdly, that the Falsehood Principle should be imposed on atomic predicates for times before all members of the relevant n-tuple exist (so the author and his father would not be said to stand in the relation 'is a descendant of' at any time before the author existed).

There is nothing especially unworkable in these suggestions from the technical point of view, but there is a danger of confusion in the grounds that may be advanced on their behalf. For instance, it is clear that 'Julius Caesar' only entered linguistic usage when the name was bestowed, and was not used by speakers to refer to Julius Caesar before he existed (this does not mean that one cannot use a name to refer to an object which does not yet exist – consider the use of 'Brasilia' before the pre-planned city was built – but only that 'Julius Caesar' was not a name of this sort). However, when we ask whom 'Julius Caesar' denotes in present-day English, we are asking about its referent in a particular language, present-day English, and one attitude we can take is that it is eternally true of that language that, in it, 'Julius Caesar' denotes Julius Caesar, for the facts about that language are unchanging; in much the same way, it is eternally true that the year of the Battle of Waterloo is 1815. What is sometimes called evolution of, or change in, a language, is best regarded, from this point of view, as change *of* language (but see Deutsch [1989] for a very different account).

The idea that the quantifiers might be read as neutral between the present and past tense is not unnatural, but it would require the addition of an operator 'Now' or a second existence predicate if we wished to get the effect of an assertion that there exists now such-and-such a thing. Since the system we have defined is expressively equivalent to the suggested revised system, the former is somewhat simpler.

The third suggestion, of a Falsehood Principle for the future, provides a good illustration of the idea of simple versus complex properties with which we made play in connection with the modal Falsehood Principle. Even if there is some intuitive support for the thought that the author and his father do not stand in the relation 'is a descendant of' at times before the author exists, it is less clear that we have any difficulty in the idea that the author's father and

the author stand in the relation 'is a parent of' at times before the author exists. But a defender of the future-oriented Falsehood Principle could say that 'is a parent of' is not itself simple, but is analysable into some complex expression involving the future tense and predicates for simple properties or relations (it could be said that '*a* is a parent of *b*' holds before *b* comes into existence in virtue of the future holding of certain simple relations between appropriate entities, e.g., cells, that exist together then). Nevertheless, however defensible this line of argument may be in this case, the same point applies to tense logic as to modal: the claim that some such analysis is always correct is a philosophical view which should not be built into logic. So our conclusion is that the system of tense logic defined in the clauses just given has the same preferred status as the quantified S5 system.[9]

Possibility semantics for quantified S5

We will complete this part of our discussion of semantics for modal logic by sketching how the possibility semantics we gave for sentential S5 in the last part of Chapter 1 may be extended to quantified S5. Again, our aim is to provide a semantics which validates exactly the arguments validated by the orthodox S5 semantics, this time first-order semantics. Our new semantics is to permit possible worlds to be incomplete. We are again thinking of possibilities as being completely specified by sentences beginning 'it is possible that', so the resulting incompleteness will have two aspects. First, such a sentence will mention only a few out of all possible objects

9. Just as it is possible to generate different systems of modal logic by imposing different structural constraints on the accessibility relation, so it is possible to generate different systems of tense logic by imposing different structural constraints on the precedence relation: PREC is the accessibility relation of tense logic. PREC's structural properties can be made to correspond rather clearly with intuitive or philosophical doctrines about the structure of time. The most intuitively appealing conception of the time sequence is that it is linear, without beginning and end, and continuous (this means that it contains no gaps: contrast the rational numbers, where there are gaps corresponding to irrationals like $\sqrt{2}$). On this view, the time sequence is formally indistinguishable from the real number sequence, and so if we impose the requirement that the PREC relation be formally indistinguishable from the relation 'is (strictly) less than' on the real numbers, we obtain a tense logic for time as thus conceived. However, there are many alternative structural constraints we could impose on PREC, to obtain discrete time, or time with a beginning and/or end, or circular time, and so on; and we can investigate how changing the structural constraints alters the set of formulae validated by the semantics, though we will not pursue these topics here. The pioneer of tense logic was A. N. Prior; see, for instance, Prior [1967]. A full survey is given in Rescher and Urquart [1971].

and use only a few of the predicates of the language; and second, even with respect to the objects mentioned in it and the predicates used, many of the latter will be such that the possibility-specification does not determine whether or not they apply to the mentioned objects. So we will think of a possibility ρ as an entity which has assigned to it, first, a domain of objects; and second, for each *n*-place primitive predicate *F* belonging to some subset of the predicates of *L*, a specification of an extension to *F* at ρ, and also of a counterextension to *F* at ρ, the extensions and counterextensions being drawn from the set *D* of all possible objects. When an *n*-tuple of objects is in the extension of an *n*-place primitive predicate *F* at a possibility ρ, this means that *F* is determined to be true of that *n*-tuple at ρ; and if an *n*-tuple is in the counterextension of *F* at a possibility ρ, this means that *F* is determined to be false of that *n*-tuple at ρ. In general, for each ρ and *F*, many *n*-tuples of objects drawn from *D* will be neither in the extension or the counterextension of *F* at ρ: *F* is not determined to be true of them and not determined to be false of them. This is how incompleteness arises.[10]

Referents will be assigned to constants of *L* as before, in a once-for-all manner, so when the extension and counterextension specification for a set of predicates is given for a particular possibility, certain atomic sentences are immediately made true or false at that possibility; one might even replace the assignment of extensions and counterextensions to predicates by an assignment of truth-values to atomic sentences, with the same effect. This means that the special concept of possibility semantics, refinement, may be defined as in the sentential case: a possibility τ refines ρ ($\tau \gg \rho$) iff every atomic sentence for which ρ is defined has the same truth value at τ as it has at ρ. So let us say that a *p*-model for quantified S5 has the following components:

(i) a set Ω of possibilities;
(ii) a set *D* of objects;
(iii) for each ρ in Ω, and for some subset of predicates and relation symbols in *L* (excluding the identity symbol), an assignment of an extension and a counterextension at ρ to

10. In English we sometimes use the phrase 'the possibility that φ', where the expression φ is logically complex. On the present treatment of possibilities, this use of 'the' is improper: for such a φ, there can be many distinct possibilities that φ, since many distinct assignments of extensions and counterextensions to atomic predicates may yield possibilities at which φ is true.

each symbol in the set; the extensions and counterextensions are drawn from D, and need not involve only, nor all, the members of $Ext(E, \rho)$ (the extension of the existence predicate at ρ fixes the domain of ρ, which is why we do not need to specify a function d from possibilities to subsets of D);[11]

(iv) for each individual constant in L, an assignment of a referent from D;

(v) a selection of a particular possibility ρ in Ω for $\rho^\star$, the actual possibility; again, we stipulate that the actual possibility must be complete, which is to say that for each predicate and relation symbol of degree n in the language, any n-tuple of objects drawn from D is in either the extension or in the counterextension (though not both) of the symbol at ρ.

Finally, we say that a p-model is a structure conforming to (i)–(v) in which the refinement relation induced by the assignments of clauses (iii) and (iv) satisfies the Refinability condition; this, in the present context, is the condition that for any atomic $\pi(t_1,\ldots,t_n)$, if $\pi(t_1,\ldots,t_n)$ is undefined at ρ in Ω then there are τ and ζ in Ω such that $\tau \succcurlyeq \rho$ and $\zeta \succcurlyeq \rho$ and $\pi(t_1,\ldots,t_n)$ is true at τ and $\pi(t_1,\ldots,t_n)$ is false at ζ; clearly, this requires the n-tuple of objects $\langle Ref(t_1),\ldots,Ref(t_n)\rangle$ to be in the extension of π at τ and the counterextension of π at ζ.

As with the sentential case, the complexity of possibility semantics manifests itself in the evaluation clauses. Because we can think of first-order possibilities as assignments of truth-values to atomic sentences, we can take over the clauses for negation and conjunction which we gave in on pages 19–19. But we need new clauses for atomic sentences (including existence sentences) and identity sentences, and also for the quantifiers. Here are the details:

(vi) $\pi(t_1,\ldots,t_n)$ is true at ρ iff $\langle Ref(t_1,\ldots,Ref(t_n)\rangle$ is in the extension of π at ρ, and is false at ρ iff $\langle Ref(t_1),\ldots,Ref(t_n)\rangle$ is in the counterextension of π at ρ;

(vii) $t = t'$ is true at ρ iff $Ref(t) = Ref(t')$;

11. This is for variety – we could have taken the same course in our development of standard possible worlds semantics. Note that a sentence-letter is a 0-place predicate letter whose only possible extension is T and only possible counterextension is F.

Clause (vii) does not mention possibilities on its right-hand side, so the same atomic identity sentences are true at every possibility, which ones are true depending only on the assignment of clause (iv). This simply repeats a feature of possible worlds semantics, that because individual constants are rigid designators, the truth-values of atomic identity sentences are settled independently of the contents of the possible worlds.

To deal with the existential quantifier, we have to preserve the usual analogy with disjunction.[12] As will be recalled from page 19,

(viii) $\varphi \vee \psi$ is true at ρ iff for every τ which refines ρ there is some ζ which refines τ such that φ is true at ζ or ψ is true at ζ.

An existentially quantified sentence can be thought of as equivalent to a disjunction of related sentences with names instead of a quantified variable: our example on page 31, 'Someone is a spy', illustrates this: in the domain of the example, the existential sentence is equivalent to 'Smith is a spy or Jones is a spy or Robinson is a spy'. More generally, in the context of free logic, the truth of an existentially quantified sentence $(\exists v)\varphi v$ at a world or at a possibility is equivalent to the truth of a single disjunction with multiple disjuncts, each of the form $E(t)$ & φt, where there is one such disjunct for each object in the domain of the world or possibility. This means that if infinitely many objects exist at the world or possibility, then the disjunction is of infinite length; but no matter, for we have already included enough names in the language to name each object. Note also the appearance of the existence predicate in each disjunct, which ensures that whether or not the quantified sentence is true at the possibility turns only on facts about objects which would exist were the possibility realized. Thus a simple way of stating the clause for the existential quantifier is in effect to say that a quantified sentence is true at ρ iff the associated disjunction is true there, at the same time applying (viii) to analyse what is involved in the disjunction being true at ρ. This yields the clause:

(ix) $(\exists v)\varphi v$ is true at ρ iff for any τ refining ρ there is some ζ refining τ and some name t such that $E(t)$ & φt is true at ζ (here φt is of course $\varphi v[t/v]$).

Furthermore, it is not difficult to verify that if a formula ψ is a dis-

12. Lloyd Humberstone pointed this out to me.

junction, even an infinitary one, each of whose disjuncts is of the form $E(t)$ & φt, then the negation of ψ is equivalent to a conjunction (infinite if ψ is infinite) each of whose conjuncts is of the form $E(t) \rightarrow \sim\varphi t$. We already have a clause for &, so if $\forall$ is introduced by the usual definition '$\sim\exists\sim$', we can infer that the following is the correct clause for $\forall$:

(x) $(\forall v)\varphi v$ is true at ρ iff for all names t, $E(t) \rightarrow \varphi t$ is true at ρ.

Readers should pause to confirm that they understand the reasoning here, which will involve working out the clause for material implication by unpacking its definition in terms of ~ and &.

Lastly, what becomes of the modal operators? Although we can just take over the sentential clause for $\Diamond$, it is not trivial that the sentential clause for $\Box$ is still acceptable. Readers who worked through the proof that $\Box\varphi$ is true at ρ (in the sentential case) iff φ is true at every τ in Ω would have discovered that the following principle is required: for any (not necessarily atomic) sentence φ, if φ is false or undefined at ρ, then ρ has some refinement at which $\sim\varphi$ is true. However, it turns out that this principle is correct in the first-order case too, so both modal operator clauses carry forward.[13]

With these clauses, we may demonstrate the equivalence of possibility semantics with the possible-world semantics for quantified S5. Each world-model is already a *p*-model. To show that for each *p*-model there is a world-model that verifies the same sentences, the simplest procedure is again to take some adequate set of postulates and/or inference rules, such as those given in Fine [1978], and to check that they are correct for any possibility in any model. By this soundness result, if Σ is a set of sentences with a possibility model, then Σ is S5 consistent, and so by the completeness of the postulates and rules, Σ has an S5 possible-worlds model.

It is even clearer in the first-order case that possible worlds semantics represents a considerable simplification over possibility

13. Suppose $\Box\varphi$, that is, $\sim\Diamond\sim\varphi$, is true at w. Then for each u refining w $\Diamond\sim\varphi$ is not true at u. Unpacking, this means that there is no v such that $\sim\varphi$ is true at v. Hence every possibility ρ has some refinement τ such that φ is true at τ. But this does not show that φ is true at every possibility ρ. However, if φ is not true at some ρ, then by the principle mentioned in the text, ρ has a refinement τ such that $\sim\varphi$ is true at τ; and we have just shown that τ must have a refinement ζ at which φ is true. Since it is easy to prove by induction that a formula true at a possibility is true at every refinement of that possibility, it follows that φ and $\sim\varphi$ are both true at ζ, a contradiction.

semantics. The point of our developing the latter system is not to make practical use of it, but simply to establish that it is in principle possible to do without the condition that the entities with respect to which we evaluate the formulae of a modal language must be complete with respect to that language. The theoretical dispensability of the completeness requirement will be of some interest later.

CHAPTER 2: FIRST-ORDER MODAL LOGIC

Chapter 3
The *De Re*/*De Dicto* Distinction

Two kinds of formula

IN EVALUATING formulae of modal and tense logic in Chapter 2, we found that the order in which operators and objectual quantifiers are arranged is significant. We also found that in evaluating formulae which contain operators within the scope of objectual quantifiers, the identity of the objects in the domains of the various worlds is important. Thus, if '$(\forall x)\Box(x$ is made of matter)' is to be true at a world w, this requires that the objects which exist at w are made of matter at all worlds; but if '$\Box(\forall x)(x$ is made of matter)' is to be true at w, the identity of the objects made of matter at other worlds is irrelevant, so long as, at each world, all the objects which exist there are made of matter there. So if we consider an arbitrary world u, the truth of the first sentence at w requires that the existents of w be made of matter at u, and so to determine the truth-value of that sentence, we have to be able to ascertain the composition at u of the existents of w. But no such cross-reference between worlds is needed to ascertain the truth value of '$\Box(\forall x)(x$ is made of matter)' at w. The contrast between the two formulae here is a special case of a more general contrast between formulae which are '*de re*' (literally, 'about the thing') and formulae which are '*de dicto*' (literally, 'about the statement'). So the rough idea is that '$(\forall x)\Box(x$ is made of matter)' says something about certain things, specifically, that any actual object is necessarily material; while '$\Box(\forall x)(x$ is made of matter)' says something about the statement '$(\forall x)(x$ is made of matter)', specifically, that it is necessarily true.

We can give a precise definition of this distinction in terms of the syntactic structure of formulae:

- A formula with modal or tense operators is *de re* iff it contains a modal or tense operator $\mathbb{R}$ which has within its scope either (i) an individual constant, or (ii) a variable not bound by a quantifier within $\mathbb{R}$'s scope. All other formulae with modal or tense operators are *de dicto*.

Hence $\mathbb{P}(\exists x)Fx$ and $\Box\sim(\forall x)Fx$ are *de dicto*, while by (i), $\Diamond Fa$, and by (ii), $\Box Fx$, $(\forall x)\Box Fx$, and $\Diamond(\exists x)(Fx \,\&\, \Diamond Gx)$, are all *de re*, in the last case because there is at least one modal operator in the formula (the last $\Diamond$) which has a variable within its scope that is not bound by a quantifier also within its scope.[1]

The difference between *de re* and *de dicto* formulae, as we see from the example about matter, is a difference between formulae which are, and formulae which are not, sensitive to the identities of objects at various worlds. In evaluating a *de dicto* formula φ in a model, we eventually reach subformulae of φ whose main connectives are modal or tense operators which have within their scope complete sentences which themselves do not contain any individual constants or modal or tense operators. $\Box(\forall x)Fx$ is already of this form, and in $\Diamond(\forall x)Fx \rightarrow \Box(\forall x)Gx$ or $\mathbb{P}\mathbb{P}(\exists x)Fx$, we reach such subformulae after one application of the evaluation rule for the main connective of each formula itself. Having reached such a subformula, one then applies the evaluation rule for the governing modal or tense operator, which in turn will require one to evaluate the formula within the scope of that operator at various worlds or times. This last formula is just a quantified sentence or a propositional combination of quantified sentences, and so in evaluating it at a world or time one is concerned to discover just whether some, or all, of the objects which exist at the world or time, satisfy certain non-modal conditions. And this last step can be effected without regard to the identity of the existents at that world or time. Note how this semantic account of the difference between the *de re* and the *de dicto* motivates our decision to count sentences with individual constants within the scope of modal or tense operators as being *de re*. For the interpretation of such sentences is sensitive to the identity of objects at worlds and times: in evaluating '$\Box$(Jones is made of matter)' at the actual world, we have to evaluate 'Jones is

1. This definition of '*de re*' is what Fine calls the 'strict' sense, and is the natural one when individual constants are treated as rigid designators; see Fine [1978a] p. 143.

made of matter' at every world, which requires us to be able to identify Jones at each world.

The distinction is especially clear in the temporal case. If '$\mathbb{F}$(Someone travels to Jupiter)' is true at the present moment, then the sentence 'Someone travels to Jupiter' must be true later, but there are no constraints on the identity of the person whose travelling to Jupiter at a later time makes the quantified sentence true at that time. But if 'Someone will travel to Jupiter' (i.e. 'There is someone who in the future travels to Jupiter') is to be true now, then this requires that at a later time t' some person who exists now travels to Jupiter then, i.e., at t'. So we see that in evaluating *de re* sentences, we rely on facts about transworld or transtemporal identity, facts to the effect that a certain individual at one world or time is identical to a certain individual at another.

Quine's view

Some philosophers, of whom the most influential has been W. V. O. Quine, have argued that *de re* sentences of modal language are problematic in a way in which *de dicto* ones are not.[2] Quine's arguments focus mainly on sentences in which an objectual quantifier binds a variable across a modal operator, sentences which are *de re* in virtue of their satisfying clause (ii) of the definition of '*de re*' on page 47. Put briefly, Quine's view is that (a) the modal operator $\Box$ is best understood as a disguised predicate of sentences, and (b) on such a reading, *de re* sentences are illegitimate, since quantification into expressions enclosed in quotation marks by a quantifier outside those marks does not make sense. Let us explain these two components of his position in turn.

To say that an operator on sentences is a disguised predicate of sentences is to put forward a hypothesis about the 'real' semantic structure of sentences containing the apparent operator. So Quine is saying that such a sentence as

(1) Necessarily, everything is made of matter

is more perspicuously written as

(2) 'Everything is made of matter' is necessarily true.

(2) is a subject-predicate sentence of the simplest sort: it contains a name followed by a predicate. In this case the name is a name of

2. See Quine [1961], paper viii, and Quine [1966], paper 13.

an English sentence, since the effect of putting quotation marks around an expression is to produce a name of that expression; while the predicate is the expression 'is necessarily true'. On our own way of formalizing (1), the adverb 'necessarily' is treated as a sentence operator, but there does not seem to be any very great difference between our method and Quine's.[3] Quine's preference for (2) is based mainly on his preference for truth-functional sentence operators in sentential logic, and as we saw in Chapter 1 (page 3), modal operators are not truth-functional; but let us not pursue the rationale for this preference here.

(1) is a *de dicto* sentence or, more strictly, receives a *de dicto* formalization, on our approach. But if instead we take a *de re* sentence and apply Quine's interpretation of the modal operator, we obtain something which looks decidedly odd. For instance,

(3) $(\exists x)\Box(x$ is made of matter)

becomes

(4) $(\exists x)$('x is made of matter' is necessarily true).

To see why Quine thinks that (4) is nonsense, consider the following assertion, concerning the English word which names the capital city of France, to the effect that it contains five letters:

(5) 'Paris' contains five letters.

To form a name of the word, we surround it with quotation marks, and then we can make a predication of the word. If the quotation marks were deleted from (5) we would still have a subject-predicate sentence, but it would say that Paris (the city) contains five letters, whatever that might mean. Now suppose that we wish to generalize existentially from (5) to assert that something (some word) contains five letters. The correct way to write the result of applying existential generalization is

(6) $(\exists x)(x$ contains five letters).

The name in (5) is replaced by the variable 'x', but the name in (5) is not the word 'Paris' but rather the expression "'Paris'", including the single quotes (we use double-quotes to form a name of this

3. Quine's approach makes the development of modal logic more difficult, but it can be done; see Schweizer [1993].

expression); possible confusion here arises from the fact that in English a quotation-name of an expression contains a display of the expression itself. Suppose, then, that we had fallen prey to confusion, and instead of (6) had written

(7) $(\exists x)$('x' contains five letters).

What does (7) say? It does not say that something contains five letters, since that is the import of (6). The best we can do for (7) is to say that it is composed of a redundant string of symbols, '$(\exists x)$', followed (in redundant parentheses) by a sentence of English,

(8) 'x' contains five letters

which falsely asserts of the third last letter of the English alphabet that it contains five letters. The moral is that by surrounding the variable 'x' with quotation marks we form a name of a letter, and even though the letter 'x' is itself displayed in its quotation-name, it cannot be bound by a quantifier situated outside the quotation marks. However, if we return to (4), we see that such impossible variable-binding is precisely what it attempts. In fact, construing (4) as we construed (7), we have to say that (4) consists in a redundant string of symbols '$(\exists x)$' followed in redundant parentheses by a subject-predicate sentence of English,

(9) 'x is made of matter' is necessarily true.

(9) is presumably false, for the reason that 'is necessarily true' is true only of meaningful sentences, and 'x is made of matter' is not a meaningful sentence, since its subject-expression 'x' has no meaning. So, by his own lights, it is hardly surprising that Quine rejects all *de re* constructions.

But this difficulty for the interpretation of *de re* sentences arises from the supposition that we ought to treat □ as a disguised metalinguistic predicate, a peculiar prescription in view of the fact that there are formally analogous operators which should not be so treated. The sentence

(10) Everything is always made of matter

makes good sense: it is true at a time t iff everything existing at t is made of matter at all times (which most objects are not, since there are times at which they do not exist). Since (10) makes sense, it cannot be rewritten as

(11) $(\forall x)$('x is made of matter' is always true)

since (11) is just as bad as (4) and (7). So the appearance that 'Always' is an operator must be taken at face value, even though it is not truth-functional (for reasons analogous to those underlying the non-truth-functionality of '□'). Why, then, should we balk at treating □ as an operator?

Quine has a reason for distinguishing modal operators from tense operators. As soon as such operators are admitted, *de re* sentences can be formed and, as we have seen, evaluation of a *de re* sentence in tense logic presupposes facts about the transtemporal identity of individuals, while evaluation of *de re* modal sentences presupposes facts about the transworld identity of individuals. Quine's view is that there is such a relation as transtemporal identity, that is, there are real features of things in virtue of which transtemporal identity obtains or fails to obtain across time between individuals with pasts and futures. But the same cannot be said for transworld identity:

> …our cross-moment identification of bodies turned on continuity of displacement, distortion and chemical change. These considerations cannot be extended across worlds, because you can change anything to anything by easy stages through some connecting series of possible worlds (Quine [1976], p. 861).

These remarks embody the crux of Quine's case against *de re* modality, but this case is apparently not a very strong one, at least in so far as it attempts to favor cross-moment identification of bodies. For it is equally true of the temporal case that you can change, if not anything to anything as time passes, at least certain things to startlingly different things, and the philosophical difficulties which arise in virtue of this phenomenon appear to be precisely parallel to those which arise in the modal case. However, we will not go into these matters in detail until Chapter 7; for the moment, let us regard the quoted passage as a challenge, a challenge to give an account of transworld identity at least as good as an account of transtemporal identity which appeals to "continuity of displacement, distortion and chemical change" as criteria for the holding and failing to hold of this relation. What are the modal analogues of these conditions? This is the central question which will be addressed by the later chapters of this book.

For the remainder of this chapter, we will consider what options are available to the philosopher who doubts that an acceptable account of transworld identity is possible, and who therefore doubts the legitimacy of *de re* modal sentences. We shall distinguish three different positions which are motivated by such scepticism about transworld identity:

(I) The first position, Quine's, is one on which *de re* sentences are rejected outright as meaningless. But less nihilistic reactions are possible.

(II) Since the problem is alleged to arise because of a supposed opaqueness in the concept of transworld identity, we could save *de re* modality if we could recast the semantics of quantified S5 so that evaluation of *de re* sentences does not involve a transworld identity relation. Such a recasting, the basic idea of which is due to David Lewis [1968], will be described later in this chapter.

(III) The third position is one on which every *de re* sentence is provided with a *de dicto* equivalent. This may be effected by either of two procedures: we can impose restrictions on the class of admissible S5 models, by adding some further clauses to those on pages 26–27, such that for each *de re* sentence σ there is a matching *de dicto* sentence σ' such that σ' has the same truth-value as σ in each model in the restricted class; or else we can formulate certain principles in modal language from which we can prove that each *de re* sentence has a *de dicto* equivalent. Intuitively, the modal principles would be true in exactly the models counted as admissible by the extra model-theoretic clauses, so they would "select" this class. One can therefore think of such principles as stating something about the abstract structure of modal reality, that feature of its structure which permits *de re* sentences to have determinate truth-conditions without any presupposition that there are determinate facts about transworld identities and non-identities; the idea is that the truth-conditions of any *de re* sentence would be given by any of the *de dicto* sentences equivalent to that *de re* sentence relative to the principles. We investigate this position next.

CHAPTER 3: THE *DE RE*/ *DE DICTO* DISTINCTION

The conditions under which every *de re* sentence has some *de dicto* equivalent are carefully investigated in Fine [1978a]. 'Conditions under which' adverts to a set of sentences embodying some postulates which make the provision of *de dicto* equivalents possible. So we want a set of sentences Σ such that for any *de re* sentence σ there is some *de dicto* sentence σ' such that in S5C we have:

$$\Sigma \vdash (\sigma \leftrightarrow \sigma').$$

Let us say that such a Σ *permits sentence elimination*. Alternatively, we want some model-theoretic conditions such that for each *de re* sentence σ there is some *de dicto* sentence σ' such that in any model satisfying the conditions, σ and σ' have the same truth-value (here 'model' means 'model for quantified S5', as defined on pages 26–27. For example, a simple hypothesis is that every *de re* sentence has a *de dicto* equivalent if the domains of all the worlds are the same. The corresponding Σ could therefore be $\{\Box(\forall x)\Box Ex\}$, example 2.24 on page 35.

This suggestion is technically incorrect, however, and suffers a philosophical flaw as well. Presumably, one who seeks a set of sentences Σ that permits sentence elimination must prefer that the sentences in Σ be themselves *de dicto*, since he regards *de re* sentences as problematic. Unfortunately for such a person, Fine has proved that, in S5C, there is *no* set of *de dicto* sentences which permits sentence elimination. Thus sentence elimination by a *de dicto* set Σ demands that the underlying logic be other than S5C. Fine has further shown that the underlying logic must be a system he calls S5BF, where 'B' is for 'Barcan' (constant domains) and 'F' is for 'flat'. A model is said to be *flat* iff in each world w in W the individuals all have the same non-modal properties.[4] In effect, this shows that the goal of sentence elimination is unachievable, since S5BF is not a system which it would be reasonable to regard as embodying the logic of our operators for broadly logical possibility and necessity.

The need for S5BF arises from the requirement that the sentences in Σ all be *de dicto*. Fine suggests that we relax this requirement. Then if we find a set Σ which permits sentence elimination and which contains *de re* sentences, the *de re* sentences could be regarded as "strictly speaking, meaningless. They are merely stip-

4. See Theorems 27 and 28 of Fine [1978a], pp. 299–301.

ulated to hold in order that the other *de re* sentences may be interpreted by means of their *de dicto* equivalents" ([1978a], pp. 277–8). It is not obvious that this position is tenable but, even if we admit it, the sets of sentences Σ which permit sentence elimination in S5C, our preferred modal logic, are highly unattractive. Fine gives one example of such a Σ, in which there are three sentences, all of them *de re*. We give below the corresponding restrictive conditions on the class of admissible S5C models, since the model-theoretic formulations are easier to grasp. In this example, the conditions permit us to eliminate not just *de re* sentences, but also *de re* formulae with free variables. The three conditions are:

(N) For each world w, there are infinitely many objects in D which do not exist in w.

(P) The extension of an atomic predicate or relation symbol at w is drawn only from $d(w)$.

(H′) The model is *homogeneous:* for any two n-tuples of distinct possible objects drawn from D, if an arbitrarily complex non-modal formula with exactly n free variables is true of one n-tuple at every world, then it is true of the other at every world. Thus, for instance, the necessary properties of any two objects must be the same (Fine [1978a], p. 286).

Although this is only one example of a Σ which permits sentence elimination or rather, an example of the corresponding model-theoretic formulations, it appears to be not uncharacteristic. But it is highly unsatisfactory that merely in order to legitimize *de re* modality one has to embrace such a curious metaphysical thesis as (N); indeed, it is not at all obvious what independent considerations one might bring to bear to decide (N) one way or the other. And (H′) is even less attractive. Consider the formula

(12) ($Ex \rightarrow x$ is not a musical performance).

It might reasonably be held that no human could have been a musical performance and hence that (12) is true of every human at every world. But (12) is false of every musical performance at the actual world and, in general, is true of a musical performance at a world iff that performance does not take place (does not exist) at that world; so musical performances and humans seem to differ as to whether or not (12) expresses a necessary property of them.

However, (H′) forbids such differences; and we can see that (12) is indisputably not a necessary property of musical performances. Thus one who accepts (H′) has to say that for every object *x*, there is a world at which (12) is false of *x*, that is, a world at which *x* exists and is a musical performance. In particular, each actual human is a musical performance at some world at which he exists. This is a *reductio ad absurdum* of (H′). In conclusion, then, the third position we distinguished as motivated by scepticism about the coherence of the notion of transworld identity appears to lead to the postulation of theses which are wildly at variance with our intuitive judgements about what is possible and impossible. The effort to preserve *de re* modality by the method of providing *de re* sentences with *de dicto* interpretations yields poor results.

Counterpart theory

In evaluating a *de re* formula such as □*Fa*, the role of transworld identity is to determine, for each world *w*, which object is relevant to the truth or falsity of *Fa* at *w*. Our evaluation clauses say that *Fa* is true at *w* iff the referent of '*a*' is in the extension of *F* at *w*, so the relevant object is the referent of '*a*', which, of course, is *a*; thus *Fa* is true at *w* iff the object which is identical to *a* is in the extension of *F* at *w*. However, one might abstract this notion of relevance from our particular evaluation method, and experiment with other relations besides identity for fixing which object is relevant to a formula at a given world. The general scheme is that *Fa* is true at a world *w* iff the object relevant to '*a*' in '*Fa*' at *w* is in the extension of *F* at *w* (is in *Ext*(*F*, *w*)); and we have the option to consider other ways of spelling out "relevant to '*a*' in '*Fa*'" besides 'identical to *a*'.

This line of thinking motivates *counterpart theory*, originally proposed in Lewis [1968]. Instead of saying that □*Fa* is true at *w* iff, at each world *u*, the thing identical with *a* (at *u*), i.e. *a*, is in *Ext*(*F*, *u*), we say (roughly) that □*Fa* is true at *w* iff, at each world *u*, anything which is a counterpart of *a* (at *u*) is in *Ext*(*F*, *u*).[5] Note that on this revised account, the qualification 'at *u*' is no longer redundant; only *a* can be identical to *a* at *u*, but perhaps something other than *a* can be *a*'s counterpart at *u*. Of course, this change in terminol-

5. This is phrased with an eye on the version of counterpart theory to be introduced below, and is not how Lewis would put it. In his terminology, □*Fa* is true at *w* iff, for each *u*, anything in *u* which is a counterpart of *a* is in *Ext*(*F*). Only things which are "in" *u* are relevant ('in' means 'exists in') and the extensions of predicates are not relativized to worlds.

ogy is of no help in the present context unless Quine's objection to transworld identity, that there is no account of it analogous to the account of transtemporal identity, is met when we introduce the crossworld relation of counterparthood to play the role originally played by transworld identity in the evaluation of *de re* sentences. But Lewis has an account of counterparthood which, it seems, does meet the objection. He writes: "The counterpart relation is a relation of similarity…Your counterparts …resemble you more closely than do the other things in their world" ([1968], p. 114). So we can give the following criterion:

(C) For x in $d(u)$ and y in $d(v)$, y is a counterpart of x at v only if nothing in v is more similar to x as it is in u than is y as it is in v.

Some comments are in order. First, when $u = v$, we take (C) to imply that x is its own *sole* counterpart. Secondly, it is consistent with (C) that when $u \neq v$, x in $d(u)$ has more than one counterpart in $d(v)$, since two or more objects in v may be equally similar, as they are there, to x as it is in u, although more similar than all the other objects in $d(v)$. Thirdly, note that the criterion states only a necessary condition for counterparthood between existents at two worlds. Should the condition also be sufficient? The problem is that on any given resolution of similarity, there will always be at least one thing in a world v at least as similar as is anything else in v to x as x is in u. So if the condition were sufficient, every object would have at least one counterpart in every world. However, it is plausible that in some worlds all of the things which exist are so dissimilar from a as it actually is that it is difficult to allow even the most similar of these to count as 'representatives' of a at that world, i.e. as a's counterparts there. So we do not want the similarity condition to be sufficient.

The other feature of (C) to remark is that it concerns only existents at worlds. To see the point of this restriction, consider the sentence $\Box Ea$. If a is a contingent existent, this sentence should be false, so we need to be able to construct models at whose actual world it is false. In the orthodox semantics, such a model is one with a world w at which a does not exist ($a \notin d(w)$), and the natural translation of this idea into counterpart-theoretic terminology is that a is a contingent existent in a model iff there is some world in the model *at* which none of the things which are a's counterparts

there exist (this is the interpretation of the more general thought, in the terminology of two paragraphs back, that none of the things relevant to '*a*' in '*Ea*' at that world should exist at it). Since (C) concerns only existents, it is consistent with an object's having a non-existent counterpart at a world, and thus leaves room for whatever stipulation we may wish to make to effect this.

These remarks motivate the following reformulation of the criterion of counterparthood into two parts, which together give us a fuller account of a counterparthood relation based on similarity. First, we impose the stipulation that the domains of worlds are *disjoint*, which mean that if there is some x which belongs both to $d(u)$ and to $d(v)$, then $u = v$. Then we say:

(C_1) For any object x in $d(u)$, if x has a counterpart in $d(v)$ (i.e., a counterpart which exists at v), then for all y, y is a counterpart of x at v iff y is in $d(v)$ and nothing in $d(v)$ is more similar to x as it is at u than is y as it is at v.

Again, we take (C_1) to imply that x is its own sole counterpart at u. To complete the account, we must now deal with the case where x has no counterpart in $d(v)$, i.e., no counterpart which exists at v. The simplest stipulation is that an object is its own sole counterpart at a world at which it has no existing counterpart:

(C_2) For any object x in $d(u)$ and any distinct world v, x has no counterpart in $d(v)$ iff x has exactly one counterpart at v and that counterpart is x itself.

We shall need to make a number of other stipulations about the counterpart relation, but we have already said enough to draw attention to a few points. First, (C_1) and (C_2) only work together given the stipulation that the domains of worlds are disjoint: if x could be in both $d(u)$ and $d(v)$, by (C_1) it would be its own sole counterpart at both worlds, contradicting (C_2). Second, it should be observed that our counterpart relation is a three-place relation, the relation 'b is a counterpart of a at w', which we write as *Cbaw*; and according to (C_2), this relation can hold even if b does not exist at w, provided $b = a$. Third, (C_2) strengthens our earlier statement of what it means in counterpart theory for an object x to be a contingent existent; in our first formulation, we said this means that at some world v, none of the counterparts of x at the world exist at v. But by (C_2) there is only one such counterpart, the very

same object; so contingent existence means that at some world the counterpart at that world does not exist at the world. Fourth, such a counterpart *x has* counterparts at other worlds, but they will be the objects determined to be *x*'s counterparts by the way *x* is at the world in which it exists (otherwise we would need a four-place counterpart relation, '*x* at *u* is a counterpart of *y* at *v*').

Let us now turn to the details of the model theory. We want to give essentially the same semantics for quantified S5 as we gave in Chapter 2, pages 26–27, except that *de re* sentences will be evaluated using a counterpart relation between objects, and the domains of worlds will be disjoint. An important constraint we impose on the model theory we are about to construct is that it should contain the standard model theory as a special case. When this constraint is met, the counterpart-theoretic approach becomes a generalization of the standard approach in a precise sense. First, for every standard S5 model we can construct a counterpart-theoretic equivalent,[6] in effect embedding standard models in the class of counterpart-theoretic ones. Second, we can then expand the class of counterpart-theoretic models to include models with no standard-semantics equivalent, by allowing the counterpart relation to have different formal properties from those of identity, so that, for instance, an object can have more than one counterpart at a world, two objects can share a counterpart at a world, or counterparthood can be, say, non-symmetric.[7]

If this is as much generalization as we countenance, certain features of the standard model theory should carry forward. In particular, the treatment of quantifiers should be actualist, $\Box$ should express strong necessity, and the underlying non-modal first-order logic should be that of free logic for the empty domain. With this in mind, we define a counterpart-theoretic model for quantified S5 for the language *L* (CTS5 *L*-model, for short) to consist in the following seven components:

6. Given a standard model $\mathbf{M} = \langle W, D, d, Ext, Ref, w^\star \rangle$ we define a counterpart-theoretic model $\mathbf{M}'$ by $W' = W$, $D' = \{\langle a,w\rangle\colon a \in D,\ w \in W\}$, $d'(w) = \{\langle a,w\rangle\colon a \in d(w)\}$, $Ext'(F,w) = \{\langle\langle a_1,w\rangle,\ldots,\langle a_n,w\rangle\rangle\colon \langle a_1,\ldots,a_n\rangle \in Ext(F,w)\}$, $Ref'(\underline{a}) = \langle a, w^\star\rangle$, $C = \{\langle\langle a,u\rangle, \langle a,w\rangle, u\rangle\colon a \in d(w) \cap d(u),\ w, u \in W\} \cup \{\langle\langle a,w\rangle, \langle a,w\rangle, u\rangle\colon a \in d(w),\ a \notin d(u),\ w, u \in W\}$. Then for every $w \in W$, the same sentences hold at w in $\mathbf{M}$ as hold at w in $\mathbf{M}'$.

7. In this context, symmetry is understood as meaning that for all x, u, y, v, if $x \in d(u)$ and $Cyxv$, then $Cxyu$ (symmetry is guaranteed if $y \notin d(v)$).

(i) a non-empty set W of possible worlds;
(ii) a non-empty set D of possible objects;
(iii) a function d which assigns to each world w in W a subset $d(w)$ of D; d is subject to the constraints that every x in D is in some $d(w)$ and that if $u \neq v$ then $d(u)$ and $d(v)$ are disjoint, i.e., $d(u) \cap d(v) = \varnothing$;
(iv) for each object x in D and for each world w in W, a specification of which objects in D are x's counterparts at w; this specification is subject to the constraint that if x is in $d(u)$, then for all other worlds v, x is a counterpart of x at v iff no y in $d(v)$ is a counterpart of x at v; furthermore, in these circumstances x is the sole counterpart of x at v; and, finally, if x is in $d(u)$, then x is the sole counterpart of x at u;
(v) for every n-place predicate F of L and for each world w, a specification of which n-tuples of objects drawn from D are in $Ext(F, w)$;
(vi) for each individual constant in L, an assignment to it of a referent from D (again, we assume that L has a name for every member of D);
(vii) a selection of a particular $w \in W$ as the actual world $w^\star$.

In the usual set-theoretic way of thinking of relations, a counterpart relation conforming to (iv) is a set of triples of objects, each triple containing members of D in its first two places and a member of W in its third. If we let C stand for this relation, then 'b is a counterpart of a at w' means that $\langle b,a,w\rangle$ is a member of C.

It remains to give the evaluation clauses for the operators in counterpart theory. Here there is a complication with no analogue in the standard semantics, for the counterpart relation is only relevant to the evaluation of *de re* sentences, and then only at a particular stage of the evaluation. For example, compare sentences (13), (14) and (15):

(13) $\Diamond(\exists x)Fx$,
(14) $\Diamond Fa$,
(15) $\Diamond\Diamond Fa$.

The truth-value of (13) at a world w does not depend on the counterpart relation at any stage of its evaluation ((13) is *de dicto*) but only upon whether $(\exists x)Fx$ is true at some world. The truth-value

of (14) at a world w depends immediately on the counterpart relation since (14) is true at w iff for some u, some counterpart of a at u is F at u. But the truth-value of (15) does not depend *immediately* on the counterpart relation: (15) is true at w iff for some u, $\Diamond Fa$ is true at u, but it is only after unpacking this latter condition that the counterpart relation is invoked.[8] So the modal operator $\Diamond$ is going to require a clause with two cases. Say that an occurrence of an individual constant t in a formula φ is *immediately* within the scope of a modal operator μ in φ iff t is within the scope of μ in φ and there is no modal operator μ' in φ such that t is within the scope of μ' and μ' is within the scope of μ. For instance, the first, but not the second, occurrence of 'a' is immediately within the scope of the initial $\Diamond$ in $\Diamond(Fa \,\&\, \Diamond{\sim}Fa)$. Then as the examples indicate, we need to distinguish the case when a modal operator has individual constants immediately within its scope from the case where it does not.

Suppressing the obvious for the sentential connectives, the evaluation clauses are:

(viii) an atomic sentence of the form $\pi(t_1,\ldots,t_n)$ is true at a world w iff $\langle Ref(t_1),\ldots,Ref(t_n)\rangle$ is a member of $Ext(\pi, w)$;

(ix) an identity sentence $t = t'$ is true at a world w iff $Ref(t) = Ref(t')$;

(x) $(\exists v)\varphi v$ (respectively, $(\forall v)\varphi v$) is true at w iff for some (respectively, all) $a \in d(w)$, $\varphi[\underline{a}/v]$ is true at w;

(xi-a) in $\Diamond\varphi(t_1,\ldots,t_n)$ let $t_1,\ldots,t_n$ be exactly the occurrences of individual constants that are immediately within the scope of the displayed $\Diamond$; then $\Diamond\varphi(t_1,\ldots,t_n)$ is true at w iff there is a world u and for each i a counterpart c_i of $Ref(t_i)$ at u such that $\varphi(\underline{c}_1,\ldots,\underline{c}_n)$ is true at u; here $\varphi(\underline{c}_1,\ldots,\underline{c}_n)$ is the sentence obtained from $\varphi(t_1,\ldots,t_n)$ by substituting the name $\underline{c}_i$ of the counterpart c_i of $Ref(t_i)$ at u;

(xi-b) if $\Diamond\varphi$ contains no occurrences of individual constants that are immediately within the scope of the displayed $\Diamond$ then

8. The alternative is to require that for $\Diamond\Diamond F\underline{a}$ to be true at w, $\Diamond F\underline{b}$ must be true at some u, for b a counterpart of a at u, which means in turn that $F\underline{c}$ must be true at some v, for c a counterpart of b at v. But the condition for $\Diamond F\underline{a}$ to be true at w is that $F\underline{c}$ must be true at some v, for c a counterpart of a at v. If counterparts are not 'inherited', we can have $\langle c,b,v\rangle \in C$ while $\langle c,a,v\rangle \notin C$, resulting in a counterexample to $\Diamond\Diamond F\underline{a} \models \Diamond F\underline{a}$. So the underlying propositional modal logic is forced to be non-S5-ish.

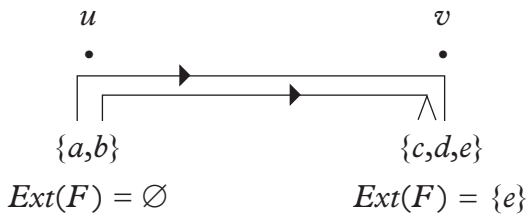
$\Diamond\varphi$ is true at w iff there is some world u such that φ is true at u;

(xii-a) in $\Box\varphi(t_1,\ldots,t_n)$ let $t_1,\ldots,t_n$ be exactly the occurrences of individual constants that are immediately within the scope of the displayed $\Box$; then $\Box\varphi(t_1,\ldots,t_n)$ is true at w iff for every world u, and for every sequence of objects $c_1,\ldots,c_n$ such that for each i, c_i is a counterpart of $Ref(t_i)$ at u, $\varphi(\underline{c}_1,\ldots,\underline{c}_n)$ is true at u;

(xii-b) if $\Box\varphi$ contains no occurrences of individual constants that are immediately within the scope of the displayed $\Box$ then $\Box\varphi$ is true at w iff φ is true at every world u.

Here is a simple illustration of this semantics, in which we give a counterexample to

(16) $\Diamond(\exists x)Fx \rightarrow (\exists x)\Diamond Fx$.



In this model, we have $W = \{u,v\}$, $D = \{a,b,c,d,e\}$, $d(u) = \{a,b\}$, $d(v) = \{c,d,e\}$, $w^\star = u$, and $C = \{\langle d,a,v\rangle, \langle c,b,v\rangle, \langle d,b,v\rangle\}$ (so b has two counterparts at v – the arrows point from an object to its counterparts in the domains of the other worlds).[9] $\Diamond(\exists x)Fx$ is true at $w^\star$ because $(\exists x)Fx$ is true at v, in turn because $e \in d(v)$ and $F\underline{e}$ is true at v (see clause (x) above). But $(\exists x)\Diamond Fx$ is false at $w^\star$ because its truth would require, by clause (x), the truth of either $\Diamond F\underline{a}$ or $\Diamond F\underline{b}$. However, at neither u nor v is there a counterpart of a which belongs to $Ext(F)$, so by clause (xi-a), $\Diamond F\underline{a}$ is false at $w^\star$; similarly, $\Diamond F\underline{b}$ is false at $w^\star$.

The similarity analysis of counterparthood suggests that the branching illustrated in our picture is justifiable. The use to which counterpart theory will be put later in this book does not depend

9. We suppress 'reflexive' arrows indicating that a is its own counterpart at u, etc., and also arrows indicating a non-existent counterpart – the absence of arrows from c, d and e into $d(u)$ automatically requires that c is c's counterpart at u, etc. Thus symmetry as defined in note 7 fails in this model.

on the similarity analysis (which I am not endorsing) but the alternative account I will develop is also one that justifies branching. Since branching is a major departure from the standard semantics, there is no reason to expect that the system of validities delivered by the counterpart-theoretic apparatus just set up will be the same system as is obtained on the orthodox semantics. We shall continue to call this latter system 'quantified S5', and the set of formulae valid according to the counterpart-theoretic model theory will be referred to as 'CTS5'.

Objections to counterpart theory

The version of counterpart theory just outlined improves on Lewis's own version and thus avoids certain technical objections to his approach.[10] However, some technical and some non-technical objections still remain. A non-technical objection is that it may seem that certain object-language modal sentences which are intuitively true will come out false according to criterion (C_1) on page 57; thus, it seems reasonable that the Englishman Jones' life could have been very different from the life he has actually led (say, if his parents had emigrated to Australia in his youth) while other actual individuals, in the same possible situation, could have lead lives quite similar to Jones' actual life. But in a world which realizes these states of affairs, (C_1) will permit only the latter individuals to be Jones' counterparts there, and so the modal judgement about Jones will be false at the actual world: there is no world in which some counterpart of Jones leads a life very different to Jones's actual life while other actual individuals have counterparts leading lives rather like Jones's actual life. Of course, it is possible to reply that in criterion (C_1) we are using the notion of similarity in some technical way, in which it does not mean simply overall similarity in obvious respects. But if the counterpart theorist makes this reply, then it is incumbent upon him to explain exactly what the technical sense of similarity is in which, in our example, the individual in Australia in the imagined world is more similar to Jones as he actually is than any of those other individuals who lead lives in the imagined world which are very similar, in the non-technical sense,

10. Lewis uses a two-place counterpart relation and a binary relation *In* (see note 5 on page 55) which conflates satisfaction of a predicate at a world and existence in the world. As a result, his semantics fails to accommodate contingent existence; see Lewis [1968], p. 119.

to Jones' actual life. At this point, a dilemma arises for the counterpart theorist; for if such an elucidation of the technical sense of similarity cannot be given, then the motivation for counterpart theory, that it employs a crossworld relation less problematic than transworld identity, is undercut; while if such an elucidation can be given, then unless it entails that counterparthood is not a one-one equivalence relation, the same elucidation could presumably be applied to transworld identity, which eliminates the motive for developing counterpart theory. We will return to this dilemma in Chapter 7.

A less telling objection to counterpart theory, due to Kripke and Plantinga, is that it misrepresents the content of ordinary modal judgements. On Kripke's view, the counterpart theorist holds that if we say 'Humphrey could have won the election', then

> we are not talking about something that might have happened to Humphrey but to someone else, a 'counterpart'. Probably, however, Humphrey could not care less whether someone else, no matter how much resembling him, would have been victorious in another possible world (Kripke [1972], p. 334, n. 13).

But, as Hazen has forcefully pointed out ([1979], pp. 319–325), this objection, and similar ones due to Plantinga ([1974], pp. 115–116), are unfair. According to the counterpart-theoretic semantics developed in the previous section, the sentence

(17) Humphrey could have won the election

has the truth-condition expressed by

(18) For some w, some counterpart of Humphrey at w wins the election at w – in symbols, $(\exists w)(\exists x)(\mathrm{C}xhw \,\&\, Wxw)$.

(18) is as good a candidate for being "about" Humphrey as any sentence of the orthodox semantics – after all, it mentions counterparts of *Humphrey*. To be sure, there is no mention of counterparts at all in (17), but there is nothing about worlds in (17) either, although worlds are quantified over in (17)'s orthodox possible worlds truth condition; so it is not only counterpart theory that introduces some new "ideology". However, in the quoted passage, it is not the truth-condition which counterpart theory does ascribe to (17) which Kripke criticizes, but rather one it does not:

(19) Some counterpart of Humphrey could have won the election at some world (in symbols, perhaps something like this: $(\exists x)[Cxhw^\star \,\&\, \Diamond Wx]$).

(19) is not a well-formed sentence of either the modal object language or the counterpart theorist's metalanguage, since it contains both the three-place metalanguage predicate of counterparthood and the modal object-language operator '$\Diamond$'. Certainly, (19) does not represent the content of (17), but the counterpart theorist does not say it does. The Kripke-Plantinga objection misses the target.[11]

The standard technical objection to counterpart-theoretic semantics concerns its handling of the logic of identity. In quantified S5 (the orthodox semantics)

(20) $a = b \rightarrow \Box(a = b)$

is valid. It is tempting to explain why as follows: if the antecedent of (20) is true at $w^\star$, then Ref(a) = Ref(b), and since the reference of constants is the same at every world, Ref(a) = Ref(b) at every world, and so '$\Box(a = b)$' must also be true at $w^\star$. Indeed, (20) would be valid even if we allowed the reference of a constant to change from world to world, provided co-designating constants at one world co-designate at every world. But even the treatment of constants as rigid designators is in fact not sufficient to guarantee the validity of (20), for if we could make sense of the idea that one

11. The trouble with the quoted passage is that it tries to use a correct modal intuition – whether or not someone else could have won the election is not what matters to whether or not Humphrey could have won it – to refute a possible- worlds *analysis* of what the possibility that Humphrey wins it consists in. But if we want to contradict the analysis, we must either phrase it in modal operator terms to get a logical comparison with (17), or phrase (17) in possible worlds terms to get a logical comparison with (18). But, according to the counterpart theorist, (18) *is* the possible worlds phrasing of (17), so no contradiction will be forthcoming. The Kripke-Plantinga objection seems merely to be a protest at replacing the standard possible worlds semantics with the counterpart-theoretic one. Similarly, someone might protest at an analysis of tense operators as quantifiers over times if it also replaced ordinary continuants with an ontology of instantaneous individuals. But it is not an *objection* to such a semantics simply to observe that the truth-value of 'Humphrey will win the election' (where 'Humphrey' denotes the present stage of a certain sequence of person-stages) turns on what happens in the future to 'someone else', if the 'someone else' is the right entity, a future stage in the same sequence of person-stages. Kripke's attitude to counterpart theory is hard to fathom, since in n. 18 of [1972] he recommends its use to solve a problem known as Chisholm's Paradox, a recommendation we follow in Chapter 7.

object could have been two, then at some world which realizes this possibility, '*a*' could perhaps attach to one of the two objects and '*b*' to the other, so that $a = b$, though true at $w^\star$, is false at a world in which the single object of $w^\star$ is two. But it is clear that on the orthodox semantics with transworld identity, this possibility cannot be represented, and we can see why without reference to the treatment of constants. Note that on the standard semantics we also have the validity of

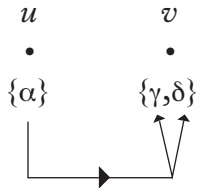
(21) $(\forall x)(\forall y)(x = y \rightarrow \Box(x = y))$,

which says that if x and y are the same at $w^\star$, then they are the same at every world, and again is just a consequence of the fact that one object does not 'become' two or more objects at other worlds. Moreover, this is true for any possible x and any possible y; if we pick an object at one world for x and the same object at another world for y, then at any world, x and y are the same. So the following strengthening of (21) is also valid:

(22) $\Box(\forall x)\Box(\forall y)\Box(x = y \rightarrow \Box(x = y))$.

Kripke has argued that these formulae are *intuitively* valid, so that there would be something wrong with a semantics on which they have counterexamples. Consider the example of Hesperus and Phosphorus, otherwise known as Venus. Although it is *a posteriori* that Hesperus is identical to Phosphorus (the names were originally associated with different appearances of Venus, and it was a scientific discovery that these were appearances of a single planet) it is surely *a priori* that if these planets are the same, then necessarily they are the same. It is not denied that we can conceive of a world in which 'Hesperus' is a name of one planet and 'Phosphorus' a name of another, but this is not a world in which *Hesperus* (Venus) and *Phosphorus* (Venus) are different planets. Since Hesperus, Phosphorus, and Venus are all the same planet, a world in which Hesperus is a planet distinct from Phosphorus is a world in which Venus is two planets. So the intuitive validity of formulae (20)–(22) turns ultimately on the intuition that one thing cannot be two (formulae (20)–(22) are versions of a thesis sometimes known as the *Necessity of Identity*).

As we already remarked, it is consistent with (C_1) that a as it is in u has two existent counterparts b and c at v; this would arise if b



and *c* are similar enough to *a* to be counterparts of *a* at *v*, if there is no choosing between them in terms of this degree of similarity, and if they are more similar to *a* is it is in *u* than is any other object in *v*. But a model in which this situation is realized is essentially a model of one thing's having the possibility of being two. The picture is in the margin. Suppose that '*a*' and '*b*' are both names of α; then $a = b$ is true at $w^\star$ (= *u*). But $\Box(a = b)$ is false at $w^\star$, since by clause (xii-a) on page 61, its truth would require that at every world *w*, any counterpart *x* of α at *w* and any counterpart *y* of α at *w* are identical, while at *v* we have two different counterparts of α; or putting the same point more precisely in the terms of (xii-a), $\Box(a = b)$ is false at $w^\star$ because at *v*, one of the identity sentences containing names of counterparts of Ref(*a*) at *v* is false. So we have obtained a counterexample to (20), a counterexample which arises because the counterpart relation need not be one-one. Unsurprisingly, readers will find that if the counterpart relation is allowed to depart from the formal properties of identity in other respects, such as transitivity, further counterexamples to orthodox S5 validities involving identity can be obtained. There will be more examples later.

It may be suggested that the remedy to this problem is simply to stipulate, in the clauses of the counterpart-theoretic model theory, that the counterpart relation is a one-one equivalence relation. But structural stipulations unmotivated by the elucidation of the nature of counterparthood, such an elucidation as (C_1), depart from the *raison d'être* of the semantics, which, to repeat, is to provide a model theory which can deal with *de re* sentences without appeal to the allegedly problematic relation of transworld identity. A structural stipulation which goes beyond what is entailed by the elucidation simply imports an unelucidated component into the content of the relation, and Quine's challenge applies again; and there is certainly no case to be made that (C_1) by itself entails that counterparthood is a one-one equivalence relation.

A defender of counterpart theory may therefore choose to query the correctness of formulae (20)–(22) themselves. But this has the appearance of a desperate measure, since, as Kripke has pointed out, there is a powerful argument for these formulae ([1971], pp. 135–141). The formula

(23) $\Box(\forall x)\Box(x = x)$

is a validity of quantified S5. But if we combine (23) with Leibniz's Law, according to which, if x and y are the same, they have the same properties, we can deduce (20). We can embody Leibniz's Law for a and b in a schema all of whose instances are valid:

(24) $a = b \rightarrow (\varphi(v)[a/v] \leftrightarrow \varphi(v)[b/v])$

where $\varphi(v)$ takes arbitrarily complex object-language predicates with one free variable as substitution instances. If we substitute for $\varphi(v)$ any expression which stands for a genuine property of individuals we obtain a valid (not merely true) instance of (24). The property of being necessarily identical to a is a genuine property of individuals (or else first-order modal logic is not worth doing) and is expressed by any one-place predicate of the form $\Box(a = v)$. Substituting in (24) yields

(25) $a = b \rightarrow (\Box(a = a) \leftrightarrow \Box(a = b))$

and (20) follows from (25) *via* (23) (if there is some validity which φ implies to be equivalent to ψ, then φ implies ψ). So a counterpart theorist who proposes to reject (20) must find fault with this argument, which means he must reject (23) or Leibniz's Law. But (23) seems unobjectionable, and Leibniz's Law irresistible.

The upshot of our discussion is this. CTS5 is not the same system as quantified S5, and the difference arises because the counterpart relation, if given a "similarity"-style elucidation like (C_l), has structural properties inconsistent with those of identity. The counterpart theorist must therefore either find a better elucidation of the counterpart relation, or give reasons why the validities of quantified S5 he rejects should in fact be rejected; and it seems that this second option in turn involves giving reasons why (23) or Leibniz's Law should be rejected. We shall leave matters there for the moment, but in Chapter 7 counterpart theory will be taken up again, and we shall see one rather plausible way in which counterpart theory might pursue its case.

CHAPTER 3: THE *DE RE*/ *DE DICTO* DISTINCTION

Chapter 4
Metaphysics for the Semantics

Semantics and explanation

IN THIS chapter, we begin the investigation of an assortment of philosophical problems which arise in connection with modality. We will subsequently be concerned mainly with the philosophical justification of a variety of modal theories or theses whose meaning will be assumed to be well-understood. But in view of the material of the previous chapters, the first philosophical issues to demand our attention are issues about the semantics of the modal operators themselves. These issues are sufficiently general not to turn on which of the various approaches already presented we adopt, so we restrict our attention here to the orthodox possible worlds semantics for S5, without accessibility.

We have regarded possible worlds semantics as a tool for fixing the powers of the logical operators, for determining which modal logical arguments are valid and which invalid. We saw that there is not always a unique answer to the question 'Valid or invalid?', for there are some arguments, involving iterated modalities, about which we perhaps have no very firm intuitions, which are valid in some systems and invalid in others. But there are also central cases, which any semantics has to get right if it is to be taken seriously; for instance, there is example (A) from Chapter 1 (page 4), repeated here in the margin.

(A)

$$\frac{\Diamond P \quad \Diamond Q}{\Diamond(P \,\&\, Q)}$$

However, it would be misleading to suggest that the authority of possible worlds semantics derives merely from its getting the cases about which we do have firm intuitions right, its dictates about the peripheral cases being a matter of indifference. There are algebraic approaches to questions of validity which will also do that (Hughes and Cresswell [1968], Goldblatt [1993]), but, in comparison with

these, there is undeniably a sense in which possible worlds semantics is the 'natural' semantics. But in what, precisely, does this naturalness consist? A tempting reply to this question is that the naturalness arises out of the treatment of the modal operators as quantifiers over possible worlds: there must be a sense in which this treatment is the correct treatment. It is in virtue of this that we can say that the possible worlds semantics does not merely pronounce that (A) is invalid, it also *explains why* it is invalid: we understand what is wrong with (A) when we are introduced to the existential quantifier treatment of ◊, which a engages our prior understanding of what is wrong with example (B).

(B)

$$\frac{(\exists w)Pw \;\; (\exists w)Qw}{(\exists w)(Pw \;\&\; Qw)}$$

The philosophical problem for this view is to elucidate the sense in which the quantifier treatment is correct, in such a way that the invalidity of (A) is explained by relating it to (B).

The most obvious suggestion about the sense in which the quantifier treatment is right is the suggestion that in translating a sentence such as

(1) $\Diamond P$

by the principles employed in Chapter I into the sentence

(2) $(\exists w)Pw$

we are translating one sentence into another with the same meaning (recall that we read (2) as 'there is some world w such that P holds at w'). Then the relationship between (A) and (B) which permits the invalidity of (B) to explain the invalidity of (A) would just be that the sentences in (A) mean the same as their translations in (B). So the quantifier treatment is right because it maps sentences into synonyms (since this is intended to be a substantial claim, we shall refer to the possible worlds translations of modal sentences simply as their renderings, which is a more neutral term than 'translation' *vis à vis* the question of synonymy). However, the correctness of the quantifier treatment cannot consist just in its preserving meaning: there must be an asymmetric element in this synonymy relationship, otherwise we would not be able to say that the invalidity of (B) *explains* the invalidity of (A). ◊ is the mysterious operator, the one whose logical powers are being investigated, while the existential quantifier is already understood.

To capture this idea of asymmetry, let us say that (2) 'articulates' or 'exhibits' the 'real' meaning of (1). On this view, the sig-

nificance of the quantifier treatment of modal operators is akin to the significance which philosophers have generally attributed to regimentations of 'problematic' sentences in standard first-order logic. A classic example is that of Russell's Theory of Definite Descriptions. Definite descriptions, phrases of the form 'the *F*' appear to be terms for referring to objects, much like proper names. For various reasons' Russell wished sharply to distinguish descriptions from genuine names and so proposed an analysis on which such a sentence as:

(3) The author of this book is Scottish

is said to have the 'real' meaning

(4) There is exactly one author of this book and he is Scottish.[1]

According to Russell, (3) has the 'surface' structure of a subject-predicate sentence in which the subject term is 'the author of this book', i.e., the sentence has the same structure as 'Graeme is Scottish'. If we use the symbol 'ι' to form definite descriptions, reading $(\iota x)Fx$ as 'the x which is such that Fx', or, simply, 'the F', then (3) would be formalized according to its surface structure as:

(5) $S[(\iota x)Ax]$

a sentence with the same form as Sg, while (4), of course, becomes:

(6) $(\exists x)(Ax \,\&\, (\forall y)(Ay \rightarrow x = y) \,\&\, Sx)$.

Thus the surface structure of (5) is misleading as to its 'real' meaning, which is exhibited by (6). In particular, (6) shows that (5) does not really contain a subject term (and hence (4) shows the same about (3)) since there are only quantifiers, predicates and connectives in (6); this is the result which Russell wanted. Furthermore, by attributing the logical form of (6) to (5), we can explain the logical powers of the operator ι from which definite descriptions are formed. For instance,

(7) Someone authored this book

1. See Russell [1918] Lecture VI for an original source, and Neale [1990] for recent work of descriptions.

appears to follow from (3), intuitively speaking, and if (3) means what (4) means, we have an explanation of why the inference is valid. The suggestion is, then, that possible worlds semantics explains validity and invalidity in the same way; that is, the relationship between (1) and (2) is the same as that between (5) and (6).

Before looking at some of the consequences of this view, let us pause to ask if there is any relation between (1) and (2) weaker than synonymy which could do the same job. Our minimum requirement is that any candidate R be such that when modal sentences and possible-worlds sentences stand in R, then a modal argument and its R-corresponding possible worlds argument are either both valid or both invalid. More precisely, we insist on a relation which meets this condition: if σ is a sentence of modal language and σ' its rendering in possible worlds language according to the quantifier treatment of modal operators, then the hypothesis that σ stands in the candidate relation to σ' should be sufficient to guarantee that σ behaves the same way in a modal argument as σ' does in the rendering of that argument in possible worlds language. Thus, for instance, we are asking whether there is some relation other than synonymy such that if $\Diamond P$ and $(\exists w)Pw$ stand in it, then it follows that the behavior of $\Diamond P$ in the argument (A) displayed above is the same as the behavior of $(\exists w)Pw$ in (B). The idea of 'same behavior' here is still rather intuitive (we will make it more precise later), but the sense of the question is clear enough to see that there is one rather trivial answer to it: we can simply *define* a relation of behavioral equivalence as that relation which two such sentences stand in when they behave in the same way in pairs of corresponding arguments like (A) and (B). Thus, if our possible worlds renderings of modal sentences map sentences onto behavioral equivalents, it is trivially true that a modal argument is valid iff its corresponding possible worlds argument is valid. But it is quite clear that the fact that this relation holds between modal sentences and their possible worlds renderings does not ground the ability of possible worlds semantics to *explain* the validity and invalidity of modal arguments. For it would be an equally substantial question *why* our method of producing possible-worlds renderings *via* the quantifier treatment of the modal operators yields behaviorally equivalent sentences; and again, the answer which strongly suggests itself is that a possible worlds rendering of a modal sentence σ is synonymous with σ. If this answer is incorrect,

then the semantical systems of the earlier chapters may have done no more than engender an illusion of understanding.[2]

So the view at which we have arrived is that possible worlds semantics explains validity and invalidity because (a) the quantifier treatment of the modal operators produces synonyms, and (b) a possible worlds rendering of a modal sentence exhibits the real meaning of that sentence (the synonymy relation has a preferred direction). But this view has a disturbing feature, in that the quantifier treatment is ontologically radical: it introduces entities of a certain sort, possible worlds, which are apparently not introduced by modal sentences themselves.

At this point, the analogy with the Theory of Descriptions breaks down, for Russell's motivation for that theory was a certain kind of ontological conservatism. Briefly, Russell held that a subject-predicate sentence would be meaningless if its subject term did not succeed in picking out some object. Now, if one holds that sentences with the structure (5),

(5) $S[(\iota x)Ax]$

are genuine subject-predicate sentences, then when faced with such a sentence as 'the present King of France is bald', the Russellian must either say that the sentence is meaningless, which flies in the face of the facts, or that there is such an entity as the present King of France, which also appears to fly in the face of the facts. But one could grasp the second horn of the dilemma and say that there is such an entity as the present King of France, a non-existent entity, and that there are non-existent objects generally. This would be an ontologically radical move. Another way, preferred by Russell since he did not wish to introduce non-existent objects, is to prevent the argument to the dilemma from getting started, by denying that 'the present King of France is bald' is really a subject-predicate sentence

In our case, we are moving from the ontologically conservative (1), '$\Diamond P$', to the ontologically radical (2), '$(\exists w)Pw$', although we could say that the explicit ontological commitment in (2) is at least implicit in (1). Furthermore, there does not appear to be any way of avoiding this commitment, if possible worlds semantics *explains* validity and invalidity for modal arguments. (2) says that *there is* a

2. The phrase is Quine's. For further discussion, see Scott [1971].

possible world of a certain sort; if this is not literally true while (1) is literally true, then the invalidity of the argument (B) is irrelevant to the question of whether (A) is valid or invalid, since sentences in (B) do not mean what sentences in (A) mean, on the view of explanation of meaning we are presently canvassing. So let us accept the extra ontology apparent in possible worlds discourse; let us agree that well-formed instances of (2) are literally true or literally false, and that there are some literally true instances, since there are some literally true instances of (1). That is to say, we agree that there are possible worlds. We shall say that by this agreement, we are *realists* about possible worlds, since we take them to be real things.

We may distinguish *absolute* realism from *reductive* realism. An absolute realist is one who holds that the notion of a possible world cannot be further analysed; so (2) is as far as we can go in exhibiting the content of (1) in better understood terms. David Lewis is the absolute realist *sans pareil*, but his position includes two extra ingredients which are not essential to absolute realism. First, Lewis holds that each possible world is a thing of the same kind as the actual world, and second, that physicalism is true of the actual world. Hence every possible world is a physical system, isolated in a special way from every other (Lewis 1986, pp. 69–86). But other views are conceivable, on which possible worlds are some kind of *sui generis* abstract object; and an exception might, or might not, be made of the actual world (Davies 1981, p. 200).

A reductive realist is one who holds that possible worlds can be identified with (constructions out of) other entities, themselves held to be less problematic than worlds. Three such positions are that worlds are maximal consistent sets of propositions (or propositions of a certain sort), that they are maximal states of affairs, and that they are maximal possibilities.[3] It follows from this that the absolute/reductive distinction is not the same as another common in the literature, between *actualism* and *possibilism*. An actualist identifies the realm of existence with what is actual, while a possi-

3. That worlds are sets of propositions is argued in Adams [1981]; that they are propositions of a certain sort in Prior and Fine [1976, pp. 116–178]; that they are maximal states of affairs in Plantinga [1974, pp. 44–5]; and that they are maximal possibilities in Humberstone [1981]. An idea which has not been worked out in the literature is that worlds are mental constructs of some sort (McGinn [1981]); this would be a version of reductive realism about incomplete possibilities.

bilist holds that in a more fundamental sense of 'exists', non-actual objects ('merely possible' objects) are among the existents. (For a temporal parallel, consider the contrast between the view that only presently existing things exist, or presently existing things plus things which existed previously, and the view that in a more fundamental sense of 'exists', things which are still to be are among the existents.) Reductive realists are usually actualists – actualism is part of the motivation for the reduction – and Lewis is a possibilist. But the combinations of absolute realism with actualism and reductive realism with possibilism are conceivable.

Either variety of realism about possible worlds is of course opposed to anti-realism about them.[4] An anti-realist says that worlds do not exist, and so is an actualist of a more radical kind than any reductive realist. For an anti-realist, any possible worlds sentence which has an existential quantifier over worlds as its main connective must be strictly and literally false; assuming the non-existence of worlds to be necessary, this formulation is correct even if a possibilist existential quantifier is admitted (recall that the possibilist quantifier Σ has the clause that '$(\Pi v)\varphi v$ is true at a world w iff for every a in D, $\varphi[\underline{a}/v]$ is true at w', i.e., there is no restriction to those a in $d(w)$. So the anti-realist cannot take the attitude towards possible worlds semantics outlined above: he cannot say that possible worlds sentences exhibit the real meanings of modal sentences in a peculiarly perspicuous way. The most interesting philosophical question about the semantics of modal logic is whether it is possible to develop an anti-realist view that is consistent with our intuition of naturalness in the quantifier treatment of the modal operators, and which can deal with the thought that the invalidity of (A) is somehow explained by the invalidity of (B). But

4. I am not using the label 'anti-realism' in its contemporary sense to mean a position which denies that the meanings of undecidable sentences are their (*ipso facto* verification-transcendent) truth-conditions. For an explanation of this kind of anti-realism, see Dummett[1975b]. My use is the old-fashioned one, in which an anti-realist about *F*'s is one who denies the existence of *F*'s. That these two notions of anti-realism are connected is argued in Dummett [1981]; see especially pp. 66–69, where Dummett characterizes a view such as ours (according to which there really are no such things as worlds)as springing 'from the perception of a genuine and important fact…that we do not need to invoke the notion of reference, as applied to such terms [for possible worlds], in order to explain how a sentence containing such a term is determined as true or false…an understanding of those statements [possible-worlds statements] involves an implicit grasp of their relation to statements of the reductive class [modal statements]'.

if some kind of realism were quite satisfactory, this question would be purely hypothetical; to give it some practical urgency, then, let us see whether there are reasons to have qualms about realism.

Realism about worlds

A very natural consideration in favor of absolute realism about worlds arises from semantic parallels between tense and modal operators. In tensed languages, according to clauses (ℙ) and (ℙ) on page 38, we have operators not obviously of quantificational form which manifest themselves in the surface structure of English as tenses of verbs; we also have explicitly quantificational expressions such as 'sometimes' and 'always'; and the tenses are treated as relativizing the semantic values of the expressions on which they operate to entities (times) over which the quantificational expressions also range. In the modal case, we treat the subjunctive mood in surface English analogously, where explicitly quantificational expressions such as 'in all possible circumstances' function in a manner similar to that of 'sometimes' and 'always'. These syntactic parallels might also be extended to include expressions for places; although nothing corresponds to tense or mood, there are quantificational expressions like 'everywhere' and names of places like names of times (dates). Moreover, in English there is a variety of spatial and temporal indexicals like 'here', 'there', 'then' and 'now', whose reference in a particular utterance is determined by the place or time at which the utterance is made. Someone impressed with the parallel drawn so far may then press it further by suggesting that 'actually' plays a similar indexical role, its reference in an utterance being the world of utterance.[5] Thus there exist the materials for the view that realism about worlds is as well motivated as realism about places and times. Just as we can speak of places and times forming their own manifolds or spaces, so we can say that worlds are the points of a logical space.

There can be no objection to the introduction of such a metaphor, but it supports realism about worlds (granted realism about places and times) only if the similarities upon which the metaphor relies for its appropriateness relate features of logical space to features of space and time manifolds which themselves are inconsis-

5. The treatment of 'actually' as a context-dependent operator is developed in Lewis [1970] and integrated into a general theory of context-dependence in Kaplan [1989]. For further discussion, see Adams [1984], Davies [1983], and Forbes [1983].

tent with anti-realism about places and instants. However, crucial features of places and times which appear to underpin the plausibility of realism about these entities have no parallel in the logical space of possible worlds. For places and times, there is a distinction between the item and its occupier, a material object in the case of a place, and an event in the case of a time. (Strictly, it is regions and intervals which are occupied by objects and events, but places and instants can be 'abstracted, from these; and it is reasonable to hold that possibilities and refinement correspond to regions and abstraction.) It seems crucial to our ability to distinguish places and times from their occupiers that we have the conception of the same place, or time, being occupiable by something distinct from its actual occupier; even in the temporal case, someone who denies that a particular token event e which occurs at a time t could have occurred at another time, will not deny that events could have occupied t other than those which do occupy it. However, this conception is quite inapplicable to logical space: given a world, one cannot distinguish a location and a content contingently located there, no matter which component one identifies with the world itself. How does this difference arise?

It is apparently sufficient for the distinction between location and occupier to be applicable that there be some contingent relational structure amongst the occupiers which either determines or is determined by the locations of the occupiers (for a relationalist, a type of reductive realist about space, these relations determine the locations of objects, while for an absolutist, the converse is true; see further Forbes [1987a]). To see that relations weaker than determination may not be sufficient, consider the case of color space (Aleksandrov *et al.*, [1983], Vol. III, pp. 151–3). Any color can be regarded as a combination of red, green, and blue in specific intensities with numerical values x, y, and z respectively. Thus any color can be given coordinates $\langle x, y, z\rangle$ with respect to the three axes red, green, and blue. Moreover, we can define distance on this three-space in such a way as to reflect real phenomena of color perception. If we say that a *threshold of distinction* for a given color is the amount of continuous alteration needed before a human being perceives a change in color, then the distance between two colors can be identified with the smallest threshold of distinction which can be laid between them. This distance relation is contingent, since humans could have had better or worse powers of sen-

sory discrimination. Nevertheless, we do not seem to be able to abstract a color space from the colors which fill it, even though in some sense this space has actually unoccupied regions, such as Hume's missing shade of blue. The problem is that the distance relations and the coordinates of the colors are quite independent: there is no inclination at all to think that if the distance relations had been different, a different color would have had the coordinates actually possessed by, say, the color of the jerseys of the Tulane Green Wave.

There are no natural relations on possible worlds corresponding to distance relations, and although relations could be introduced, such as the relation '*w* is more similar to *u* than is *v*' (for a fixed method of resolving respects of similarity), there is nothing contingent about such relations. Hence there is no means by which we might distinguish a possible world from what is true at it: the content of a world is in no sense something which occupies the world. And this means that the appropriateness of the metaphor of logical space does not reside in similarities which motivate equal degrees of realism about places and times, on the one hand, and worlds, on the other. For our ability to separate a place, or a time, from its occupier, is crucial to realism about places and times, be it absolute or reductive, as is the applicability of a distance relation to the places and times themselves. The means by which the conceptual separation is effected is by holding the distance relation between the points constant while changing the distance relation between the occupiers (by moving them around, or by deletion with or without replacement). This procedure attributes necessity to the facts about distances between points themselves, which gives them identity criteria, and therefore 'objecthood', independent of that of the category of occupiers. But this apparatus is not available for worlds.[6]

However, a reductive realist who is also an actualist will find this limitation of the metaphor of logical space not particularly dismaying, since on this view, realism about worlds is derivative from realism about the entities from which worlds are held to be constructed. A more general objection has to be pressed against such

6. In McGinn [1981] it is argued that this failure of analogy constitutes a reason to reject the reality of possible worlds. This paper contains an interesting defense of anti-realism about worlds conjoined with realism about modal reality.

a realist, and one who is unpersuaded by the objection we are about to produce will find the anti-realism which avoids it concomitantly unmotivated. The main objection against both sorts of realism about worlds is the nominalist-actualist objection from epistemology. According to this objection, knowledge of properties of objects requires experience of these objects or of their effects, which in turn requires that these objects or effects be within the range of our sensory faculties. But only objects which are both concrete and actual are, or can be involved in producing effects which are, within the range of these faculties. However, the realist holds that the expression $\Diamond P$ attributes a property, that of P's holding, to an object, a world, which is non-actual according to the absolute or possibilist-reductive realist, and non-concrete according to the actualist reductive realist. Hence realism renders it generally impossible to know whether or not $\Diamond P$ is true.

The strength of this objection depends upon the plausibility of nominalism and actualism, but the problem of how it is possible that we have knowledge of propositions from certain areas of discourse is a powerful consideration in favor of these positions.[7] Since we cannot conduct a general discussion of the issues here, let it suffice to say that whatever force the epistemological objection has is reflected by a corresponding urgency in the development of an adequate anti-realist view of possible worlds.

Two problems for anti-realism

The challenge for the anti-realist is to give an interpretation of the appealing features of possible worlds semantics which shows how these features can arise even though there are no such things as worlds; he cannot just ignore the semantics, given the intuitions we have about its naturalness, for this is a phenomenon which surely requires explanation. Furthermore, his interpretation must posit some semantic relationship between sentences of modal language, which we shall call L_m, and their renderings in possible worlds language, which we shall call L_w; for without such a relationship, it must seem positively miraculous that the semantics agrees with our intuitions about validity and invalidity. And we saw earlier that the obvious candidate for this relationship is synonymy. It follows that the component of the realist view about the semantics which the anti-realist must attack is the claim that the synonymy relation

7. See further Benacerraf [1965] and Field [1980], pp. 1–19.

is asymmetric in the direction which makes the possible worlds sentences stand to modal sentences as Russell's interpretations of the sentences of a language with the operator ι stand to the sentences of that language: the anti-realist has to allow synonymy, but deny that the possible worlds renderings exhibit the real meanings of the modal sentences. In particular, since he is an anti-realist about worlds, he has to say that objectual quantifiers, when they range over possible worlds, do not have their literal meaning, the meaning they have in ordinary first-order languages; in turn, then, the sentences of possible worlds language do not mean what they appear literally to mean.

What, then, is the meaning of such sentences? The simplest maneuver available to the anti-realist here is to reverse the direction of the asymmetry in the synonymy relationship. Instead of saying that the meaning of a modal sentence is given by its L_w rendering, we can say that the meaning of an L_w-sentence is given by its rendering in (reverse translation into) L_m; so in the simplest case we say that (2):

(2) $(\exists w)Pw$

should be understood as having the meaning (1) has,

(1) $\Diamond P$

or that (2) has a meaning imputed to it by (1). Since there is no literal assertion of the existence of possible worlds in (1), it follows that there is no literal assertion of the existence of worlds in (2), despite appearances. However, there are at least two pressing problems for this anti-realist position. The first is the problem of validity. If L_w-sentences have non-literal meaning, there must be an element of the incidental in the methods of possible worlds semantics for determining validity and invalidity: there has to be a more direct method. Furthermore, whatever this method is, one has to be able to derive from it an explanation of why possible worlds semantics is successful. We have said that the invalidity of (B) in some sense *explains* the invalidity of (A), its modal counterpart, and in deriving or explaining the invalidity of the latter from that of the former, we are assuming, by anti-realist lights, that however the meaning of the sentences in (B) differs from the apparent meaning they have, the meaning which results from interpreting their quantifiers literally, this difference does not change the logic

(B)

$$\frac{(\exists w)Pw \;\; (\exists w)Qw}{(\exists w)(Pw \;\&\; Qw)}$$

of the quantifiers in these non-literal occurrences. The fundamental account of validity for L_m must justify this assumption; and if it can succeed in doing this, we can account for the intuition that the semantical status of (A) is illuminated by its rendering as (B) simply in terms of the great familiarity of first-order languages.

The second problem which faces the anti-realist doctrine we are investigating is the problem of reverse translation. In proposing that each possible worlds sentence be ascribed the meaning of the modal sentence of which it is a rendering, we are proposing an elimination by paraphrase of the ontology of possible worlds, as opposed to a reductive identification of worlds with other entities. But this elimination is possible only if every meaningful possible worlds sentence is a rendering of some modal sentence, and as L_m and L_w presently stand, this is not so. In fact, much of the difficulty lies in an expressive weakness in L_m, a weakness which the anti-realist must show how to remedy to make his position plausible. We deal with these problems in turn in the next two sections.

Validity: other approaches

To justify attribution of normal logic to quantifiers when they bind world variables in L_w-sentences, we have to show that the valid/invalid classification consequent upon this attribution is in agreement with the classification delivered by criteria which apply directly to L_m-sentences, without detour through non-literal renderings of them. So, first, we have to find criteria of this sort; and there are two types of criteria we might hope to develop, criteria from proof theory and criteria from alternative semantics. Let us begin with proof-theoretic criteria.

In developing possible worlds semantics in Chapters 1 and 2, we assumed a fund of intuitions about the correctness or incorrectness of particular arguments. What precisely is the source of these intuitions? One possible account is that competent speakers of English have native intuitions about what follows from what in their language, and when presented with the argument-schemata of formal logics, intuitions about the particular connectives occurring in these arguments are isolated and activated, so that it is just obvious to such a speaker, at least in simple cases, whether or not the conclusion of any English instance of the schema would follow from its premises. According to this view, on which the bedrock 'pretheoretic' intuitions are intuitions about what follows from what, the fundamental method of encapsulating the meaning of a

logical constant is to give a rule for when a sentence with that constant as its main connective follows from other sentences, and also a rule which determines what other sentences follow from it. So this view looks to the 'natural deduction' rules governing a connective for the embodiment of the essence of a native speaker's mastery of the connective.[8]

To give a simple illustration, consider the connective &. The typical semantic account of validity for propositional languages presupposes that the meaning of & is fixed is some semantic way, usually by a truth-table or fundamental truth-value matrix. On the present view, the meaning of & should rather be given by an introduction and elimination rule for the connective, thus:

&–Int: if A has been proved from premises X and B has been proved from premises Y then A & B follows from premises $X \cup Y$;

&-Elim: if A & B has been proved from premises Z, then A follows from Z and B follows from Z.

More formally, the rules may be written:

&I: if $X \vdash A, Y \vdash B$, then $X \cup Y \vdash A$ & B.

&E: if Z ⊢ A&B, then $Z \vdash A$ and $Z \vdash B$.

In terms of these rules, we can justify the truth-table for & in the light of a view about the truth-predicate associated with Quine. According to Quine, although use of the truth-predicate involves 'semantic ascent', so that instead of making a statement about the world we predicate a property of a sentence,

> …the truth predicate serves, as it were, to point through the sentence to reality…Thus ascent to a linguistic plane of reference is only a momentary retreat from the world, for the utility of the truth predicate is precisely the cancellation of linguistic reference…The truth predicate is a device of disquotation (Quine [1970], pp. 11–12).

Suppose we put P & Q (strictly, '{P & Q}') for Z in &-E. Then, since P & $Q \vdash P$ & Q, we infer by &E) that P & $Q \vdash P$ and P & $Q \vdash Q$. Ascending to a linguistic plane of reference, we conclude that

8. The main proponent of this view was Gentzen; see Gentzen [1969]. The chief modern development of Gentzen's approach is Prawitz [1968]. See Peacocke [1987] for a sophisticated philosophical account of logical constanthood in this tradition.

the truth of P follows from the truth of P & Q, as does the truth of Q, and this gives us the three F entries in the truth-table for &. Similarly, putting P for X and Q for Y in &I, we can conclude that the truth of P & Q follows from the two-premise set $Z = \{P$ is true, Q is true$\}$, using descent followed by ascent; so we obtain the T entry in the table.

On this view, the semantics of a connective is answerable to its rules of proof. So what function does the semantics play, if it is the inference rules which are fundamental? We can say that the semantics provides a tool for establishing that an argument-schema is incorrect, where, in this context, 'incorrect' means 'not establishable by the rules of proof'. That a schema is incorrect, in this sense, if there is a counterexample in the semantic sense, is established by a soundness proof for the semantics, and that every incorrect schema has a semantic counterexample is established by the completeness proof.

But according to Michael Dummett, the position on which the semantics is answerable to the rules of proof rather than conversely

> …obliterates the distinction between a semantic notion of logical consequence…and a merely algebraic one…Semantic notions are framed in terms of concepts which are taken to have a direct relation to the use which is made of the sentences of the language…algebraic notions define a valuation as a purely mathematical object…which has no intrinsic connection with the use of sentences…It is quite impossible that it should be an utter illusion that semantic accounts of the logical constants supply an explanation of their meanings, and that such accounts have no more significance than a purely algebraic characterization of a logical system (Dummett [1978], p. 295).

However, we have the materials at hand to rebut this objection, at least in the case of our example, the connective &: a distinction between semantic and algebraic clauses for & may be said to be manifested in the fact that the semantic account of & follows from the rules of proof together with the Quinean manipulations of the truth-predicate and the classical assumptions that each sentence is either true or false and not both. So the semantic account is intimately related to the use which is made of sentences in the language, even if we think of the fundamental facts about the use of connectives as being recorded in their rules of proof.

The proof-theoretic approach already runs into problems with negation, if it is classical negation whose truth-table we want to

derive. But how does it apply to the modal operators? To simplify matters, let us suppose ◊ to be introduced by definition, and concentrate on □. The elimination rule for □ is straightforward:

□E: if $X \vdash \Box A$, then $X \vdash A$.

However, it is more difficult to say what the introduction rule should be. Intuitively, the idea is that if A follows from a set of sentences all of which are necessary, then A is itself necessary. But what does 'necessary' mean here? An obvious suggestion is that it means 'has □ as its main connective', which gives the rule:

□I: if $X \vdash A$ and every sentence in X has □ as its main connective, then $X \vdash \Box A$.

One application of □I establishes that $\Box\Box P$ follows from $\Box P$, but in fact the system defined by the two rules above is just S4, so we cannot use the rules to establish the S5 thesis that $\Box\sim\Box\sim P$ ($\Box\Diamond P$) follows from $\sim\Box\sim P$. However, if we decide to count $\sim\Box\sim P$ as a necessary sentence, on the grounds that every occurrence of a sentential letter in it is *within the scope* of a □, then the inference goes through and we obtain S5; that is, S5 is the system defined by the rule □E as above, and

□I*: if $X \vdash A$ and for every sentence S in X, each sentence letter occurrence in S is within the scope of some occurrence of □, then $X \vdash \Box A$.

We can also obtain quantified S5 by replacing 'sentence' with 'atomic predicate or relation symbol' in □I*. Of course, it is not an objection to the proof-theoretic approach that it delivers different systems on different construals of 'necessary sentence', since possible worlds semantics also delivers different systems. Perhaps, indeed, the proof-theoretic approach has the advantage, delivering a narrower range of systems, which might on that account be regarded as the natural ones.

For modal operators, however, it is harder to meet Dummett's point that the proof-theoretic approach renders the possible worlds semantics indistinguishable from algebraic semantics, and thus fails to explain our intuition of naturalness in the former. For we cannot relate the possible worlds semantics to the use which is made of modal sentences by an argument using no more than the resources needed to derive the truth-table for & from its deduction

rules: such manipulations with the truth predicate will not take us from expressions containing modal operators to expressions containing quantifiers over possible worlds. However, we can instead appeal to the anti-realist thesis about possible worlds sentences, that the meaning of these sentences is the meaning which belongs to the modal sentences of which they are renderings; then we can employ Quine's principles to move from (8) to (9):

(8) $\Box P$ is true

(9) $\Box$(*P* is true)

and then apply the anti-realist thesis to (9) to obtain a sentence which has the meaning which (9) has,

(10) In every possible world, *P* is true.[9]

There is no plausible claim which could substitute for the anti-realist thesis here which would permit the production of a clause from an algebraic definition of validity as an equally natural competitor to (10), so again, the natural semantics can be related to our use of modal sentences in a way that others cannot be.

A technical problem for the proof-theoretic approach is to show that possible worlds semantics agrees with it about the validity and invalidity of modal arguments. To do this, we should prove that any modal argument is correct according to the inference rules (say, for S5) iff its translation into possible worlds language is valid according to ordinary first-order semantics.[10] This result establishes that even though world quantifiers in L_w do not mean what

9. It may seem that (9) does not follow from (8) because a sentence which in fact expresses a necessary truth could have been used to express a contingent truth. However, following Peacocke [1978], pp. 477–8, we can regard the predicate 'true' as short for 'true-in-*L*' for a fixed language *L*, and then stipulate that the semantic properties of a language are the same at every world. Thus if $\Box P$ is true-in-*L* there is no world where *P* expresses a contingent truth *and* is a sentence of *L*.

10. There is a minor technical complication here, since we have not been translating modal sentences into ordinary first-order sentences, but rather into sentences of a *two-sorted* first-order language. In a model for such a language, two domains are specified, one for the range of variables of the first sort and the other for that of the second. Each *n*-place predicate of the language is also sortally characterized by place – each place is stipulated to be occupiable by a particular sort of variable. An atomic sentence formed from such a predicate is true only if it has names of objects from the appropriate domain in the appropriate places. For further details, see Enderton [1972], pp. 277–86.

quantifiers usually mean in first-order language, the difference is not sufficient to prevent us from employing our familiarity with first-order logic in assessing modal arguments. And, in fact, the argument which is needed here is fairly straightforward.[11] However, there is a more substantial philosophical difficulty in the way of the present approach. The thought that the rules of inference embody the meaning of a connective is supported by the idea that the rules explain what operation on meanings the new connective performs, so that if it is introduced into an already understood language, then one can straight away understand sentences with one occurrence of the new connective, since one grasps the operation (knowing the rules) and also grasps the meanings operated upon (since these are from the already-understood language in this case); and one's understanding of sentences with more than one occurrence of the new connective is built up from there. But inspection of the introduction rules for $\Box$ given above reveals that they do not help in the step from a $\Box$-free language to sentences with one $\Box$, since they only specify how to reason with $\Box$ in a language to whose lexicon it already belongs. This suggests that the rules do not embody an operation on meanings of the required kind.

Perhaps this difficulty can be overcome recursively. But at this point, it is simpler to introduce another approach to which an anti-realist might turn for his primary account of what it is for a modal argument to be valid. This second approach, which involves alternative semantics rather than proof theory, can be motivated by considering how we went about engaging intuitions about correctness and incorrectness of formal inference schemata such as example (A) on page 4. Our procedure there was to choose particular English substitution-instances of the given schema such that, for an invalid schema, the possibility of the premises of the instance being true while its conclusion is false, is evidently demonstrated. This method gives rise to a conception of validity for formal languages which is the main semantic rival to the usual model-theo-

11. In one direction the result is obvious, since by inspection of the rules $\Box$E and $\Box$I* on page 83, we can see that they correspond to valid $\forall$-inferences in a first-order language with one world-variable (which is treated semantically as a name of $w^\star$ when it is not bound). For the converse, a conceivable difficulty is that for some set of L_m-sentences $\{X,A\}$, X a set, we have $\mathrm{T}(X) \vdash \mathrm{T}(A)$ but every first-order proof of this sequent has a line which is not the translation of any L_m-sentence. But it is not difficult to prove that this does not happen.

retic approaches, the *substitutional* conception (Quine [1970], pp. 49–56). Applying it to the modal case, the basic idea would be that a schema of quantified S5 is valid iff uniform substitution of expressions from an interpreted language with *n* free variables for the *n*-place atomic relation symbols of the schema, and names for the individual constants, always yields arguments whose premises cannot all be true if their conclusion is false. As it stands, this notion of validity is implicitly relativized to the language from which the substituends are chosen; at present, we have in mind a regimented fragment of a natural language. Unfortunately, if the fragment is sufficiently 'weak', the wrong results will be obtained; for instance, if we restrict ourselves to arithmetical expressions and names of numbers, then we will validate the inference of $\Box P$ from P, since all truths of arithmetic are necessary truths. A better definition is therefore that a schema is valid for L iff it is not possible that there is some extension of L from which substituends can be chosen in such a way that it is possible for the premises of the resulting instance to be true while its conclusion is false (Peacocke [1981], pp. 137–8).

Several nice questions arise about how the modal substitutional account is to be understood. It is no objection to it that it uses modal operators to define validity for modal languages, unless it is also objectionable that quantifiers are used to define validity for quantificational languages. But since the definition speaks of what is possible and not possible for extensions of the language in question, a complete account would have to address itself to such topics as the existence-conditions of languages: does a language exist at a world only if it is the actual language of some population at that world?[12] And, as with the previous case, we would like to show that a modal argument is valid by the substitutional account iff its translation into possible worlds discourse is valid by ordinary first-order semantics. However,to show this we would need soundness and completeness proofs for the modal inference system, where soundness and completeness are also conceived of substitutionally. Soundness, the more important property, is easy to establish, but completeness is highly problematic.[13]

12. There is a detailed treatment of this and other questions about the substitutional account in the doctoral dissertation of S. G. Williams (Oxford University 1984).

13. Here I am indebted to Williams, *op. cit.*

In sum, then, the anti-realist is not without resources to explain what it is for modal arguments to be valid or invalid, even though possible worlds semantics is rejected as giving the fundamental account of these concepts for modal systems.

The meanings of possible worlds sentences

The anti-realist thesis that the meanings of L_w-sentences are imputed to them one by one by the inverse of the translation scheme which carries L_m-sentences into L_w-sentences figured in the previous discussion, as providing the anti-realist with a reply to Dummett's objection. But the thesis itself requires some defense. The problem is that there are L_w-sentences which are apparently meaningful but which are not reverse-translated by any L_m-sentences; nor can these sentences be eliminated merely by deleting some vocabulary from L_w, since some of the problematic L_w-sentences employ only vocabulary which appears in L_w-sentences which do translate L_m-sentences, so that deletion would result in these L_m-sentences having no possible worlds rendering. As an example of such a problematic L_w-sentence, Hazen ([1976], p. 38) has given:

(11) $(\forall w)(\exists x)(Exw \ \& \ Exw^{\star})$.

(11) says that in every possible world there is some object which also exists in the actual world, and cannot be reverse-translated into L_m because no expression of that language as it presently stands can have the force of '*x* actually exists' if that phrase is within the scope of a $\Box$ or $\Diamond$.

However, (11) has a perfectly natural English rendering, 'Necessarily, some actual object exists', which does not contain any of the vocabulary of possible worlds. So what the example shows is that L_m as it stands does not have all the resources required to formalize English modal discourse. Hazen suggests that L_m be supplemented with an 'actuality' operator, written 'A', to obtain an expanded language (which we shall still call L_m) in which we can express (11), i.e., in which we can formalize 'Necessarily' something actual exists,:

(12) $\Box(\exists x)\mathsf{A}Ex$.

A similar example involves the L_w-sentence:

(13) $(\exists w)(\exists x)(Exw \ \& \ {\sim}Exw^{\star})$.

(13) says that in some world there is something which does not exist in the actual world, which in modal English is 'there could have been things other than there actually are', a truth which underlies some of our counterexamples to Barcan and converse Barcan formulae on pp. 25–26. To express this with the new operator, we write:

(14) $\Diamond(\exists x)\sim \mathsf{A}(Ex)$.

So far, the introduction of the actuality operator has merely been a syntactic maneuver. To ensure that (12) and (14) do have the readings (11) and (13) respectively, we need to give an evaluation clause for the new operator in possible worlds semantics. But there is no mystery about what the clause should be:

(i) a formula of the form $\mathsf{A}(\varphi)$ holds at a world w in a model M iff φ holds at the actual world $w^\star$ of M.

The reader may confirm, by applying the relevant evaluation clauses to (12) and (14), that they do have the possible worlds imports stated in (11) and (13).[14]

However, having added A to L_m, Hazen is still able to find problematic L_w-sentences. For instance, if instead of a constant for the actual world in (11) we have another quantified world variable,

(15) $(\exists u)(\forall w)(\exists x)(Exw \ \& \ Exu)$

then our simple 'actually' operator is of no avail. It is tempting to render (15) in modal English as 'it could have been that necessarily, something is actual', but this sentence has at least one interpretation on which it is simply equivalent to 'necessarily, something is actual'. To obtain the effect of (15), we have to make 'actual' refer back to the state of affairs introduced by an evaluation of the initial 'it could have been that': we wish to speak of what would have been actual if that state of affairs had obtained. It is possible to hear 'it could have been that necessarily, something is actual' as involving such a back reference, and in that case we would be justified in introducing further operators into L_m to express this reading. Such operators, *indexed* 'actually' operators, have been devised by Christopher Peacocke (Peacocke [1978], pp. 485–7). The idea is to attach a numerical index to the $\Diamond$ or $\Box$ which we wish the later

14. For more on 'actually', see Davies and Humberstone [1980].

occurrence of 'actual' to pick up' and then to index the corresponding 'actually' operator with the same number. So

(16) $\Diamond_1\Box(\exists x)\mathsf{A}_1 Ex$

is the L_m-sentence which (15) translates.

In possible worlds terms, indexed 'actually' operators are required when one wishes to assert that things meeting a certain condition at one world also meet some condition at some perhaps distinct world; the problem which the indices solve is that of expressing the condition met at the first world in such a way that in evaluating the condition at another world, one is taken back to that first world. Obviously, then, one cannot give an evaluation clause for indexed operators in the straightforward manner of (i) above, since in evaluating a formula one needs some method of keeping track of which worlds are introduced for which indexed □'s and ◇'s as one proceeds. However, the details are not too complicated (cf. the postscript to this chapter, page 104 ff). Later, we shall see that these indexed operators are indispensable in the formulation of some plausible modal theses about entities of various sorts; and on page 54 we already have a thesis which needs them for its exact expression, the thesis that there are infinitely many non-existents at each world.[15]

Further generalization of the basic actually, operator A can be motivated by consideration of such a sentence as

(17) My car (*a*) could have been the same color as yours (*b*) actually is.[16]

In possible worlds terminology, (17) says that in some world, my car has the same color as the color which yours has in the actual world; (17) thus involves a crossworld comparison of the color of two objects. One could give a formalization of (17) by quantifying over colors:

(18) $(\exists C)(\Diamond(a \text{ is colored } C) \,\&\, b \text{ is colored } C)$.

15. The formula we want is in a modal language with the appropriate operators which is based on the infinitary language $L_{\omega_1\omega_1}$. We write

$\Box_1\exists\Diamond\{x_i\colon i \in N\}[\wedge(\sim\mathsf{A}_1 Ex_i \,\&\, \wedge\{x_i \neq x_j\colon i < j, i, j \in N\}]$.

Here $\exists\Diamond X$ abbreviates an infinite sequence of alternated ◇'s and ∃'s, one for each variable in X, the sequence beginning with ◇.

16. The example is due to Peacocke.

But, to quote Lewis's complaint, "that's not how the English does it" ([1986], p. 13). To arrive at a formal representation of the mechanism embodied in the English, we use a two-place predicate C such that 'Cab' means 'a is same-colored with b'. In possible worlds language, we wish to compare a in a world u with b in a distinct world v, so that we can say that a in u is same-colored with b in v.[17] That is, we require C to be a four-place predicate in L_w; then (17) has the L_w-regimentation:

(19) $(\exists w)Cawbw^{\star}$.

To achieve the effect of (19) in L_m, one needs an 'actually' operator which can associate particular objects with the actual world, independently of the order in which the objects are referred to in the sentence. So the 'actually' operator which is to associate a particular object with the actual world should be indexed with the name of that object. By allowing such term indexing in L_m we can express (19) by

(20) $\Diamond \mathsf{A}^{b}_{x} Cax$.[18]

A semantics for a language with such operators may be found in the postscript to this chapter (page 104 ff).

We also have little difficulty in understanding a weaker version of (17), in which 'actually is' is replaced by 'could have been', as involving a comparison between two worlds neither of which is actual. This requires a combination of numerical and term indices, as in

(21) $\Diamond_1 \Diamond \mathsf{A}_1 \binom{b}{x} Cax$.

Should the anti-realist object to these new operators, on the grounds that they are really nothing but devices for disguised quantification over worlds? The objection appears doubtful, for each successive step in introducing the operators was motivated by the production of an English sentence which required, or had a reading which required, the operator introduced at that step, if we are to do things 'the way the English does it'. For the English examples no more give the appearance of quantification over worlds than do modal sentences of English formalizable in the

17. For more on crossworld relations, see Salmon [1981], Ch. 13.

18. The idea and notation are due to Peacocke.

original version of L_m, without any of the new operators, where the anti-realist denies that there is quantification over worlds. Thus it is unclear why anyone should think that the new operators import the unwanted ontology.

By introducing the various types of 'actually' operator, the anti-realist can assign interpretations to a much wider range of L_w-sentences than is possible with L_m as it first stood. Nevertheless, there are still some L_w -sentences which are uninterpreted, and which will remain so no matter how we might extend the techniques of the previous paragraphs. These are the sentences which are, intuitively, 'about' worlds, such sentences as $(\forall w)(w = w)$, 'every world is identical to itself'. Now it is possible for the anti-realist to rule such sentences out of L_w by restrictions on the methods by which well formed formulae can be built up,[19] but the fact remains that in formulating possible worlds model theory, one makes stipulations which, if formalized in first-order language, would be L_w-sentences with no L_m interpretation. So given that such sentences are used, what attitude should the anti-realist take to them? The most appealing suggestion is that he should regard them in the same way as certain mathematical sentences were regarded by Hilbert, that is, as instruments or uninterpreted stipulations which enable us to establish facts of interest about the interpreted sentences.[20] For the anti-realist, the justification for the use of such sentences, that is, for the making of such stipulations, is that a semantical theory conforming to them is in agreement over questions of validity, with the fundamental account of validity in L_m, whichever account he prefers; the function of the uninterpreted sentences is just to ensure that this agreement obtains.

Finally, we should recall that the anti-realist treatment of L_w-sentences was originally premised upon the thought that 'Possibly, *P*, and 'There are some possible circumstances in which *P*, are synonymous, the meaning of the former fixing that of the latter. But in our discussion of anti-realism, we have replaced possible circum-

19. For example, rather than regard our two-sorted L_m as a notational convenience for a single-sorted language with predicates for being a world and being an individual (so that $(\forall w)(w = w)$ is regarded as an abbreviation of $(\forall x)(Wx \rightarrow x = x)$) we can think of the sort distinctions as built into L_w. Then if we stipulate that = is a predicate of sort $<D,D>$ only, $(\forall w)(w = w)$ is not even well-formed.

20. See Dummett [1978], p. 219, Smorynski [1977], pp. 822–5, and Field [1980] for discussion of this kind of 'instrumentalism'.

stances by possible worlds, and it is a fair complaint that this detracts from the plausibility of the synonymy claim. For possible worlds are complete ways things might have been, and there is apparently nothing in the meaning of 'Possibly, *P*' which corresponds to this element of completeness. However, in Chapters 1 and 2 we developed the materials for an anti-realist response to this objection, for we showed there how to give a semantics for L_m which preserves the quantifier treatment of $\Box$ and $\Diamond$ but does without the completeness assumption. The premise upon which the anti-realist treatment really relies might be better expressed as postulating a synonymy between 'Possibly, *P*' and 'There is some possibility that *P*', where, again, the latter is a mere paraphrase of the former. We already know that possibility semantics is equivalent to possible worlds semantics, and so all the apparatus of this and the previous section can be transferred to that system.[21] Possible worlds are complete possibilities, that is, possibilities which have no proper refinements, and we can express this condition in the language of possibility semantics. Again, the resulting sentence will have no L_m interpretation, since it is 'about' possibilities, but the stipulation that every possibility conform to the condition is an example of an uninterpreted sentence which is of considerable instrumental value, since it allows us to replace the clause for negation which is the cause of the complexities in possibility semantics and obtain the familiar possible worlds semantics as a result.

So we may conclude that anti-realism about possible worlds does not demand the wholesale rejection of the semantical methods of the earlier chapters, and thus continue to use these methods so long as we find them useful, without having to believe that there literally are such things as worlds.

21. For reverse-translating L_w into L_m, see the Appendix. In reverse-translating a sentence about possibilities, when a possibility quantifier ($\exists\rho$) is eliminated, so are all occurrences of $\sim(\exists\tau \Rightarrow \rho)$, $\sim$ replaces this expression in front of the complete subformula within the latter's scope, and τ is eliminated from this subformula.

Chapter 5
A Modal Theory: The Essences of Sets

Essential properties and essences

THE study of non-modal predicate calculus is often labelled 'mathematical logic', a title which adverts to one motivation for the development of modern logic, which was to codify the canons of inference employed by mathematicians in constructing and developing mathematical theories. By analogy, one might expect to find in modal logic the canons of inference employed in the construction and development of modal theories. However, modal theories themselves, at least for the entities which will be of interest to us, are in a somewhat underdeveloped state in comparison with mathematical theories;[1] we will therefore be studying individual modal theses rather than full theories. Moreover, the subject-matter of these theses, typically philosophical, is not often amenable to illuminating formal articulation, so we will not be treating the theses as axioms and trying to prove theorems from them. The role of modal logic is more to make the theses absolutely precise than to facilitate the deduction of substantial consequences from them; it also helps provide a framework for clear discussion of the grounds for accepting or rejecting these theses. It is this activity which we will pursue in this and the next three chapters, with a view to meeting Quine's challenge to explain transworld identity. For this challenge does not evaporate upon adoption of anti-realism about

1. One exception is modal set theory, for which see Fine [1981b]. There is another sense in which the modal *logics* we surveyed in Chapter 1 might be regarded as theories. The logic determined by the class of general models, known as K, could be taken to be the basic logic, and other (adequately axiomatizable) logics thought of as theories expressing in modal terms the fact that accessibility has such-and-such a structure. But this viewpoint does not sit well with an anti-realist attitude to the semantics.

worlds; as we shall see, an elucidation of transworld identity can be regarded as an elucidation of the boundaries of possibility for ordinary things, an enterprise which must be taken seriously by any anti-realist who thinks there are objective modal facts, even if they are not about worlds.

The modal theses with which we will be concerned are theses attributing essential properties and individual essences to things of given categories, so we start with an explanation of this terminology. First, an *essential property* of an object x is a property without possessing which x could not exist. In other words, if P is an essential property of x, then for all possible worlds w, if x exists in w then x has P in w. Essential properties may be as complex as you please, but there are also simple and highly trivial examples of them. For instance, the property of *existing* is obviously essential to every object x, regardless of what kind of a thing x is. The same is true of the property of being *self-identical*. However, as we have set up our system of quantified S5, there is a difference between these two properties, for the latter is also a *necessary* property of x. We may say that a necessary property is a property possessed by x in every world (not just in worlds where x exists). Then for material objects, the paradigm of contingent existents, existence will not be a necessary property, but self-identity will be, since even if an object a does not exist at a world w, our semantics makes $a = a$ true at w. However, having drawn this contrast between necessary and essential properties, we will not refer to it again, since from now on we will be concerned only with essential properties.

The essential properties of a thing will typically depend upon what category of thing it is, and perhaps also on some more particular facts about the thing itself, so a thesis attributing an essential property will say, roughly, that for things of such-and-such a category, if certain particular facts about them are thus-and-so, this remains the case in every world where the thing exists. In possible-worlds terminology, what we want to say is that for any possible object which is of category C in some possible world, if that thing has a certain (perhaps complex) property P at that world, then at every world where that thing exists it still has P. As a modal schema, this possible-worlds formulation translates into:

$$(\mathrm{S}) \quad \Box(\forall v)\Box[(Cv \,\&\, Av) \rightarrow \Box(Ev \rightarrow Av)].$$

Note that (S) begins with a universal quantifier governed by a $\Box$, so

instances of (S) will speak of every object in every possible world, that is, all possible objects. The point is that one does not wish to limit the attribution of an essential property to just the things of a given kind which happen to exist. The schematic letter *C* stands for a predicate which identifies the category of thing to whose members we wish to attribute the essential property; sample instances of *C* are 'number', 'set', 'organism', and 'event'.[2] What follows in *Av* specifies a particular property which the consequent of the main conditional asserts to be essential. If we write out the trivial instance of (S) which says that existence is an essential property of every possible object, we obtain:

(1) $\Box(\forall x)\Box[(Tx \,\&\, Ex) \rightarrow \Box(Ex \rightarrow Ex)]$.

Here *Tx* stands for '*x* is a thing'; so in effect there is no restriction on the category of thing to which we wish to attribute existence as an essential property, and we could just as well have suppressed the symbols '(*Tx* &' in (1). In possible worlds terms, (1) says that for all possible *x*, if *x* is a thing and exists in some world, then in any world in which *x* exists, it exists. This way of reading (1) is justified by the equivalence of (1) with

(2) $\Box(\forall x)[\Diamond(Tx \,\&\, Ex) \rightarrow \Box(Ex \rightarrow Ex)]$,

which brings out the function of the second □ in (S). In future, when the context makes it clear what category of object it is which we are talking about, we suppress the category predicate. Of course, in the examples to come, the formula *Av* will be more complicated and more interesting than *Ex*, and is therefore likely to contain variables other than *v* bound by quantifiers following the initial $\forall v$. So (S) is really a simplified version of a general schema for an attribution of an essential property to things of a category *C*. The general schema is:

(S′) $\Box(\forall v)\Box(\forall u_1)\ldots\Box(\forall u_n)\Box[(Cv \,\&\, Av,u_1,\ldots,u_n) \rightarrow \Box(Ev \rightarrow Av,u_1,\ldots,u_n)]$.

The reader who finds the point of this schema difficult to grasp in the abstract should return to it once we have presented its first sub-

2. My use of 'category' here is not governed by any underlying philosophical theory of categories. Roughly, I distinguish categories when there are interestingly different things to be said about their members' individual essences.

stantial instance, the thesis of Membership Rigidity for sets, later in this chapter (page 107).

So much for the idea of an essential property. The other notion we mentioned as an object of investigation was that of an *individual essence*, and this concept can be defined in terms of essential properties. An individual essence of an object x is a set of properties I which satisfies the following two conditions:

I(i) every property P in I is an essential property of x;

I(ii) it is not possible that some object y distinct from x has every member of I.

Note that in possible worlds terminology, I(ii) says that there is no world w such that some y other than x has every P in I at w. As with essential properties, there are trivial examples of individual essences, of which the most obvious involves self-identity: the property of being identical to a is an essential property of a and furthermore, it is not possible that any object other than a has that property. This fact suggests that we might define a special sub-class of individual essences, which we can call *non-trivial* individual essences. We say that a non-trivial essential property of x is any property essential to x other than:

(a) a property x has as a consequence of some *de dicto* truth (this excludes, e.g., the property of being unmarried if a bachelor);

(b) the properties of existence, self-identity, or their weakenings (by the 'weakening' of a property P, let us mean any Q such that if x has P, then it follows by logic alone that x has Q; thus, for the number three, being identical to three and being identical either to three or to four, are both trivial essential properties, while being the cube root of twenty-seven is a non-trivial essential property, since the fact that three is the cube root of twenty seven follows only in theories containing some portion of elementary arithmetic);

(c) a property x has in virtue of a necessary truth concerning items of another category (so it is only trivially essential to material things that they are such that three is the cube root of twenty seven; the criterion also makes it trivially essential to an object x which stands in some necessary re-

lation R to a thing which is necessarily F that it is necessarily R to some F, so it does not quite capture the idea of irrelevance or independence which underlies the example given; it would take us too far afield to improve this here).

Finally, we define a non-trivial individual essence of x to be an individual essence of x none of whose members are trivial essential properties.

From now on our only interest will be in the non-trivial, so, unless explicitly stated otherwise, this qualification will be understood whenever the phrases 'essential property' or 'individual essence' are used. The motivation for investigating individual essences should be obvious, since if every object has such an essence, the problem of elucidating transworld identity can be solved. It will be recalled from Chapter 3 (page 51) that, according to Quine, there are relatively unproblematic criteria of cross-moment identification of bodies, but no such criteria of crossworld identification. However, an individual essence of an object x, in virtue of its non-triviality, would give necessary and sufficient conditions for crossworld identification of x without employing the property of being identical to x or any of its cognates. Each property in x's essence would correspond to a necessary condition, since each of them is essential to x, and the whole set of properties would give a sufficient condition, because of clause I(ii) (of course, we shall hope to find simpler conditions than those resulting from taking the whole set). The resulting transworld identity condition for x, as already remarked, would apply only to pairs of worlds u and v such that x exists in both, but this is as complete a solution to the problem as Quine's criteria of continuity of displacement, distortion and chemical change are to the corresponding problem of transtemporal identity. So our plan is to choose particular categories of object and develop a theory of the individual essences of members of the given category; we can also describe our theory as a theory of transworld identity conditions, but only as a *façon de parler*, since we are antirealists about worlds. We will begin with that category of object, sets, for which the correct account of individual essence is perhaps least controversial.

CHAPTER 5: A MODAL THEORY

The essences of sets

In this section, we are going to develop some modal intuitions about sets, with a view to formulating and then defending certain claims about the essences of sets; we shall argue that the members of a set are its essence, in the sense that if certain objects are the members of a set x in some world, then the property of having exactly those objects as members is an essence of x. Evidently, this claim can be broken apart into a necessary and a sufficient condition for transworld identity of sets, the condition of having the same members; we shall see below exactly how to say in modal language that this condition is both necessary and sufficient.

There are some preliminary matters to which we should attend. First, exactly what is a set? We shall presume a certain conception of set known as the *iterative* conception, which is best explained by contrast with another conception, according to which a set is any collection of entities which fall under some particular concept. As is well known, the unqualified version of the latter view was shown by Russell to be inconsistent. Russell found a concept c such that the 'set' which comprises those things which fall under c can be shown to be a logically impossible object; hence no set is specified by this concept, and thus the view of sets in question is refuted. The concept Russell began with is the concept of being self-membered, a concept which specifies the set of all objects which are members of themselves. If we use $\in$ for the relation of membership, this set may be written

$$\{x\colon x \in x\}$$

(read: the set of all sets x such that x is a member of x). Now this appears to be a quite reasonable specification of a set; for example, the set of all abstract objects is itself an abstract object, and is therefore a member of itself and so an element of $\{x\colon x \in x\}$. But if we next consider the concept of *not* being self-membered, which specifies the set Bertie, written

$$\{x\colon \sim(x \in x)\}$$

we can derive a contradiction by asking whether or not Bertie is a member of itself: if Bertie is a member of Bertie, then by the definition of Bertie, Bertie it is not self-membered, i.e., not a member of Bertie; while if Bertie is not a member of Bertie, then by the definition of Bertie, Bertie is a member of itself, i.e., it is a member of Bertie. So Bertie is a member of Bertie iff Bertie is not a member

of Bertie; but in classical logic, $P \leftrightarrow \sim P$ is a contradiction.

This argument shows that the conception of set as the range of items falling under a concept must be revised, and a very natural revision is suggested by the argument itself. For, on reflection, there is something strange about the idea that a set could be a member of itself. To see why, use for an analogy the idea that God is *self-created*. An obvious problem with this is that for '*x* creates *y*' to be true, *x* must bring about the existence of *y*, and *x*'s doing anything at all requires the existence of *x*. Thus God's existence is a precondition of His creating anything, including Himself: He cannot create Himself without satisfying a precondition which renders His creating Himself unnecessary, indeed, impossible. In a parallel fashion, it seems that a set cannot be a member of itself if we think of a set as a collection of entities which are somehow 'brought together', the set existing only when these entities are brought together. Then if a set was one of its own members, it would have to pre-exist itself, in order to be among the things whose being brought together constitutes the formation of the set.

These arguments against self-creation and self-membership evidently play upon some 'before/after' dichotomy. In the case of self-creation, the dichotomy arises from the *a priori* truth that creator must exist at a time before the creation first exists. In the case of sets, the use of temporal language is unavoidable but metaphorical: there is of course no physical process of bringing together certain antecedently existing objects, a process whose completion at a certain point in time marks the start of the existence of the set. But however hard it may be to explain the appropriateness of the metaphor, its naturalness cannot be denied, and it motivates a picture of the universe of sets something like the following. At the bottom level, we have all the non-sets, for which the term 'individual' is reserved. At the next level, all sets whose members are entities at the bottom level, i.e., all sets of individuals, are formed. At the level after that, all sets whose members are objects on one or other of the previous two levels are formed; at the level after that, all sets of objects from the previous three levels are formed, and so on for every finite level. The first infinite level is reached by forming all sets whose members are to be found somewhere on the finite levels, and then the process is iterated until the level of the next infinite 'limit ordinal', and then iterated again, and so on, without end. This sequence of levels is known as the cumulative hierarchy

of sets. It is cumulative because at any given stage one may at that stage form any set so long as each member of the set appears at some previous stage – one is not limited to objects from the immediately preceding stage. And because a set is formed only at a level above the levels where its members are formed, the analogy between set-formation and object-creation is preserved: no set in this hierarchy is a member of itself.

The cumulative hierarchy embodies the iterative conception of set, and has been extensively studied in the form of axiomatic theories: axioms are laid down intended to express fundamental principles, and their consequences are investigated, along with metamathematical properties of the axiomatic theory itself. The best-known of all these theories is the Zermelo-Fraenkel theory of pure sets, usually called ZF; the sets are said to be pure because the ground level of the hierarchy, the level of individuals, is taken to be empty. Thus the hierarchy begins only with the second level, where we form all sets of individuals. Since there are no individuals, the set of all individuals is the empty set; indeed, any set of individuals is the empty set, so the empty set is the only set at the second level, but the hierarchy of pure sets can be built up from there.[3]

However, the assumption that there are no individuals, or 'urelements', is rather special, and there are versions of set theory in which it is not made, for instance, ZFI, Zermelo-Fraenkel set theory with individuals, to which we shall advert further later in this chapter. But one device of ZFI is worth noting at this point. ZFI is a theory of two kinds of entities, individuals and sets, and some of its assertions are not assertions about all objects, but only about all sets, or about all individuals (nothing is both). Consider, for instance, the so-called *Axiom of Extensionality*, which in fact is a principle common to all reasonable conceptions of set. This axiom says that sets with the same members are the same set or, equivalently, that if x and y are distinct sets, then either x has a member which is not in y, or y has a member which is not in x; so the identity of a set is fixed by its membership. However, if we write the Axiom of Extensionality as follows:

3. There is an illuminating investigation of the relationship between the theory of levels, or stages, and the axioms of ZF, in Boolos [1971], where the author gives a metaphorical account of forming a set, which he attributes to Kripke, as putting a 'lasso' around its members (p. 200, fn. 7). Devlin [1979] is a very accessible formal development of the ZF theory of sets.

(E*) $(\forall x)(\forall y)[(\forall z)(z \in x \leftrightarrow z \in y) \rightarrow x = y]$

we assert that for any objects x and y, if the members of x are the same as the members of y, then x and y are the same. In ZFI, the objects comprise both individuals and sets, and individuals have no members. Thus if x and y are distinct individuals, (E*) in fact identifies them, since, as individuals, they have the same members, none at all. What we want to say is only that if x and y are sets, then they are the same if they have the same members, so we need a predicate S for being a set (and, for other assertions, a predicate I for being an individual) to formulate the axiom as we should. (E*) should have begun '$(\forall x)(\forall y)(Sx \,\&\, Sy \rightarrow$'. However, a notational convention enables us to avoid expansion into longer formulae with extra implication signs; instead of using the predicates S and I we shall reserve distinctive variables for sets and distinctive variables for individuals, using standard variables x, y, etc., as general variables, to range over both individuals and sets. For set variables, we shall use X, Y, etc., and for individual variables, i, j, etc. Formulae with the special variables may be regarded as abbreviations of formulae with the special predicates S and I. So we can rewrite the Axiom of Extensionality as

(E) $(\forall X)(\forall Y)[(\forall z)(z \in X \leftrightarrow z \in Y) \rightarrow X = Y]$.

(E) says that for any set X and any set Y, if they have the same members they are the same set. We use the general variable z for the members, since the members of a set may be either sets or individuals. Apart from this use of notation, however, no familiarity with any set theory is needed to follow the philosophical discussion below, only a grasp of the conception of sets just outlined; therefore we will not further elaborate the details of ZF or ZFI here (for ZFI I have followed Suppes [1972]).

The question to which we now wish to develop an answer is: what are the transworld identity conditions of sets? The answer we are going to give is the following:

> sets x and y in different worlds u and v respectively are the same set if and only if the members of x at u are the same as the members of y at v

but we would like to approach this answer by a route which highlights its plausibility, and we would also like to see how it might be formulated in modal language as a thesis about the essences of

CHAPTER 5: A MODAL THEORY

sets. So let us work with a concrete case. Consider the set whose members are exactly the passengers on a certain transatlantic flight' say British Airways flight BA 167 from London to New Orleans on Sunday 19th September 1982. This set, which we will call *X*, is a set of people which we have picked out by using a characteristic possessed by all and only those people, the characteristic of being a passenger on that flight. The set itself, however, is merely the collection of those people. This feature of sets, that they are 'nothing but' collections of objects, is used by set theorists to motivate the Axiom of Extensionality, (E), just given above. One might say that (E) expresses at least part of the idea that the identity of a set is completely given by the identities of its members, for if there were more to the identity of a set, there would be no reason to expect sameness of membership to yield a sufficient condition for sameness of set. By contrast, if one thinks of the characteristics we can use to pick out sets, it would be a mistake to hold that it is sufficient for these characteristics to be the same that they apply to the same objects. Thus, even if the passengers on flight BA 167/9.19.82 are all and only the members of a certain cult, the characteristics, or properties, of being on that flight and of being a member of that cult, are still distinct. There is more than one way of bringing out the difference, but the simplest is to refer to the possibility that the cult members travel on a different flight, or non-cult-members travel on that flight with them; that is, there are worlds where the extensions of the properties are different, so they cannot be the same property.

If these comments about (E) are accurate, then (E) is no mere contingent generalization about sets; rather, it is a necessary truth about them, in the 'broadly logical' sense of 'necessary', in which case we may strengthen (E) to the Axiom of Necessary Extensionality, which we abbreviate (□E):

(□E) $\quad \Box(\forall X)(\forall Y)\ [((\forall z)(z \in X \leftrightarrow z \in Y)) \rightarrow X = Y].$

Now it is tempting to think that with (□E), we are on the road to formulating in modal language our idea that for sets to be the same across worlds is for them to share, transworld, the same members. But in fact (□E) does not say anything of the kind. (□E) is a *de dicto* principle which says that (E) is true at every world *w*, which in turn means that for every *w*, sameness of membership at *w* is sufficient for identity of sets at *w*; this is an intraworld identity condition

which gives us no clue as to when a set x existing at a world u is identical to a set y existing at a distinct world v, and is quite consistent with x and y being the same set even if x's members at u are not the same as y's members at v. In other words, (□E) is consistent with a set being like a box, into which different things can be put at different times (Fine's simile); and this is exactly what we want to rule out.

The most obvious strategy for improving (□E) in this respect is to strengthen it to a *de re* principle by inserting further operators. So let us consider the simplest *de re* extension of (□E),

$$(3) \quad [\Box(\forall X)(\forall Y)\Box[((\forall z)(z \in X \leftrightarrow z \in Y)) \rightarrow X = Y].$$

(3) says that for any compossible sets X and Y, if there is some world where X and Y have the same members from among the existents of that world (remember that $\forall$ is given the actualist interpretation) then X and Y are the same set. However, it is not too difficult to see that this is false; for if X and Y are sets existing at a world u where they have different members from amongst the existents of u, and hence are different sets, we can take a world v where no member of X exists and no member of Y exists, so that X and Y do have the same members from amongst the existents of v, and are therefore the same set, by (3). But, as a matter of modal logic, X and Y cannot be distinct at u and identical at v; so the sufficient condition for identity in (3) is mistaken.

For the sake of definiteness in the discussion to come, it is as well to decide now an issue which has been ignored to date, that of whether or not the Falsehood Principle should be applied to the set-membership predicate $\in$: should we say that a sentence such as $a \in b$ is false at a world w if either a or b does not exist at w? There is good reason to favor imposition of the Principle, a reason which flows from the iterative conception and the picture of the cumulative hierarchy of levels we sketched above. If we ask what it is to be a member of a set at a world, in terms of the iterative conception, then the following answer is the most plausible: to be a member of a set at a world is to be one of the objects on which the formation operation was brought to bear when the set was formed, on the relevant level, at that world. Now forming a set is just what it is to bring it into existence, and thus if b has members at a world, it has been formed at that world and so exists at that world ('forming' applies to all members at once, rather than one by one). This

means that the truth of $a \in b$ should imply the existence of b. And so far as the members are concerned, it would be very strange to hold that an existent might be brought into existence by an operation on non-existents (if this were possible, perhaps a nonexistent God could bring himself into existence); so the existence of a seems to be implied as well. The upshot of these remarks is that imposition of the Falsehood Principle on $\in$ is well-motivated, so we take the following thesis (F) to be an axiom about sets:

(F) $\quad \Box(\forall x)\Box(\forall X)\Box(x \in X \rightarrow Ex \,\&\, EX)$.

The reader can easily determine that the imposition of (F) does not affect how things stand with (3), so we are still looking for the right way to formulate transworld identity conditions for sets.[4] Let us approach the question by another route, *via* our definitions of essential properties and individual essences; for, as we saw above (page 97), a theory of individual essence provides transworld identity conditions. Our first step, therefore, is to ask what plausible candidates there are for non-trivial essential properties of sets.

Returning to our original example of the set X of passengers on BA flight 167/9.19.82, let us compare X with the set Y of passengers on the same flight (or 'flight-type') for 9.21.82, the following Tuesday. In view of the airline's later decision to discontinue this route, we can imagine that X's members are just the travellers a, b, and c, while Y's members are just the travellers d, e, and f. By (E), X and Y are different sets. However, it is contingent that these six people travelled when they did, so we can imagine a world u where a, b, and c are the passengers on BA flight 167/9.21.82, while d, e, and f are the passengers on 167/9.19.82. Since it is the transworld identity of sets, and not flights, which is in question at the moment, let us assume that flights are the same in two worlds iff they are by the same airline, along the same route, and at the same times. Note

4. It is a consequence of (F) that the statement 'Socrates $\in$ {Socrates}' is false at worlds where Socrates does not exist. This may look strange, but it is an appearance one easily learns to live with. A quite distinct question (discussed at greater length below) is that of what set, if any, the set abstract '{Socrates}' should be said to denote at w if Socrates does not exist at w. Since names of non-existents are allowable, the abstract could be taken to denote singleton Socrates at all worlds, i.e., it could be a rigid designator. But if one reads the abstract as 'the set of all existent objects x such that x = Socrates', '{Socrates}' will be a non-rigid designator, since it will denote the empty set at worlds where Socrates does not exist. We will in fact adopt this latter, actualist, reading.

that this is a genuinely non-trivial assumption about the transworld identity conditions of flights, even although it presumes the transworld identity conditions of other entities, airlines, routes and times. Let us also introduce the labels W and Z for the respective sets of travellers in u on BA 167/9.19.82 and BA 167/9.21.82; so W's members are d, e, and f, while Z's members are a, b, and c. We are now using X, Y, W, and Z as labels for sets picked out at a particular world, X and Y at $w^\star$ and W and Z at u. Setting this out in detail, we have:

(i) X = the set of travellers in $w^\star$ on BA 167 for the 19th, = $\{a,b,c\}$;
(ii) Y = the set of travellers in $w^\star$ on BA 167 for the 21st, = $\{d,e,f\}$;
(iii) W = the set of travellers in u on BA 167 for the 19th, = $\{d,e,f\}$;
(iv) Z = the set of travellers in u on BA 167 for the 21st, = $\{a,b,c\}$.

There are then essentially two different ways of making crossworld identifications between the sets of travellers X and Y in $w^\star$ and the sets of travellers W and Z in u. We can say (A) $X = W$ and $Y = Z$, or (B) $X = Z$ and $Y = W$. Let us consider (A).

To hold that $X = W$ and $Y = Z$, as we already remarked, is to hold that sets are like boxes; just as we can put different things into the same box at different times, so we can have different objects in the same set at different worlds. However, although this is compatible with (□E), there is a strong intuition that (A) is the wrong option and that sets are not like boxes; this would seem to be part of the intuition that there is nothing more to the identity of a set than the identities of its members, a part not captured by (□E). In support of (A), we can at best cite the fact that:

(4) The set of travellers on BA flight 167/9.19.82 could have been $\{d,e,f\}$

from which we may be tempted to infer that

(5) $\{a,b,c\}$ could have had d, e, and f as its members instead of a, b, and c.

Since it is in virtue of u and worlds like it that (4) is true, from (4) we can conclude that $X = W$.

The objection to this argument, of course, is that (4) is ambiguous in the way that all subjunctive English sentences with definite descriptions are.[5] In this particular case, the ambiguity is between the readings (6) and (7):

(6) It is possible that the set of travellers on BA 167/9.19.82 has *d*, *e*, and *f* as its members;

(7) Concerning the set of travellers on BA 167/9.19.82, possibly *that* set has *d*, *e*, and *f* as its members.

In possible worlds terms, (6) says that in some world there is a set which satisfies the given description in that world and whose members are *d*, *e*, and *f* in that world. This is clearly true, in virtue of *u*, but is of no relevance to the question whether or not $X = W$, since in (6) the description 'the set of travellers on BA 167/9.19.82' does not pick out *X*, but rather, whatever set satisfies the description in a verifying world with respect to which the phrase 'it is possible that' is evaluated. Thus neither (6) nor (5) follows from (4). On the other hand, although (7) entails (5) and supports the claim that $X = W$ (because the 'concerning the *F*,…that *F*…' construction allows us to pick out *X* with the definite description 'the set of travellers on BA 167/9.19.82' and say of *X* that its members could have been *d*, *e*, and *f*), the mere existence of such a world as *u* does not suffice for the truth of (7). (7) is in fact a clear expression of the thought which intuition rebels against, that that very set *X* could have had different members.

Subsequently in this chapter, we will address the question of how the intuition of error in (7) and (5) is to be justified. For the moment, let us simply *use* the intuition to provide us with candidates for essential properties of sets: we will say that for each mem-

5. The problem is one of the relative scopes of modal operators and definite descriptions. On the face of it, the sentence 'the tutor of Alexander might not have tutored Alexander' is contradictory: it seems to say that there is a world where the individual who tutors Alexander does not tutor him. On this reading the definite description said to have *narrow scope* and the modal operator to have *wide scope:* 'it might have been that the tutor of Alexander does not tutor Alexander'. But there is an alternative interpretation of 'the tutor of Alexander might not have tutored Alexander' on which it is true – we interpret it as saying that the individual who in the actual world tutors Alexander does not do so in some possible world: 'the tutor of Alexander is someone who might not have tutored Alexander', where the definite description has wide scope and the modal operator narrow scope. Smullyan [1948] is a classic discussion of this phenomenon against the background of Russell's treatment of descriptions.

ber y of a given set it is essential to that set that y belong to it. According to this principle, the membership of a set is the same at every world in which the set exists, so we can call the principle *Membership Rigidity*, or MR for short. We formalize it as an instance of (S′), omitting the category predicate 'set' but using set variables:

(MR) $\Box(\forall X)\Box(\forall x)\Box(x \in X \rightarrow \Box(EX \rightarrow x \in X))$.

(MR) is inconsistent with the conception of sets as boxes with different members at different worlds; if $\Box$ is read as 'always', it is also inconsistent with the conception of sets as boxes with different members at different times, a conception just as unintuitive as the corresponding modal one. Indeed, taking u and $w^\star$ as terms for times, $w^\star$ standing for the present, the modal considerations which we have just worked through could all be reiterated for times, and this suggests that the explanation we give of our intuitions about the inappropriateness of the box metaphor should be able to explain why a set cannot change its members through time as well as why it cannot change them through worlds. We will see that this constraint on a successful explanation is satisfied below.

Although Membership Rigidity provides essential properties for sets, it does not yield individual essences. For a given set x, let us call the set M of properties decreed by (MR) to be essential to x the set of *membership properties* of x. That is, for each y in x, M contains the property of having y as a member; and to avoid complications with the subset relation, we will also include in M the property of having no members other than each of those objects y. Then (MR) does not by itself imply that for a given set x, the set M of membership properties of x satisfies both I(i) (that every member of M is essential to x) and I(ii) (that it is not possible that something else has every member of M). Obviously, M satisfies I(i), but the problem is with I(ii); that is, it is consistent with the truth of (MR) that the set M of membership properties of a set, even if it exhausts the non-trivial essential properties of that set, is nevertheless not an essence of that set. By inspection of I(ii), this must mean that (MR) is consistent with distinct sets having exactly the same members at certain worlds, and so we need some further principle to rule this out. The obvious candidate is ($\Box$E), which says that, in any world, sets with the same members are the same sets. But ($\Box$E) deals with only one case where distinct sets are alleged to have the same members, the case where the sets in ques-

tion are both elements of the domain of the world at which they are claimed to have the same members. Yet it is also consistent with (MR), and with (MR) and (□E) together, that there be two sets X and Y such that the membership of X at any world where X exists is the same as the membership of Y at any world where Y exists, and there is no world where X and Y both exist; for instance, someone who held the rather awkward view that individuals have transworld being but sets are worldbound individuals, would reject sameness of membership as sufficient for transworld identity, but could accept (MR) and (□E), since in virtue of the actualist treatment of quantifiers, X and Y in (□E) range over only the existents of one world at a time.

But from the nature of this counterexample to the claim that (MR) and (□E) provide essences for sets, it is clear how we should strengthen (□E) to obtain a principle which deals with it. We need a principle which says that if X is a set existing in a world u and Y a set existing in a possibly distinct world v, then if X has the same members in u as Y has in v, then X and Y are the same set. According to this principle, if sets X and Y have the same members, X in one world and Y in another, then they are the same set. So we will label the principle *Crossworld Ex tensionality*, or CE for short. Crossworld Extensionality does not have to be a component of every view which finds the box metaphor unsatisfactory, since both it and its negation are consistent with Membership Rigidity; on the other hand, it does seem to be an aspect of the intuition that there is nothing more to the identity of a set than the identity of its members, which itself is inconsistent with the box metaphor. So we can regard the two principles (MR) and (CE) as distinct consequences of this more general view about the identity of sets, however it is to be elucidated and defended.

It is not wholly straightforward to formalize (CE) in modal language, even though we found it easy to state in possible worlds language above. For instance,

$$(8)\quad \Box(\forall X)(\forall Y)\{[\Box(\forall z)(z \in X \leftrightarrow z \in Y)] \rightarrow X = Y\}$$

does not have the effect of making a crossworld comparison between the memberships of two sets, since values of X and Y will always be chosen from a single world, the one introduced by the evaluation of the initial □, and compared at a single world, the one introduced by the evaluation of the second □. Moreover, if we try

to separate the initial pair of quantifiers with a □, we obtain a principle which, though true, still does not provide the identity condition we want:

(9) $\Box(\forall X)\Box(\forall Y)\{[\Box(\forall z)(z \in X \leftrightarrow z \in Y)] \rightarrow X = Y\}$.

It is undeniable that if a set X which exists at u and a set Y which exists at v are such that, at any world, X and Y have the same members from amongst the existents of that world, then X and Y are the same set; for if X and Y are distinct and compossible, then by Necessary Extensionality and the necessity of non-identities, this difference will manifest itself as a difference in membership within the domain of any world where they both exist; while if they are not compossible, then at any world where one exists (the non-empty one, if the other is the empty set), none of the actuals of that world which belong to it will belong to the other (by the Falsehood Principle, since the other would not exist at that world). But in this argument for (9), we are envisaging worlds other than u and v at which at least one of X and Y exists, so we are implicitly assuming some transworld identity condition for X and/or Y, most likely Crossworld Extensionality, and it is the content of the latter which we seek to formalize (this point also applies to (8), and to further strengthenings of (9)).

What we have to be able to do is to choose a value a for X from one world u and a value b for Y from another world v and then keep track of u and v somehow, so that we can speak of the members of a at u and the members of b at v. For this, we use the singly-indexed 'actually' operators introduced in Chapter 4 (page 88), since the indexes enable us to keep track of the worlds where a and b have the memberships we wish to compare. The formalization is:

(CE) $\Box_1(\forall X)\Box_2(\forall Y)\{[\Box(\forall z)(\mathsf{A}_1(z \in X) \leftrightarrow \mathsf{A}_2(z \in Y))] \rightarrow X = Y\}$.

The biconditional which is antecedent of the conditional in (CE) says that being in X in the first world is equivalent to being in Y in the second, so the conditional says that if this is so then $X = Y$, and the initial string of symbols makes this a sufficient condition for identity of any two possible sets X and Y and any pair of worlds w_1 and w_2 such that X exists in w_1 and Y in w_2. Thus (CE) identifies the sets X and Y which we imagined when we showed that (MR) and (□E) are not sufficient to provide sets with individual essences.

Note that (CE) entails (□E) (the cases where $w_1 = w_2$), so one can regard (CE) as embodying the full strength of the component of the iterative conception which underpins (□E). Note also that the unindexed □ in (CE) is essential, otherwise the variable z will range over just the domain of the second world, while it can be that the intersection of that domain with the membership of X in w_1 is the same as the membership of Y in w_2, though they are different sets, since X has members at w_1 which do not exist at w_2.

In view of our definition of individual essence, we should expect to find principles analogous to (MR) and (□E) in the account of individual essence for objects of other categories. By I(i), we will look for a principle which yields non-trivial essential properties for entities of the relevant category, and then by I(ii) we may anticipate a principle like (CE) which rules out the possibility that distinct entities of the category agree on essential properties, a principle which specifies an indiscernibility condition whose holding across worlds is sufficient for transworld identity between the entities for which it holds. We will see that in fact we do find this pattern in the theory of individual essence for various categories of thing, but the indiscernibility principle is less intimately related to intraworld identity conditions than in the special case of sets, where (CE) entails (□E).

Finally, we should note an alternative way of adding to (MR) to obtain principles which fix essences for sets. In (MR) and (CE) we combine individually necessary conditions for transworld identity with a principle which makes them jointly sufficient, but we could achieve the same effect by replacing the sufficient condition for identity with a sufficient condition for existence; (CE) does not itself state an existence condition, since we are assuming, by our notational convention, that we are given sets as values of X and Y. But what might be a natural existence condition? If (MR) and (CE) are correct, then to do without (CE) we need an existence condition which implies it. A fairly natural transworld sufficient condition for existence that will imply (CE), given (MR) and (□E), is that X exists in a world w if all its members at any given world where X exists, also exist at w. So if X has certain members at u, and these members all exist at w, then by the existence condition, X exists at w, and by (MR), its members at w will be exactly the members it has at u. By (□E) no other set at w has exactly these members as well, so no counterexample to (CE) can arise. Note that in this

argument, (MR) guarantees that it does not matter which X-containing world u we begin with.

The modal sentence which most directly expresses the sufficient condition for existence just formulated, a condition we call 'Set Existence', also uses the singly-indexed 'actually' operators in its formulation:

(SE) $\quad \Box(\forall X)\Box_1\{[\Diamond(EX \,\&\, (\forall y)(y \in X \rightarrow \mathsf{A}_1(Ey)))] \rightarrow EX\}.$

This formula says that for any possible X ('$\Box(\forall X)$') and any chosen world ('$\Box_1$'), if there is some world in which X exists such that every member of X in that world exists in the chosen world, then X exists in the chosen world. No indexed operator is needed to govern the consequent of the conditional, since in evaluating (SE), EX would be evaluated with respect to the world introduced by the modal operator within whose immediate scope it lies, in this case the chosen world introduced by the indexed $\Box$. And what the argument one paragraph back shows is that (MR), ($\Box$E) and (SE) together imply (CE). However, (MR) and (CE) do not imply (SE) unless we assume a further principle. This is the principle that, given any objects at a world, all sets which can be built up from those objects also exist at that world; we might call this the principle of *automatic set formation*. If this principle in fact holds at every world, we can argue from (CE) to (SE) thus: if X's members at u also exist at v, then by automatic set formation at v, the set of them will exist at v, which by (CE) is the set X; so (SE) is verified, in fact without using (MR).

If we embed our modal theses about sets within a theory intended for structures with automatic set formation, there is little to choose between (SE) and (CE). However, automatic set formation is a controversial principle against the background of *de re* modality, at least, a more controversial principle than (CE), and has been challenged (Parsons [1977]); for someone might say that although the availability of the members of a set guarantees the possibility of forming the set, the formation operation need not actually be carried out. Thus a minimal *de re* theory of the essences of sets is better formulated with the less controversial (CE) as an axiom, since the truth of (MR) renders (SE) a conditional automatic formation principle: if X exists at w and its members at w exist at u, (SE) compels the formation at u of a set with exactly those members as its members, since it compels the existence of X at u and,

by (MR), X must have the same membership at u as it does at w. There are also other reasons to prefer (CE) to (SE). For instance, if we use (SE) we will still need one of (CE)'s consequences, ($\Box$E), so an account appealing to (SE) is less economical. Secondly, for other categories of entity, an existence condition is less easy to state than a crossworld indiscernibility condition. And, thirdly, we will discover that the justification of such principles as these turns on features of the concept of identity, so that (SE) is less directly justified by the relevant considerations than a sufficient condition for identity like (CE); indeed, in view of the controversial aspect of (SE), it may not be justified by these considerations at all.

The system MST

Although it is not to the main point of this chapter, it is of some interest to consider how *de re* modal theses about sets could be added to a standard axiomatic theory such as ZFI to obtain a system of modal set theory. We now present one such system, called MST.[6] The reader with no interest in, or knowledge of, axiomatic set theory can proceed without loss of continuity to our discussion of the philosophical basis of Membership Rigidity on page 121.

The language of MST, L_{MST}, is a first-order modal language with three predicate symbols, the one-place symbols S and I for being a set and being an individual, and the two-place symbol '$\in$' for set membership. Where possible, we will exploit our convention about variables to avoid use of S and I, so L_{MST} also has three sorts of variables, general variables, set variables and individual variables, as explained above (the abbreviating formulae, as well as those abbreviated, are in Form[L_{MST}], the set of wffs of L_{MST}). The theory has *de re* axioms dealing with transworld facts about sets, individuals, and set membership, together with *de dicto* axioms dealing with the intraworld theory of sets. Our intention is that the intraworld theory should be ZFI.

The first three *de re* axioms are (F), (MR), and (CE), repeated here:

(F) $\quad \Box(\forall x)\Box(\forall X)\Box(x \in X \rightarrow (Ex \,\&\, EX))$.

6. I developed the main lines of MST after reading Fine [1977a,b], but independently of Fine [1981b]. However, I am indebted to the last of these papers for clarification of a number of points, especially on the confusing topics of abstracts, and to Fine himself for extensive comments on an earlier version of this chapter.

(MR) $\Box(\forall x)\Box(\forall X)\Box(x \in X \rightarrow \Box(EX \rightarrow x \in X))$.

(CE) $\Box_1(\forall X)\Box_2(\forall Y)\{[\Box(\forall z)(\mathsf{A}_1(z \in X) \leftrightarrow \mathsf{A}_2(z \in Y))] \rightarrow X = Y\}$.

The next two *de re* axioms deal with *I* and *S*. We have to decide whether it is merely essential to a set that it is a set, or whether it is necessary, *mutatis mutandis* for individuals. Following our discussion of the Falsehood Principle on page 103, being a set at a world requires being formed at that world, i.e., being brought into existence at that world, and thus we should assert $\Box(\forall x)\Box(Sx \rightarrow Ex)$. But then if sets were necessarily sets, every set would be a necessary existent. We therefore adopt the following *Axiom of Set Rigidity*:

(SR) $\Box(\forall x)\Box(Sx \rightarrow \Box(Ex \rightarrow Sx))$.

If we also held the Falsehood Principle for *I*, we could infer an analogue of (SR) for individuals. Another option is to leave this question open and to take the essentiality of individuality as an axiom:

(IR) $\Box(\forall x)\Box(Ix \rightarrow \Box(Ex \rightarrow Ix))$.[7]

To continue MST, we want to add axioms to the effect that ZFI is true at every world. We therefore add to the above axioms the *de dicto* necessitations of Pairing, Null Set, Power Set, Foundation, and Infinity, but not the principle (□E) already described, since it follows from (CE). Thus the empty set, which we denote $\varnothing$, is a necessary existent, by □[Null Set] (which gives us some empty set at each world) together with (CE), which implies that there is exactly one such set at each world and that the empty set of any one world is the same set as that of any other. In fact, this follows from □[Infinity], one *de dicto* instance of Comprehension (see below), and (CE), hence □[Null Set] is strictly redundant. We add to this list the axiom '□(The existing individuals, if any, form a set)', i.e.,

(IFS) $\Box(\exists X)(\forall y)(y \in X \leftrightarrow Iy)$

since only if the individuals are finite in number is the existence of a set of them guaranteed by the other axioms.

7. The axioms to this point are in a sense 'modal' rather than set-theoretic, and are by no means all the axioms of this sort we might want to include. For instance, in a fuller treatment we should either add $\Box(\forall x)\Box(Sx \rightarrow Ex)$, 'existence is necessary for being a set', or discuss what would be excluded that is allowed as things stand.

However, there are alternative ways to define the remaining axioms, since in their non-modal versions, the defining principles, Comprehension and Replacement, are schemata, so questions arise about what restrictions are appropriate on their modal instances. Here are the non-modal principles:

> *Comprehension:* For any formula φ with free variables among $x, Z, r_1,\ldots,r_k$,
>
> $$(\forall Z)(\forall r_1)\ldots(\forall r_k)(\exists Y)(\forall x)(x \in Y \leftrightarrow (x \in Z \,\&\, \varphi))$$
>
> is an axiom.
>
> *Replacement:* For each formula φ with free variables among $x, Z, y, r_1,\ldots,r_k$,
>
> $$(\forall Z)(\forall r_1)\ldots(\forall r_k)$$
> $$[(\forall x \in Z)(\exists! y)\varphi \rightarrow (\exists Y)(\forall x \in Z)(\exists y \in Y)\varphi]$$
>
> is an axiom.

For each principle, we have to consider whether or not its MST version should restrict substitutions for φ to non-modal formulae of L_{MST}, and whether or not the values of the parameters and Z should all be chosen from the domain of a single world. If the schemata are to yield only *de dicto* axioms, we must impose both restrictions on each. The intuitive idea that, at each world, all sets which can be built out of individuals from the domain of that world should exist, provides some guidance here. For example, the idea underlying Comprehension is that given any set X, any subset Y of X definable from parameters should exist. 'Definable' means 'definable by an expression of the language', and, since the language we are working with allows modal formulae, a restriction of φ in the schema to non-modal formula would open up the possibility of there being a modally definable set which clearly exists, but whose existence is not a consequence of the axioms. As an example of such a set, we may consider the subset of X whose members are the contingently existing individuals in X. In other words, if given the existence of X,

$$(10)\quad (\exists Z)(\forall i)(i \in Z \leftrightarrow i \in X \,\&\, \Diamond{\sim}Ei)$$

is to follow from the axioms, modal φ must be allowed in Comprehension. And as Fine has pointed out ([1981b], p. 186), if

(11) $\Box(\exists X)(\forall i)(i \in X \leftrightarrow \Diamond{\sim}Ei)$

is to hold, similarly in Replacement.

The issue of parameters, that is, the issue of whether to insert $\Box$ between each $(\forall r_i)$, is less easily resolved. If a subset of a given set at a world can only be defined with parameters including non-existents at that world, it is unclear that it belongs in the cumulative hierarchy of that world, since non-existents then enter essentially into its generation, while our conception of the hierarchy is that the existence of every set in it is determined just by that of the entities at lower levels. But to allow merely possible parameters is not necessarily to admit such a set, since a condition can pick out a set without corresponding to the process which generates the set. The question is whether the Comprehension Principle will have false instances if merely possible parameters are allowed. Note that in formulating non-modal Comprehension above, we might have restricted the (values of the) parameters to Z, but instead we allow them to range over the universe of sets. So in the modal case we will allow them to range over the modal universe of sets:

> $\Box$-*Comprehension:* For any formula φ with free variables among x, Z, $r_1,\ldots,r_k$,
>
> $\Box(\forall r_1)\Box\ldots\Box(\forall r_k)\Box(\forall Z)(\exists Y)(\forall x)(x \in Y \leftrightarrow (x \in Z \,\&\, \varphi))$
>
> is an axiom.

What of Replacement? The idea behind this principle is that certain sets exist because they can be defined as the result of replacing the members of a given set, one for one or one for many, by the entities which are to make up the new set (it follows from this that the cardinality of the new set cannot exceed that of the given set), and it does not seem to matter to this intuitive idea if the definition of the replacement relation has merely possible parameters. But will theories of different strengths result, depending on how we decide this question for $\Box$-Replacement? Concerning some modal set theories based on ZFI but without (F), Fine has conjectured that when merely possible parameters are excluded, the sentence

(12) $\Box(\forall X)\Box(\exists Y)(\forall x)(x \in X \leftrightarrow x \in Y)$

is not a consequence of the axioms. (12) holds in MST, but for this

system there is an analogous problem, that of whether (13) holds:

(13) $\Box(\forall X)\Box(\exists Y)(\forall x)(\Diamond(x \in X) \leftrightarrow x \in Y)$.

(13) will not hold in MST with *de dicto* Replacement if (12) is not a theorem of Fine's systems with *de dicto* Replacement; therefore, since (13) seems true – it says that given any possible set X and any world w, the existents of w which belong to X in some world form a set Y at w – we shall adopt the stronger formulation as the Replacement principle for MST:

> $\Box$-*Replacement:* For each formula φ with free variables among x, Z, y, $r_1,\ldots,r_k$,
>
> $\Box(\forall r_1)\Box\ldots\Box(\forall r_k)\Box(\forall Z)$
> $[(\forall x \in Z)(\exists!y)\varphi \rightarrow (\exists Y)(\forall x \in Z)(\exists y \in Y)\varphi]$
>
> is an axiom.

This completes the listing of the axioms of MST.

A model for MST is a quantified S5 model in which all axioms of MST are true. More exactly, it is a 5-tuple $\langle W, D, d, v, w^\star\rangle$ such that every axiom of MST is true at $w^\star$, where W, D, d and $w^\star$ are as on page 26 and v is a valuation function defined for L_{MST}: $v(I)$ and $v(S)$ are disjoint subsets of $D \times W$ such that $v(I,w) \cup v(S,w) = d(w)$ and $\forall u$, $v(I,w) \cap v(S,u) = \varnothing$, and $v(\in)$ is a subset of $D \times D \times W$ such that $\langle a,b,w\rangle \in v(\in)$ only if $b \in v(S,w)$. Since L_{MST} contains no individual constants, the evaluation of a sentence of L_{MST} in a model $\mathbf{M}$ proceeds through the expanded language L', which contains a name for each entity in the domain D of $\mathbf{M}$. **Mod**(MST) is the class of all models of MST, and a theorem of MST is a sentence true in every model in **Mod**(MST).

A rather obvious theorem of MST is

(14) $\Box(\forall x)\Box(\forall X)\Box(x \in X \rightarrow \Box(EX \rightarrow Ex))$.

For if $a \in b$ holds at w and Eb holds at u then, by (MR), $a \in b$ holds at u and thus by (F), Ea holds at u. However, to establish *de re* theorems of MST involving other Boolean notions customary in set theory, one must define these notions carefully. For instance we define the subset relation thus:

$$a \subseteq b =_{\text{df}} Sa \,\&\, Sb \,\&\, Ea \,\&\, Eb \,\&\, (\forall r)(r \in a \rightarrow r \in b).$$

Each condition in addition to the standard one is required here; for

example, without the third conjunct, every non-existent set at w would be counted a subset at w of any existent set at w. With this definition, it then follows that

(15) $\Box(\forall X)\Box_1[EX \rightarrow \Box(\forall Y)(Y \subseteq X \rightarrow \mathsf{A}_1(Y \subseteq X))]$.

Readers may wish to develop the *de re* portion of MST further themselves. We address the question of whether the *de dicto* portion of MST is stronger than ZFI below.

In standard set theory, it is usual to augment the operators of the formal language with the abstraction operator, using which we can form certain terms, *set abstracts*, for denoting sets. To be precise, L_{MST} is expanded by addition of the new operator, and in the syntax of the expanded language, L_{MSTA}, we deal with this operator as follows:

> *Abstraction:* if v is a general variable and φ a wff, $\{v\colon \varphi\}$ is a term with each occurrence of v in φ bound.

In a standard theory, $\{v\colon \varphi\}$ may be read 'the set of all objects v such that φ', and the operator does not add to the expressive power of the language, since any sentence σ containing abstraction terms can be translated into some sentence σ' which is free of abstraction terms and which is provably equivalent in the theory to σ (see Quine [1963], Chs. 1 and 2).

However, in modal set theory, the treatment of set abstracts requires some care. First, the reading of abstracts suggested above is ambiguous, in that in a modal context 'all' may be either an actualist or a possibilist quantifier. In our version of S5, quantifiers are actualist, and so, consonant with this, we admit only actualist abstracts: we read the abstract as 'the set of all existent objects such that…'. Secondly, when abstracts are admitted (whether possibilist or actualist) we thereby admit non-rigid singular terms; for example, '$\{x\colon Ix\}$' stands for different sets (the domain of existing individuals) at different worlds, since individuals may be contingent existents. Similarly, if we had constants a and b in the language, then at a world u where a and b both exist, '$\{a,b\}$', which abbreviates '$\{x\colon x = a \vee x = b\}$', would denote the set whose sole members are a and b, but at a world v where only a exists, '$\{a,b\}$' denotes at v the same set as is denoted at v by '$\{a\}$', the set whose sole member is a; so '$\{a,b\}$' denotes different sets at different worlds. Thus 'Socrates $\in$ {Socrates}' is false at worlds where

Socrates does not exist even in theories without the Falsehood Principle, for under the actualist reading of set abstracts, the truth of that sentence would require Socrates to be a member of the empty set at such worlds; but since Socrates is not a member of the empty set at worlds where he does exist ($\varnothing$ has no existent member at any world) and since there must be such a world (every x in D is in some $d(w)$) this is inconsistent with (MR).

However, as in the standard case, the actualist abstract does not increase the expressive power of the language; in fact, the following clauses state equivalences on the basis of which abstracts may be eliminated (granted $\sim\Diamond(\exists x)(Sx \,\&\, Ix)$):

(i) $I\{x\colon \varphi\} \leftrightarrow \bot$;
(ii) $S\{x\colon \varphi\} \leftrightarrow \sim\bot$;
(iii) $y \in \{x\colon \varphi\} \leftrightarrow Ey \,\&\, E\{x\colon \varphi\} \,\&\, \varphi[y/x]$;
(iv) $E\{x\colon \varphi\} \leftrightarrow (\exists z)(\mathrm{S}z \,\&\, z = \{x\colon \varphi\})$;
(v) $t = \{y\colon \varphi\} \leftrightarrow (\forall z)(z \in t \leftrightarrow \varphi[z/y])$, for any term t;
(vi) $\{x\colon \varphi\} \in y \leftrightarrow Sy \,\&\, (\exists z)(\mathrm{S}z \,\&\, z = \{x\colon \varphi\} \,\&\, z \in y)$.

Here $\bot$ is a contradiction, so $\sim\bot$ is a tautology.[8]

It is helpful to distinguish two attitudes to abstracts. On the first attitude, the abstraction notation is simply an abbreviatory device, the abbreviated sentence being the one which results when the abstracts are eliminated from the given sentence in accordance with the clauses above. From this point of view, the syntactic formation clause we gave for the notation is redundant: an abstract is well-formed in a given sentence iff a well-formed sentence of L_{MST} results when the abstract is eliminated from that sentence. The formation clause is relevant only on the second attitude to abstracts, where we regard their addition to L_{MST} as resulting in a new lan-

8. Clause (ii) may seem strange: do we really want to agree that $S\{x\colon x \notin x\} \leftrightarrow \sim(P \,\&\, \sim P)$? The problem is that in a model of MST, the Russell abstract has no denotation at any world. But to apply the predicate S to an abstract is not to commit oneself to the existence, or even merely possible existence, of a denotation for the abstract – we can *stipulate*, for technical convenience, that 'Bertie is a set' is to be true, even though there is no actual or possible set answering to 'Bertie'. If one were to give a semantics for abstracts, treating them like definite descriptions, then on a Scott-style approach (Scott [1967]) the necessarily non-denoting abstracts would be assigned objects not in the domain D of a model of MST, and in the resulting extended model the extension of S would include such objects. This is also Fine's procedure ([1981b], p. 190). But other approaches are possible, e.g., a Russellian one, on which clause (ii) would be inappropriate.

guage. On this approach, the semantic account of a model of L_{MST} must also be extended, so that every sentence of the expanded language can be evaluated; in particular, an evaluation clause for terms formed by abstraction must be stated. The second view then faces the problem of non-denoting terms: the quantifier rules for L_{MST} with abstracts must be stated so that we avoid making inferences which are incorrect because a given abstract does not denote, and the evaluation clause must explain how truth-values are to be assigned to sentences with non-denoting abstracts. Since on the first attitude to abstracts, evaluation and inference are really carried out only with respect to unabbreviated sentences, so that our account of quantified S5 in Chapter 2 already suffices for the development of MST, the first attitude is the one we adopt.[9]

The distinction between the two attitudes is of obvious relevance to non-modal set-theory; if abstracts are primitive in the language, then the quantifier rules must be those of free logic; for if not, we could infer that there is a set which does not exist, $(\exists Y)\sim(\exists X)(X = Y)$, which is a classical contradiction, from the theorem $\sim(\exists X)(X = \{z\colon Sz \,\&\, z \notin z\})$, proved in ZFI by Russell's argument (note that in ZFI, the set of all non-self-membered individuals is just the set of all individuals). However, since the quantifier rules of MST are already free, it may be thought that in fact the complications ensuing from the second attitude to abstracts have been exaggerated. But this is not so, since an analogous problem arises in MST in virtue of the phenomenon of 'impossible sets'. These are of two categories. First, certain abstracts may fail to pick out a set at any world in an L_{MST}-model in Mod(MST) because of the way objects are distributed throughout the domains of the worlds of that model: for instance, 'the set of all possible objects x such that possibly x exists' denotes a set at a world in a model **M** if, but only if, for some w in W_{M}, $d(w) = D$. However, we have avoided such troublesome abstracts by the actualist reading of the abstraction notation. But secondly, there are abstracts which fail to pick out a set at any world in any model because it is a theorem of MST that necessarily no set satisfies the condition abstracted upon: '$\{x\colon Sx \,\&\, x \notin x\}$' and '$\{x\colon x = x\}$' are examples. So the problems for the second attitude to abstracts still arise in MST, since from the necessary non-existence of Bertie we could infer in quantified S5 the

9. I am not claiming that the second attitude is mistaken or unworkable. See Scott [1967] and Bencivenga [1976] for its implementation.

possible existence of something possibly identical to a necessary non-existent; in other words, terms for necessary non-existents in quantified S5 are analogous to terms for non-existents in classical logic.

This last remark raises broader issues about the relationship between ZFI and MST. In particular, we should ask if MST is a *conservative extension* of ZFI. MST is conservative over ZFI iff:

Con: For any sentence σ of L_{ZFI} MST $\models \sigma$ only if ZFI $\models \sigma$.

Now it may seem clear that *Con* fails, for the logic of ZFI is classical and therefore

(16) $(\forall x)Ix \rightarrow (\exists x)Ix$

is valid, hence a theorem of ZFI, but (16) is not valid in free logic. However, (16) is nevertheless a theorem of MST, since ∅ exists at every world of every **M** in Mod(MST). More generally, any closed sentence σ with no individual constants is such that $(\exists x)Ex \rightarrow \sigma$ is valid in free logic iff σ is valid in classical logic, and since we have $E\varnothing$ as a theorem of ZFI, a set theory just like ZFI but for having the quantifiers of free logic will have the same theorems as ZFI (assuming the first attitude to abstracts). However, if L_{MST} and L_{ZFI} are augmented by constants which do not always denote necessary existents in MST, then in the statement of *Con*, ZFI should be understood to be free ZFI.

The proof of *Con* is straightforward. In the light of our recent remarks, the left to right direction is immediate, since all axioms of ZFI are clearly theorems of MST, their necessitations being axioms of MST. For the right to left direction, we need only prove the equivalent claim

(★) Any sentence true in a model of ZFI is true in a model of MST.

However, (★) is also immediate, since any model of ZFI is *ipso facto* a model of MST. More exactly, let $\mathbf{M} = (A_{\text{M}}, v_{\text{M}})$ be a model of ZFI in which σ is true. Define $\mathbf{M}' = (D, W, d, v, w^{\star})$ in which $D = A_{\text{M}}$, W contains just one world, which is therefore $w^{\star}$, $d(w^{\star}) = A_{\text{M}}$ and v is v_{M}. Obviously, σ is true in $\mathbf{M}'$, so it remains only to check that $\mathbf{M}'$ is in Mod(ZFI). But this is also obvious, for there is only one world in **M**, all possible objects exist at it, and all ZFI axioms are true at it; thus the *de re* axioms of MST are automatically true – for

instance, (CE) is true since Extensionality is true. Hence MST is conservative over ZFI.

To bring our discussion of MST to an end, note finally that the above argument also establishes that MST is consistent if ZFI is. And, of course, MST is incomplete, as ZFI is, on account of Gödel's First Incompleteness Theorem.

Membership Rigidity: two unsuccessful justifications

For the special case of sets, (MR) alone refutes Quine's claim that anything can be changed to anything by easy stages, and in conjunction with (CE), provides a complete non-trivial account of the transworld identity conditions of sets. Given any possible sets *X* and *Y*, *X* existing in *u* and *Y* existing in *v*, we can say that *X* and *Y* are the same iff they have the same members in their respective worlds. Of course, the members of a set may themselves be sets, but this does not render our claim about *X* and *Y* trivial. If some of the members of *X* and *Y* are sets, one can apply the same identity condition over again, and so on, until we reach sets whose members (if any) are not sets. If we have reached the empty set, there is no need to go further, while if we have reached non-sets, we will then need to apply a different theory of individual essence to say when they are identical across worlds; but however that should go, the question about the transworld identity of the original sets *X* and *Y* has been completely resolved. In particular, if one restricts the objectual quantifiers to the pure sets, then (MR) and (CE) constitute a complete vindication of *de re* modality, a vindication which is extended as theories of essence are provided for further categories of thing.

However, it is one thing to write down some intuitively appealing principles which for a particular category of entity meet Quine's challenge, another to produce an argument to justify these principles. Quine was not denying that arbitrary stipulations could block the transformation of anything into anything by easy stages, so we have to show that (MR) and (CE) are better than arbitrary stipulations. We will begin by looking at two attempts to justify (MR), one essentially due to Richard Sharvy [1968], the other to David Wiggins ([1980], pp. 112–117). Since the role of (CE) has not been widely noticed or discussed in the literature, we will not examine other arguments about it here, but in view of both principles' apparent connection with the intuitive idea that a set is 'nothing but' the collection of its members, we might expect a defense

of (MR) to employ resources sufficient for the defense of (CE) as well. This will be true at least of our own account.

In fact, Sharvy addresses (MR) in its temporal rather than its modal guise; he seeks to explain why a set cannot change its membership through time. But it is easy to rewrite his explanation in modal terms. The crux of his view is this (p. 311):

> That is…the reason why no class can change its members: an apparent variable class would have to be identical (at any given time) to some class that does not change its members. So at that time, Leibniz's Law would be violated.

In effect, then, this is a temporal version of the modal argument Kripke gives for the necessity of identity, which as we saw in Chapter 3 (page 67), threatens to force the counterpart theorist to deny Leibniz's Law. Sharvy's idea is that someone who allows sets to change their members from time to time, or equivalently, from world to world, will also have to deny Leibniz's Law when temporal or modal properties are admitted: for of the alleged identical sets, one will have the property of possibly having different members (or in the future having different members) while the other will not. But, in fact, there is no such implication. Consider the following simple model. We let $W = \{w, u\}$ and $D = \{a,b,c\}$. Let $d(w) = \{a,c\}$ and $d(u) = \{a,b,c\}$; let the extension of S at w be $\{c\}$ and at u also; let the extension of I at w be $\{a\}$ and at u, $\{a,b\}$; and let the extension of the set-membership relation at w be $\{\langle a,c\rangle\}$ and at u be $\{\langle b,c\rangle\}$, i.e., $v(I) = \{\langle a,w\rangle, \langle b,u\rangle\}$, $v(S) = \{\langle c,w\rangle, \langle c,u\rangle\}$, and $v(\in) = \{\langle a,c,w\rangle, \langle b,c,u\rangle\}$, as in the diagram; finally, let $w = w^\star$. (MR) fails in this model. But there is no contradiction with Leibniz's Law, since there is no property which this model represents as being possessed by c and also, somehow, not possessed by it. From Sharvy's point of view, the trouble with the model is that it does not contain a set which keeps the same members through worlds: no set $\{a\}$ exists at u.[10] But such a set cannot be forced upon an objector to Membership Rigidity, so it does not follow from Leibniz's Law that his position is mistaken.

w	u
•	•
$\{a,c\}$	$\{a,b,c\}$
$Ext(\in)$	$Ext(\in)$
$= \{\langle a,c\rangle\}$	$= \{\langle b,c\rangle\}$

Wiggins has a more complex, but no more successful, defense

10. Given $\langle x,c,w\rangle \in v(\in)$ iff $x = a$, automatic set formation (page 111) and (CE) rule out $\langle b,c,u\rangle \in v(\in)$. This is in line with Sharvy's argument, but does not save it, since if I am right (see following), the considerations that justify (CE) are the same as the ones that justify (MR).

of Membership Rigidity. His idea is that the truth of (MR) derives from the necessary conditions for singling out this or that sort of entity and the consequential limitations on the extent to which an entity which must be singled out in such and such a way can be envisaged other than it actually is. So Wiggins claims that since a given set must be singled out as the possessor of such and such members, the set thus singled out cannot be envisaged with different members. Thus Wiggins gives a semantico-psychological interpretation of the idea that there is nothing more to the identity of a set than the identity of its members: to single out, or think of, a set, is to single out, or think of, a thing with *these* members. But this falls a long way short of justifying (MR). Singling out, or thinking of, a thing, is an activity of sentient beings carried out at particular worlds, and while it may be true that singling out a specific set X at two worlds u and v involves singling out its members at u and singling out its members at v, this fact does not imply that its members at u and its members at v have to be the same. Wiggins holds that since we have to single out a set *via* its members, we cannot envisage a singled-out set having different members, but this, again, is a *non-sequitur*. The necessity to which Wiggins appeals lies in the singling out relation, and it is quite consistent with there being only one way of thinking of an object, or way of singling it out, that the properties of the object which figure in this way are contingent properties of it.[11]

Wiggins also holds that someone who denies Membership Rigidity does not really have any grounds for assenting to (□E), the necessitation of extensionality. He writes (1980, p. 113):

> Suppose someone doubted the necessity of the membership relation. How could he combine the doubt with a reasoned affirmation of extensionality, or advance on behalf of extensionality such claims as 'a set is nothing more than a unity constituted by its members' (Richmond Thomason, *Symbolic Logic*, London, Collier MacMillan 1970, p. 284)? If there is no other way of identifying such a unity than via its constituents, then its identity is derivative from these…There is no sense then in the idea of a set $\{x, y\}$ with actual members x, y, turning up in another possible world lacking x or y.

Again, it seems that if there is no other way of identifying such

11. See fn. 6 of Chapter 6 for further discussion of Wiggins's views.

a unity than *via* its members, the most Wiggins is entitled to claim is that the unity's identity in the world in which it is being identified is derivative from its constituents in that world, which is consistent with its having different constituents in different worlds. Furthermore, this passage apparently claims that (□E) is some kind of consequence of (MR). But this is not so: it is Crossworld Extensionality, not Membership Rigidity, which implies (□E), and these two *de re* theses are logically independent: distinct sets can have the same members, provided they do not co-exist at any world, even though every set has the same members at every world where it exists, while indiscernible sets may always be identical across worlds, even when some set changes its membership through worlds in which it exists. Hence, despite the efforts of Sharvy and Wiggins, we still lack a justification for our views about the individual essences of sets.

The grounding of identities and non-identities

Our fundamental intuition about sets is that there is nothing more to their identities than the identities of their members. We used this intuition earlier to motivate (E) and (□E), but there is no obvious reason why this intuition should not be thought of as applying to transworld identity itself. (MR) and (CE) would then be seen simply as technical articulations of this idea. This move suggests two questions which may be pressed against the sceptic about (MR) and (CE). If there is nothing to the transworld identity of a set over and above the transworld identities of its members, we can ask of an alleged counterexample to (MR) what *makes* the two sets in the example one and the same. In the terms of our earlier case, for instance, we can ask what could make the set X of passengers on BA 167/9.19.82 in the actual world, $\{a,b,c\}$, the same set as the set W of passengers on BA 167/9.19.82 in u, $\{d,e,f\}$, and no substantive answer to this question could avoid departing from our intuition about the identity of sets. Similarly, if presented with an alleged counterexample to (CE), an example in which a set X has the same members at a world u as a set Y has at v, and X and Y are claimed to be different sets, we can ask *in virtue of what* the difference between them arises. In general, then, we are asking the sceptic about (MR) and (CE) for an account of that in which the alleged transworld identities and differences in his counterexamples consist; we are asking for the *grounds* of the claimed numerical difference or numerical identity in the putative refutations of (MR) and (CE).

It is difficult not to believe that our assent to (MR) and (CE) is motivated by such considerations as these. But we have still to find some underlying justification for these considerations, since without such a justification it is open to the sceptic to reject our demand for a description of the grounds of his identity judgements, or else to claim that the identity between, say, *X* and *W* above, is grounded in the fact that the passengers in *X* are on the same flight as the passengers in *W*. To be in a position effectively to counter either response, we need two things:

(I) We need an argument for the more general view that numerical identities and differences in transworld identity must be grounded in some way, so that the sceptic cannot reject our demand for grounds.

(II) We need a convincing demonstration that someone who denies (MR) and (CE) is deprived of any way of grounding facts about identity; in conjunction with (I), this refutes the imagined sceptical move of appealing to sameness of flight in our example.

Below, we make a start on these tasks, presenting the basics of a position which will be developed as the same issues arise in connection with the analogues of (MR) and (CE) in the theories of the individual essences of entities of categories other than set.

Let us start with (I). It is difficult to find an irresistible argument for the principle that facts about identities and differences must be grounded in some way, but the thesis that it is part of the content of our concept of identity, whether transworld or transtemporal, that there are no ungrounded facts about such identities, can be supported by illustration from a wide variety of cases.

Case 1. Consider the supposition that things could have been exactly as they are except that the steel tower in Paris opposite the Palais du Chaillot is different from the one actually there (the Eiffel Tower). To make sense of this supposition, it is not permitted to imagine that the tower is made of different metal from the metal which actually constitutes it, or that it has a different design, or designer, or history. The only respect in which the imagined situation is to differ from the actual world is in the identity of the tower. The extent to which such a difference seems unintelligible is some measure of the plausibility of the view that transworld differences must be grounded: in Dummett's terminology, the example shows

the strangeness of the idea that there can be 'bare' differences in transworld identity; rather, there must be something in which such differences consist.

Case 2. What holds for differences also holds for identities. An interesting illustration of the peculiarity of 'bare' or 'primitive' transworld identity is provided by the hypothesis that an actually untwinned human might have had an identical (monovular) twin. Some facts of life are relevant here. Identical twins are produced when the normal processes consequent upon the formation of the zygote (the cell which is formed by the fusion of the father's sperm with the mother's egg) break down – after the first mitotic division of the zygote, the resulting daughter cells separate in a non-standard way and develop into distinct embryos. The important fact is that mitotic division is physically symmetric, since each daughter cell receives a copy of the chromosomes of the parent cell, and each copy is, so to speak, semi-original. A chromosome consists of a pair of intertwined strings of DNA, and each string is itself a sequence of molecules called nucleotides. Replication is effected by the intertwined strings unravelling from each other, each string then acting as a template for the construction from materials present in the cell of a sequence of nucleotides exactly like the one from which the template string has just unravelled. Then each template string resumes the double helix structure by intertwining itself with the new string whose construction it has just directed. Thus two chromosomes are obtained from one, each new chromosome contains half the matter of the original chromosome, and the new matter in such a pair of daughter chromosomes has the same source. The two chromosomes now proceed to opposite ends of the cell nucleus, and since replication has been occurring with all the chromosomes in the cell, division of the cell down the middle produces a pair of cells indistinguishable in genetic content (if replication was error-free), each containing half the genetic material of the original cell.

Suppose, then, that *A* is an actual but untwinned human being. Could *A* have been an identical twin? Granted that by 'identical twin' we mean a pair of individuals produced by the above process, an affirmative answer to this question amounts to the claim that there is a possible world in which the zygote of *A* in that world undergoes fission as described, followed by the nonstandard separation giving rise to two individuals *B* and *C*; and, furthermore, *A*

is identical to *B* or to *C*. But we ought to be very reluctant to say that there is such a world. Of course, it is possible that *A*'s zygote divides in the required way, but from the account of mitosis just sketched, it is evident that there is nothing in such a situation to determine which of *B* and *C* is identical to *A*; nothing grounds one of the identities rather than the other. Perhaps the example is rather underdescribed. There will be some worlds in which *B* turns out to be rather more like *A* as he is in the actual world than does *C*, but there will also be worlds in which the converse is true. This claim might be rejected, on the grounds that in each world where such twins are produced, we should identify *A* with whichever of the twins is more similar to *A* as he actually is. But this procedure will make it impossible for *A* to be one of a pair of twins who are equally like, or the other of whom is the more like, *A* as he actually is, which seems wrong if we are going to allow *A* to be one of a pair of monovular twins in the first place. Hence someone who wishes to speak of transworld identity has to say that the fact about which of the twins at a given world is identical to *A* is a primitive, ungrounded fact, a bare truth about the transworld identity relationship which is entirely lacking in actual traces, be they as verification transcendent as you please. Again, the degree to which the supposed identity is hard to grasp is a measure of the plausibility of the view that real identities are grounded.[12]

Case 3. Case 2 may be regarded as a modal version of a temporal case involving the splitting of an amoeba, where it is natural to say that when the division occurs, the original amoeba ceases to exist and two new ones come into existence. It would be very strange to hold that, in fact, the parent amoeba survives the splitting and only one new amoeba comes into existence, and the strangeness of this view clearly derives from the impossibility of citing features in virtue of which the parent amoeba is identical to one rather than the other of the amoebae which results from the splitting.

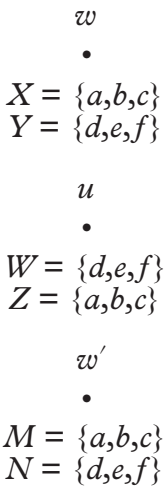
Case 4. Cases 2 and 3 are reminiscent of examples discussed by writers on personal identity. Suppose Oldman's brain has functionally equivalent hemispheres storing the same memories, realizing the same abilities and character traits, etc., and imagine that each hemisphere is transplanted into a new body giving rise to two

12. For a reader with Cartesian inclinations, this example can be reformulated to concern whether my body could have been that of one of a pair of identical twins.

individuals, Newman-1 and Newman-2 (Parfit [1984], pp. 309–12). Without appeal to some such entity as the soul, can we credibly maintain that one of Newman-1 and Newman-2 is identical to Oldman, while the other is not? Again, there are no features which could ground the putative identity, since the same features are realized in the two cases. This has not prevented the view that Oldman is identical to one or other of the new men from being held (Chisholm [1970]), but this primitive personal identity, not grounded in any feature we would normally regard as relevant to questions of personal identity, is certainly extremely puzzling.

The description of these cases is by no means the last word on the doctrine that for each instance of identity or failure of identity, there must be facts in virtue of which that instance obtains, for in the next chapter we will consider two cases which are apparently inconsistent with the doctrine, and in Chapter 7 we will deal with two more. Nevertheless, enough has been said to lend the idea some plausibility, so we end our discussion of sets by indicating what the defender of Membership Rigidity and Crossworld Extensionality can do about the second argument he needs, as characterized under (II) above. That is, suppose the sceptic agrees that there must be features in virtue of which the identity he postulates between X and W obtains, and cites the fact that these sets contain the passengers of the same flight (BA 127 on the 19th.) in the two worlds. How can this response be shown to be inadequate?

The problem for the sceptic is to produce consistent judgements about more complex cases without attributing absurd essential properties to sets. We will see the full extent of this problem in the next chapter but, for the moment, let us complicate our example by imagining a third world w' in which our six individuals all exist but no one travels anywhere on the 19th or the 21st. In this world, we may assume that the sets $\{a,b,c\}$ and $\{d,e,f\}$ both exist, so let us call them M and N. What should we say about the transworld identity relationships among X, Y, W, Z, M, and N? A sceptic who rejects (MR) because he thinks that it misidentifies the essential properties of sets may hold that M and N are not identical to any of the other four, on the grounds that it is essential to X and W to have members who travel on BA 167 on the 19th, and to Y and Z to have members who travel on the 21st. But such a sceptic simply confuses the set of travellers on that flight with the property of travelling on that flight; it might be incorrect to translate every-



CHAPTER 5: A MODAL THEORY

thing the sceptic says into what we regard as the truth by construing his word 'set' as our word 'property', and to say that he lacks the concept of set, for he may use the word as we do in non-modal discourse; but we can still regard him as conflating two distinct notions in his modal discourse.

On the other hand, if the sceptic does identify M and N with some of the others, regardless of which, then he is open to the objection that he has not really provided grounds for his judgement that $X = W$ and $Y = Z$. By making such identifications, he admits that being made up of passengers on the same flight is not necessary for identity with X; and if it is not sufficient, then obviously no grounds have been provided. And it seems unlikely that the use of contingent features of sets to provide conditions sufficient for transworld identity will lead to a coherent theory of these conditions. At best, we could make identity judgements on a case by case basis employing the contextually most salient features of the sets in the cases. It is hard to see how this position differs from one which denies both that transworld identity is a coherent notion and that there is a fact of the matter about the truth-values of *de re* modal sentences; so it would be incorrect to speak here of a theory of that which grounds transworld identity for sets.

It is obvious that these brief comments on scepticism about (MR) need elaboration, and we have not touched on (CE) at all yet. But rather than work through the relevant considerations in detail here, only to have to repeat them again in connection with entities of other categories, we will postpone elaboration of these fundamental ideas until our discussion of a more famous case in the next chapter. What we have argued here is that, *re* (I), requiring *grounds* for identity and non-identity across worlds is perfectly intelligible and well-motivated by examples, and, *re* (II), that it is not easy to provide such grounds for sets on views which reject (MR) and (CE).

CHAPTER 5: A MODAL THEORY

Chapter 6
The Necessity of Origin

Kripke's thesis

IN THE three lectures gathered together under the title 'Naming and Necessity', Kripke pursues at least two quite distinct topics. One concerns the proper account of the semantic relation of reference, while the other concerns metaphysical problems about essential properties of individuals. Because Kripke frames his own account of reference in modal terms, and uses examples involving possibilities to refute two rival accounts, it is not immediately obvious that his positions about these two issues are independent [Salmon 1981, *passim*]. But this should be clear at the present stage of our discussion; for instance, we have already seen that the necessity of identity says that one thing could not have been many, nor many one, and this has little to do with whether or not proper names are rigid designators.

Once we have a clear view of what the necessity of identity says, it appears to be a thesis hard to resist. But in the same discussion Kripke introduces another claim which is more properly termed essentialist,[1] and which is considerably more controversial:

> The question [is]…could the Queen – could this woman herself – have been born of different parents from the parents from whom she actually came? Could she, let's say, have been the daughter instead of Mr. and Mrs. Truman?…we can imagine discovering this…But let's suppose that such a discovery is not in fact the case. Let's suppose that the Queen really did come from these parents…The people whose body tis-

1. The necessity of identity is not really an essentialist thesis, at least if we follow Fine in identifying essentialism with the view that individuals may be *distinguished* by their necessary properties; *every* individual is necessarily identical with itself. See [Fine 1978b, pp. 288–9].

> sues are sources of the biological sperm and egg…Perhaps in some possible world Mr. and Mrs. Truman even had a child who became Queen of England and was even passed off as the child of other parents. This would still not be a situation in which this very woman whom we call 'Elizabeth II' was the child of Mr. and Mrs. Truman, or so it seems to me. [Kripke 1972, pp. 312–14]

It is probably these remarks, more than any other, which have brought about the renewed interest in essentialism in recent philosophy. (The title usually attached to Kripke's doctrine here, 'the necessity of origin', is a small misnomer on the strong interpretation of 'necessarily', if the Falsehood Principle is applied to 'is a child of', but 'the essentiality of origin' is a more awkward phrase.) If we generalize what Kripke says about the Queen, then he is arguing that the parents of any organism are essentially the parents of that organism. However, the identity of the parents of an organism, he says, is fixed by the identities of the bodies from which the sperm and egg come that give rise to the organism; hence it is no counterexample to his claim that a sperm-and-egg transplant would result in an organism having different parents, in one sense of 'parent'. But if it is no counterexample, then it is really the sperm and egg which matter: what is essential to the Queen is to come from the sperm and egg from which she actually came.

One cannot immediately extend this claim to every organism, since not all organisms are created by sexual reproduction. But they do all 'come from' some organic antecedent, which may be a single thing, such as the acorn from which an oak tree develops. For a general term to cover such antecedent entities, let us use the word 'propagule'; the oak tree's propagule is its acorn, while a human's propagule is his zygote, whose propagules are in turn the sperm and egg whose fusion that zygote is. Thus the relation 'x is a propagule of y', or 'Prop(x, y)' for short, can hold across either fission or fusion: we shall regard it as irreflexive, asymmetric and intransitive; and it is evidently a relation to which the Falsehood Principle applies, since an existent propagule cannot give rise to a non-existent, an existent cannot have a nonexistent propagule, and two non-existents at a world cannot enter into the biochemical reactions of development at that world which make one thing a propagule of another there. Using this relation, we can formulate a general version of Kripke's views about the Queen as an instance of the essentialist schema (S′) on page 95 in the following way:

(K) $\Box(\forall x)\Box(\forall y)\Box(\mathrm{Prop}(x, y) \rightarrow \Box(\mathrm{E}(y) \rightarrow \mathrm{Prop}(x, y)))$.

The reader should compare (K) with Membership Rigidity, and note that, since we are not yet proposing any analogue to Crossworld Extensionality, the essentialism embodied in (K) does not attribute individual essences to organisms: it is consistent with (K) that at some world some organism has exactly the propagules which some distinct organism has at some other world. In addition, it should be emphasized that the question whether or not (K) is true is quite independent of one's views about the nature of the self: it would be beside the point to dispute Kripke's claims about the Queen on the grounds that the Cartesian self who is the Queen could have inhabited any old body. Rather, a Cartesian should read Kripke's remarks as claims about the Queen's body, albeit infelicitously expressed.

In what follows, we will be concerned mainly with the justification of (K). As with (MR), we will argue, this time in some detail, that one who denies (K) must deny that facts about the transworld identity of organisms are always capable of being grounded, or else he must propound some other equally implausible conception of identity. Our arguments will apply, *mutatis mutandis*, to the defence of Membership Rigidity, and we will then turn to the question of finding some analogue of Crossworld Extensionality. In addition, rather than simply relying on the intuitive discomfort one feels with the examples of alleged ungrounded identities in the last section of the previous chapter, we will consider two cases which are harder to deal with from the point of view of someone who holds that identities must be grounded, and our treatment of these cases will support application of this principle about identity in defence of (K).

An unsuccessful defence of (K)

Kripke has not himself given a detailed argument for (K),[2] but others have attempted to do so, so we will begin this discussion by considering one of the best known accounts, due to Colin McGinn

2. This is what I wrote in the first edition, but it is misleading, since Kripke does give 'something like a proof' of a related principle about the matter of which a thing is composed. The proof is in endnote 56 of [Kripke 1972], where it is printed in 'inexplicably garbled' form (Kripke, [1980, p. 1]). An unreliable reader of endnotes, I was unaware of it until it appeared, corrected, in [Salmon 1979]. By then I had already devised my own, similar, argument about acorns and oak trees, which first appeared in Forbes [1980a].

[McGinn 1976]. McGinn's strategy is to assimilate the origin relation amongst organisms to the identity relation, so that the necessity of origin becomes a special case of the necessity of identity. We will argue that this assimilation is illegitimate.

The biological relations McGinn considers are those of *continuity* and *d-continuity*. Continuity is a temporal, transitive, one-one relation – a human being, for example, is continuous with his zygote. d-continuity is like continuity, except that it need not be one-one; for instance, an organism x is d-continuous with any number of organisms if these have undergone fusion to produce something with which x is continuous. To establish (K), it suffices to show that d-continuity is rigid, which in this context means that if a pair of objects is in its extension at one world where both of its members exist then that pair is in its extension at every world where the second member exists ('rigid' really ought to mean that the extension is the same at every world – thus '=' is rigid – or that a sequence of objects is in the extension either at every world where all the members of the sequence exist or at none such, but neither of these is quite what we want here).

McGinn argues that d-continuity is rigid in the following two steps:

Step 1. Continuity is necessary and sufficient for identity among organisms: a human and his zygote are one and the same; hence, since identity is rigid, so is continuity.

Step 2. d-continuity is like continuity in all relevant respects; so d-continuity is rigid as well.

The premise that continuity is necessary and sufficient for identity amongst organisms is obviously the crucial one, but it is not at all plausible. McGinn's reason for holding it is that he thinks it would be unmotivated to deny the identity of the zygote with the resulting adult since

> …adults are commonly identical with children, children with infants, infants with fetuses, and fetuses with zygotes. Any attempt to break the obvious biological continuity here would surely be arbitrary. [p. 133]

However, there is a non-arbitrary reason for denying the identity which does not involve denying any continuities. We have seen that the zygote reproduces by copying its own genetic content and then

by dividing itself symmetrically in two. We have also seen that it would be inconsistent with the repudiation of bare truths about identity through time to hold that the zygote is identical with one or other of the resulting daughter cells. Since it cannot be identical to both, it follows that the zygote ceases to exist upon the completion of replication. But then it contradicts Leibniz's Law to identify the zygote with the resulting adult (or even the resulting embryo), since adults (embryos) outlast their zygotes, indeed, do not even temporally overlap them.

This argument is not irresistible. It is a familiar point that the indiscernibility of identical s does not give us a criterion of identity which can be used to resolve hard cases, since it often happens that we must first ascertain whether a certain *F* is identical to a certain *G* before we can decide whether or not the *F* in question has a property the *G* has. Here we have a case in point. Someone who holds that the zygote and the adult are the same will say that, after replication, the zygote no longer exists in just the Pickwickian sense in which the child the adult once was no longer exists: the person who was the child still exists, as an adult now, and the person (or human being, if zygotes aren't persons) who was the zygote still exists, as an adult; just as it is *strictly* false to say that the child no longer exists (becoming an adult is not equivalent to dying) so it is false to say that the zygote no longer exists. That is, a defender of McGinn's position could claim that 'zygote' is a *phase* sortal for humans, rather than an ultimate sortal for organisms: each human being has a zygotehood which precedes his infancy and childhood, and the zygote, infant, child, and human are all one and the same thing.

It would certainly be unusual to use 'zygote' as a phase sortal, since in the mouths of biologists it is a predicate for a certain natural kind of cell, and it is an odd feature of McGinn's argument that it fails unless we stick to this idiosyncratic sense. But is there anything *wrong* with using 'zygote' as a phase sortal? There is certainly a disanalogy between the notion of zygotehood, on the one hand, and childhood, on the other, since zygotehood ends with the division, and hence ceasing to exist, of a specific entity, while there is no similar phenomenon in the passing away of childhood. Still, there is no *a priori* reason why phases should not end in the clear-cut way the purported phase of zygotehood does.

The real problem with this line of defence of McGinn is that

when we treat 'zygote' as a phase sortal to save the identity claim he makes, the argument no longer suffices to establish what it aims to establish, that it is essential to a given human being to have developed from the very same cell as that from which he actually developed: briefly, the phrase 'the zygote' will no longer refer to the entity, the propagule cell, to which it should refer if the argument for (K) is to succeed. For with talk of a phase of a human being there is an attendant distinction between the human and whatever makes him up during that phase. During childhood, the human (or his body, if you prefer to read (K) as a thesis about bodies instead of persons) is made up of certain cells, but, by Leibniz's Law, he is not identical to the sum of the cells which make him up at any moment or during any period of his childhood, since the human can, and usually does, outlast any such sum; this would be true even if, as is not the case, exactly the same cells make up the human at each moment of his childhood. So if a human has a zygotehood before his childhood, then, by the same reasoning, he is not identical to the sum of the cells making him up during that phase; that is, he is not identical to the single cell in question (nothing turns on its being one cell, for the same would be true if humans originated from a two-cell entity resulting from an association of sperm and egg in which each continues to exist as a separate entity). More specifically, if 'zygote' is used as a phase sortal, it will be correct to say that the human is identical with the relevant zygote, but the zygote to which he is identical is not the same thing as the propagule cell, for the cell ceases to exist when it divides, while the zygote does not; this is the crucial point – the cell is now an entity which constitutes the zygote, rather than the zygote itself, and the zygote no more ceases to exist when zygotehood ends than does the child when he becomes an adult.

So on this defence of McGinn it will be true, and therefore necessary, that the human is identical with the zygote he was – the human could not have been a different zygote, nor a different child – but for all that McGinn has said, the human could have originated from a different cell, since a different cell could have made him up when he was a zygote, just as different cells could have made him up when he was a child. However, the thesis which McGinn is supposed to be arguing for is the thesis that it is essential to the human to have originated from the very same propagule, i.e., the very same cell. So our conclusion is this: if we use 'zygote'

as a sortal for cells, then it is false that a human is identical to his zygote; but if we use it as a phase sortal for humans (as we may legitimately do, if we wish), then although it is true that a human is (necessarily) identical with the zygote he was, this is insufficient to establish the modal relationship to the propagule itself of which (K) speaks. Either way, McGinn's argument fails to support (K).

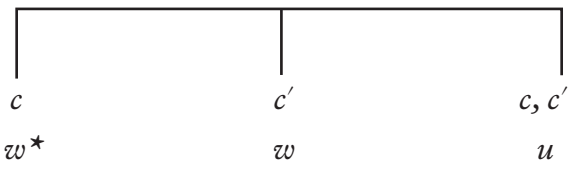
Finally, a more general objection to McGinn's justification of (K) is that it does not extend in any obvious way to explain essentialist principles which are intuitively of the same kind as (K). For instance, we cannot justify Membership Rigidity on the grounds that sets are somehow identical to their members (which one?). Since any acceptable account of (K) should employ the same resources as are required for the explanation of true principles analogous to (K) about entities of other categories, we should try to develop a defence of (K) which satisfies this condition.

The case of the moveable oak tree

The simplest way to bring out the principles underlying (K) is to investigate the consequences of denying that origin is essential. So let us consider the case of an oak tree, for instance the oak tree which stands in the cloisters of New College, Oxford, and the acorn *c* which is that oak tree's propagule. In what follows, we use 'the cloisters of New College, Oxford' as a rigid designator of a place. The most favored case for a sceptic about (K) seems to be the following. Imagine a possible world *w* in which there is an oak tree which grows in New College cloisters and which resembles the actual oak tree as far as is possible compatible with the supposition that this tree in *w* grew from an acorn *c′* distinct from *c*. Thus the *c′*-tree in *w* very quickly comes to be constituted of the same matter as the tree in the actual world, has the same morphology, etc. Suppose also that in *w*, *c* does not grow into any tree at all, or better, does not even exist. In sum, in the actual world there is a *c*-tree, in *w* a *c′*-tree, and the trees are indiscernible across these two worlds except only with respect to origin. Then according to the sceptic about (K), it is sheer dogmatism to insist that these trees are numerically distinct. Can the identity of the propagule acorn really have that much significance?

We shall show that it does. Let us agree on some straightforward and uncontroversial possibilities for the acorn *c*, the propagule of the tree in the cloisters in the actual world. Let us agree that *c* could have been planted on the other side of the clois-

ters and could have developed into an oak tree there. We make no assumption about the identity of such a tree, nor about how much or how little any such tree would resemble the *c*-tree in the actual world. So consider the class of worlds where this happens. Some of these will be otherwise almost indistinguishable from *w*, since the planting of the acorn *c* on one side of the cloisters does not render it impossible that *c′* is planted on the other side and grows into an oak tree exactly like the *c′*-tree in *w*. Choose one of these two-tree worlds, *u*, say. Thus in *u*, as in *w*, the *c′*-tree is the one which bears a high degree of resemblance to the *c*-tree in the actual world, while in *u* the *c*-tree need not bear much resemblance to the actual tree beyond having its propagule. If we link trees which have a very high degree of resemblance in almost every respect by a continuous line, and label trees by their acorns, we get this picture:



Here the *c*-tree in the actual world is linked to the *c′*-tree in *u* rather than the *c*-tree, reflecting overall resemblance rather than origin.

We may now press the sceptic about (K) to answer the following question: which, if any, of the two trees in *u* is identical to the actual tree?[3] This question has only three possible answers consistent with the formal properties of identity:

(i) the *c*-tree in *u* is the actual tree;
(ii) the *c′*-tree in *u* is the actual tree;
(iii) neither of these trees in *u* is the actual tree.

We will now argue that whichever answer the sceptic returns, his earlier claim that the *c′*-tree in *w* is the actual tree commits him to the existence of bare facts about transworld identity, or to some even less plausible view if he tries to qualify his position.

Suppose the sceptic returns answer (i), that the *c*-tree in *u* is the actual tree. Then, since the trees in *u* are distinct, the *c′*-tree in *u* is not the actual tree, and so the *c′*-tree in *u* is not the same tree as the

3. *Contra* J. L. Mackie, we do not have to make any assumption about which of the two trees is the *better* claimant; see Mackie [1974, p. 560].

c'-tree in w, which, according to the sceptic, *is* the actual tree. But there is no difference whatever between these c'-trees; they have the same propagule, and by choice of u, they have the same shape, the same location, the same matter, and so on. Hence the sceptic must posit a transworld numerical difference where there is nothing in virtue of which this difference obtains. His position is exactly like that of someone who holds that a set X existing at one world can have exactly the members there that a set Y has at another world and yet not be the same set as Y. On each position, a very plausible sufficient condition for transworld identity, being indistinguishable in every 'intrinsic' respect, is violated. For sets, this principle is just (CE), since the only intrinsic feature of a set is its membership, but for trees, the identity condition is actually somewhat weaker than the analogue of (CE) which we will endorse for organisms, since organisms have more intrinsic features than those which will figure in this principle. So, in effect, the sceptic about (K) is in a worse position than is someone who just rejects a certain account of the individual essences of organisms, since the identity principle he is in conflict with here is not so controversial as the yet-to-be-formulated principle analogous to (CE), which will give a stronger sufficient condition for transworld identity.

Under this pressure, the sceptic may retreat to answers (ii) or (iii), on which, indeed, no ungrounded identities or nonidentities will arise in connection with the three worlds just considered. But we can show that under very plausible assumptions, these three worlds may be supplemented with a fourth, whose existence the *antiessentialist* cannot consistently contest, and with this fourth world in the picture, ungrounded facts about identity reappear. The plausible assumption we need is that the c-tree in the actual world could have been just as the c-tree in u is, that is, it could have grown where the c-tree in u grows, and could have had that shape, that matter, etc. Although this assumption is hardly controversial, in the next section we will discuss the position of an essentialist who refuses to grant it (thereby making it essential to the actual tree to differ in some way from the c-tree of u). For the moment, let us grant it, and from the class of worlds where the actual tree is just like the c-tree in u, let us choose a world v which differs from u as little as possible compatible with there being no c'-tree in it, and no other tree in the cloisters. Using the same conventions as before, we may extend the diagram above to the following:

The *c*-trees of *u* and *v* are joined by a continuous line and are *completely* indistinguishable in all intrinsic respects, by choice of *v*. But, according to the sceptic, they are distinct trees, since by hypothesis the *c*-tree in *v* is the actual tree, while on either answer (ii) or answer (iii) to our original question about *u*, the *c*-tree in that world is not the actual tree. So, once more, the sceptic must posit a transworld nonidentity where there is nothing which grounds the difference, if he concedes that there are such worlds as *w*, *u*, and *v*. At the very least, then, we can defend (K) by saying that with a few other assumptions (K) is entailed by our intuition that identity must be grounded, the intuition we manifest if we agree that there is something wrong with the identity claims in the examples of the last section of Chapter 5.

Intrinsic and extrinsic grounding

The argument of the previous section is the central component in the defence of (K), but it uses one unexplained notion, that of an *intrinsic* feature of an entity. The need for some such concept can be brought out by noticing that there is a response to our argument which pays some attention to the principle that identities must be grounded, yet which allows (K) to be rejected. Someone could say that there is something in virtue of which the *c′*-trees in *w* and *u* are distinguishable, and something in virtue of which the *c*-trees in *u* and *v* are distinguishable, and hence something in virtue of which these are pairs of numerically distinct trees, for the *c′*-tree in *u* shares the cloisters with a *c*-tree while the *c′*-tree in *w* does not, and the *c*-tree in *u* shares the cloisters with a *c′*-tree while the *c*-tree in *v* does not. We wish to say that these are not "relevant" differentiating features because they are not intrinsic. But is this justifiable?

To hold that these relational differences between the trees are relevant to questions of identity is certainly unintuitive. How can what goes on concerning *another* acorn affect the identity of *this* tree? How can it be, for example, that if certain things had happened which actually did not, then a certain object would have existed, but if another *causally isolated process* had also occurred, everything else remaining the same (so far as is possible), no such

object would have existed? For instance, on answer (ii), the c'-tree in u is the actual tree, hence the c-tree in u is a distinct tree t. In v, something happens that prevents c' from developing into a tree, and as a result t does not exist in v, since the c-tree in v is the actual tree. In natural terminology, we can say that we are countenancing the possibility that transworld identity is not always intrinsically determined, but is sometimes *extrinsically* determined. Consider someone who holds that even if certain features of objects and events are causally isolated at w from a given object x existing at w, these features may still be relevant to questions about the transworld identity of x with objects at other worlds; we say that such a person holds that transworld identity is extrinsically determinable, or extrinsic, for short. In fact, the thesis that identity is intrinsic, or indeed the thesis that it is extrinsic, are qualifications of the thesis that it is grounded, in the sense that they impose constraints upon the conditions which may be regarded as grounding an identity or non-identity. Hence to say that identity is intrinsic is to say that whether or not the identity relation holds across worlds between x and y must be settled by intrinsic features of x and y at the relevant worlds (Quine's criteria of continuity of displacement, distortion and chemical change would be examples of intrinsic grounds for transtemporal identity).

We can convince ourselves that our concept of identity does not permit its applications to be extrinsically grounded not only by testing our intuitions against examples' but also by contrasting identity with relations which are explicitly extrinsically grounded. For instance, if the counterpart y at a world w of some object x were selected by the criterion of maximizing similarity, it would be an extrinsic relation, since whether or not a given object at w is a counterpart of x would turn on whether or not other objects at w are more similar to x than it is. It is the extrinsicness of this relation, ultimately, which explains why a counterpart-theoretic semantics based on it ascribes the wrong truth-values to certain modal judgements, as illustrated in the last section of Chapter 3.

The intuition that identity relations are not extrinsic is especially strong in the case of identity through time. Suppose that throughout a period of time we are continuously observing a scene in which, it seems to us, there is a continuously existing object of some sort F which is undergoing no perceptible change. If identity through time were extrinsic, this sensory information would not

even be *prima facie* evidence that the *F* observed at the beginning of the period was the same as the one observed at the end: we would also have to know how things were at the end of the period with causally unrelated *F*'s existing then.

There is a certain kind of alleged counterexample to these claims about transtemporal identity. Suppose the ownership of a church building changes hands and what was an Episcopalian chapel becomes a Buddhist monastery. The signing of the documents, etc., which constitutes the change in ownership, may be regarded as a process causally isolated from the building (suppose it takes place on another continent). Nevertheless, before the signing we have one church, and after it, another; so whether the same church exists throughout the period of time in question may be determined by causally isolated factors.

However, there is a fallacy in this argument, of the same kind as one of those diagnosed by Wiggins [1980, pp. 30–35] in an argument for the thesis that identity is relative. When we ask whether or not the same church exists throughout the period of time, we must decide whether 'same church' means 'same building' or, more strongly, 'same building used by the same religion'. If in the above story we mean just 'same building', there is no counter-example to the intrinsicness of identity, since the identity of the building (fixed by its location, design and the materials it is built from, say) does not alter when its ownership changes. And if we mean 'same building used by the same religion' there is still no counterexample, since the transfer of ownership is part of a process that results in the building being taken over by Buddhist monks.

The intrinsicness of transtemporal identity, like its property of being grounded, may also be supported by an example about personal identity. Consider again the case of the split brain operation (Case 4 on p. 127) and suppose that it is actually performed and gives rise to Newman-1 and Newman-2 from Oldman. The operation might have been performed differently: the half-brain which now sits in Newman-2's head could have been thrown away instead, so that only one person would have resulted, with the half brain now in Newman-1's head. Suppose someone agrees with us that neither of the new men is Oldman, for the reason that identity must be grounded, but says that if only one half-brain had been transplanted, the one in Newman-1's head, the resulting individual would have been Oldman. Then the actual Newman-1 is enti-

tled to think, "Thank goodness that other half-brain wasn't thrown away, otherwise I wouldn't have existed." But in this thought the non-Cartesian can only grope for the reference of 'I', the entity which would not have existed if Newman-2 had not; for the person who would have existed would have been exactly as Newman-1 actually is in every physical and psychological respect.[4]

If we are willing to reject any view which commits its holder to the possibility of an extrinsic grounding for a fact about identity or nonidentity, we are in a position to fill a *lacuna* in the argument of the previous section. It will be recalled that the sceptic who denied that the *c*-tree in *u* is the actual tree was shown to embrace bare truths about transworld identity by an argument which involved a world *v* at which the actual tree originates from its actual propagule at a place different from its place of origination in the actual world. Since a sceptic about (K) need not be an anti-essentialist, the defence of (K) could be blocked here by the doctrine that the place of origination of an organism is essential to it. Of course, this is highly counterintuitive, but the relevant question is whether a location essentialist ('L-essentialist') is making a *mistake*, or whether he is merely adopting some acceptable convention

4. In Chapter 1 of Nozick [1981], the author advances the 'closest continuer' theory of personal identity, which falls prey to this objection. For if only Newman-1 results from the operation, he is the closest continuer of Oldman, and therefore identical to him, while if both Newmen result, there is no unique closest continuer and so no identity. Rather than regard the objection as damaging to his theory, however, Nozick prefers to say that we are here uncovering an antinomy in our notion of personal identity (it is the concept, rather than the philosopher's theory of it, which is at fault!). He admits (p. 47) that this must seem very *ad hoc* and tries to dispel this reaction by drawing parallels between personal identity and other concepts for which there are both intrinsic and extrinsic analyses, where the latter appear superior. For example, Nozick holds that whether a belief is knowledge depends not just on the reliability of the methods by which the believer acquires the belief, but also on there being no non-reliable method whose role in the acquisition of the belief outweighs that of the reliable ones; so certain methods of belief-acquisition need not be sufficient for knowledge in a given case, even if in some other cases they do suffice, since those other cases satisfy the extrinsic condition that no additional method of a certain sort played such-and-such a role. But this comparison accomplishes nothing if personal identity just *is* a concept which requires an intrinsic account, as our intuitive reaction to the example in the text seems to imply. It is a significant difference between knowledge and personal identity that there is no analogous difficulty with an extrinsic account of knowledge: there is nothing problematic in saying that although Jones's belief that *p* is knowledge, if certain superstitions of his had played the main role in acquiring that belief, it would not have been knowledge. So the charge against Nozick that his antinomy claim is *ad hoc* still stands.

as an alternative to the one we happen to employ.

We find L-essentialism counterintuitive because of the lack of any very intimate connection between an organism and the place at which it originates; intuitively, someone who fixes transworld heirlines in terms of the locations of objects at worlds seems also to impute extrinsicness to transworld identity. But is there any connection between this sort of extrinsicness and the kind discussed above? There does seem to be a connection, a single underlying phenomenon, for we imagined the sceptic appealing to the properties of causally isolated entities to ground his identity claims, and the L-essentialist does the same. Certainly, the objects and features located at the place of an organism's conception need not be causally isolated from it, but these objects and features are quite distinct from the place itself. The place itself does not enter into any causal relation with the organism, because, at least on the prescientific conception of place, places are of a nature such that necessarily they do not enter into any causal relations whatsoever. So L-essentialism is incompatible with the condition that identity be intrinsically determined.

There are other maneuvers, of increasing complexity, which the sceptic about (K) may attempt. It is not very illuminating to pursue these, so one example will suffice. In our arguments against the sceptic, we have availed ourselves of worlds in which two propagules each give rise to organisms, and this suggests that there might be room for a qualified scepticism about (K) consistent with identity's being an intrinsically grounded relation. The sceptic could say that an organism can have one set of propagules in one world and a different set in another, if, but only if, there is some overlap between the two sets; then assuming the same propagule cannot function twice over, there will be no world in which each set independently gives rise to an organism. The simplest application of this qualified scepticism would be to creatures which reproduce sexually, such as humans, where there is an organism, the zygote, which has two propagules, the sperm and egg. Thus, if α is a zygote in w and β a zygote in u such that α's propagules in w are s and e while β's propagules in u are s' and e, then the sceptic could identify α and β. The problem he faces is that if he can make this identification, then he should be able to identify β with the zygote γ in a world u, where γ's propagules in u are s' and e'; but then, by his own principles, he cannot identify α and γ (contradicting the

transitivity of identity) since there are worlds where both *s* and *e* and *s′* and *e′* fuse to form zygotes. It is perhaps possible to pick and choose one's identity judgements to avoid outright contradiction here, but hardly while remaining faithful to the principle that identity judgements must be grounded.[5]

Essences and bare particulars

These arguments establish that one who subscribes to the principle that the facts about transworld identities and non-identities must be intrinsically grounded will also have to subscribe to (K), on pain of commitment to rebarbative consequences. Moreover, similar points can be made about sets in connection with Membership Rigidity (p. 107) and Crossworld Extensionality (p. 109). (CE) gives a sufficient condition for transworld identity between sets which grounds such identity in intrinsic features of sets (their membership), and since the effect of (MR) is to render the condition necessary as well, there can be no other intrinsic theory of the essences of sets which disagrees with this one, unless, as does not seem to be the case with sets, there is another family of intrinsic features which can be appealed to. More specifically, the reader who doubts the relevance of the arguments of the previous two sections to the case of sets will find that these arguments work effectively when instead of considering the trees which develop from the acorns *c* and *c′*, we consider the singleton sets whose sole members are, respectively, *c* and *c′*. If a sceptic about (MR) posits a transworld identity between these two singletons, then by introducing a world where they both exist in virtue of *c* and *c′* both existing, we may press against him the question about the identities of the singletons in that world. The possible answers are the analogues of (i), (ii), and (iii), on page 137, and analogous moves can be made against each answer. By intrinsicness, the identity of a set at a world cannot be sensitive to what other sets (outside the transitive closure[6] of the given one) exist at that world, and if the set/attribute distinction is to be made properly, the identity of a set cannot be sensitive to properties of its members, or its members' members, etc. Finally, a qualified scepticism which allows a many-membered set to change its members one at a time through a

5. See sections 4–6 of Forbes [1980b] for further details.

6. The 'transitive closure' of a set *x* is the set which contains all the members of *x*, all the members of the members of *x*, all the members of the members of the members of *x*, and so on, until we reach elements of *x* which are not sets.

sequence of worlds succumbs to the argument of the previous section, since we eventually reach a world where by the transitivity of identity we still have the same set, but none of its members at that world are members of it at the world with which we started, giving rise to ungrounded identity again.[7]

The parallel with sets highlights the fact that we have still to complete our account of the individual essences of organisms, for until we have a principle analogous to (CE), it is left open that distinct organisms at different worlds have the same propagules. Mechanically transcribing (CE) into an analogous principle would yield a principle we might label 'Propagule lndiscernibility' (PI): if x at world u has the same propagules as y at world v then $x = y$. But this is incorrect. It is presumably true that more or less anything can develop into more or less anything, given sufficiently sophisticated engineering, so taking the acorn c which grows into a certain oak tree in the actual world, we can consider a world where c is treated in such a way that it develops into a small vegetable. Then (PI) entails that that oak tree could have been, say, a cabbage, and therefore that there are entities which can be oak trees in some world and cabbages in others. But we lack any conception of what such entities could be: they seem unattractively similar to the scholastic notion of 'bare' individuals, subjects of properties which can be abstracted from all the properties which "inhere" in them. Bare individuals do not necessarily involve ungrounded transworld identity (see footnote 14), but are surely unintelligible in their own right. It would be possible to save (PI) by insisting that an acorn which grows into a small vegetable at a world *ipso facto* is not the same acorn as the actual one or, more strongly, that no propagule of a vegetable at one world can be a propagule of a tree at another. But it is unclear what could support this elimination of contingency in developmental outcome, since it may take only a very slight chemical change some time after the propagule has come into existence to produce the unnatural outcome.

The conception of a 'bare' individual arises in analogous temporal cases. Suppose a quantity of some polysaccharide is treated

7. Ungrounded identity also features in the case where $\{a\}$ and $\{b\}$ are identified, for some a and b which are not compossible. Here we cannot embarrass the sceptic with a world where both sets exist, granted that the existence of a set requires the existence of its members, but the sceptic will not be able to produce adequate grounds for the identity he posits.

in such a way that the sugar chains break down into their simple components. For instance, some cellulose decomposes into glucose. On one view, there is no transtemporal identity between the two quantities of substance in this example, but on another, there is: a single quantity of substance which is cellulose at one time becomes glucose at some later time. This second view appeals to the notion of an entity which can be different kinds of polysaccharide at different times, a notion which does not seem to answer to any concept we possess. Whatever the proper account of it is, there is a distinction we draw between changes which one and the same individual can undergo and changes which constitute the destruction of one individual and the creation of a new one; it may well be that this distinction is not completely defined, so that there are cases which its sense does not determine to be of one sort or of the other, but there are many more cases which the sense of the distinction settles: the zygote's mitotic division would be a case in point. The limits on changes which are changes in a single individual are marked by what Wiggins has called 'substance' concepts (or 'sortal' concepts),[8] and the fact that we do not have the idea of a bare individual manifests itself in the modal case as well as the temporal one, in the former by imposing limitations on what changes from world to world can be regarded as mapping out the contingent properties of a single individual; these limitations make the sort to which a thing belongs, or the kind of substance it is, essential to it. So we can modify (PI) to a principle of propagule-and-sort indiscernibility (PSI), which reads as follows:

8. See Wiggins [1980, Ch. 3]. I agree with much of what Wiggins has to say about identity through time, but not with his extension of his account to transworld identity. He holds that, given a sortal specification of what a thing is, we cannot conceive of that thing in a way which implies that it fails to satisfy the sortal, for this would be to conceive of it 'as having a different principle of individuation (different existence and persistence conditions) from its actual principle' (*op. cit.*, p. 122). But why can we not 'just suppose' that the oak tree could have been a cabbage? We need a theory according to which our conception of the thisness of an individual is formed in the temporal case and then projected to transworld identity, to fix the boundaries of significance on *de re* hypotheses about the individual. Note that Wiggins obtains his essentialism about sets by counting the nature of a set's membership as a component of its principle of individuation; but it seems to me to be of a piece with this position about sets that the spatio-temporal route of a material object be counted essential to it. See the second section of Chapter 9.

(PSI) If x is an organism at u with exactly $z_1,\ldots,z_k$ as propagules and y is an organism at v with exactly $z_{k+1},\ldots,z_{2k}$ as propagules then x and y are the same organism iff (i) $z_i = z_{k+j}$, $1 \leqslant i \leqslant k$ for some $j \leqslant k$, and (ii) the sort of x at u is the same as the sort of y at v.

(PSI) is deliberately vague, in that it does not specify exactly how the sort of an organism is to be defined. In the time of a single world, the same individual can undergo a change of sex, but it is less clear that an individual of one sex could have been, from the outset, an individual of another (again, Cartesians may take the individuals here to be just the bodies). This appears to be the kind of boundary problem which our concepts are not sufficiently well-defined to settle, so the full story about individual essences of organisms is correspondingly left unfinished. But the form of the account is quite clear, and we may leave matters there, since it would serve no useful purpose to fix a boundary by stipulation.

The branching conception of possible worlds

Whatever account of an individual's essence we give, we rely to some extent on a parallel with identity through time for support for the thesis that the account is genuinely sufficient for transworld identity; for one can always say, concerning any non-trivial condition we argue to be sufficient, that distinct possible objects satisfying the condition are conceivable. On such a view, for instance, it would be held that distinct possible organisms can have exactly the same origin and be of exactly the same species and sex, one in one world and the other in another. The view in question is known as *Haecceitism*, since it attributes to each individual a primitive identity or thisness, as opposed to the kind of essentialism defended above, according to which non-trivial conditions sufficient for transworld identity can be given.

Some philosophers might say that Haecceitism is incorrect as far as transtemporal identity is concerned, for it seems possible to give criteria for transtemporal identity which are both necessary and sufficient. For instance, a continuity account, a version of which was adverted to by Quine, has some degree of plausibility. The classical conception of transtemporal identity as spatio-temporal continuity may be stated as follows:

(C) For any sortal F and any objects x and y, x and y are the same F iff x is an F and y is an F and for any times t, t', if

> x exists at t and y at t', then for each time t'' between t and t' there is a region of space occupied by an F at t'' such that the interior of the sum of these regions (for all t'') is a continuous region of space.[9]

A Haecceitist about identity through time would have to claim that even if p is a continuous path through space which is occupied at every instant of the period of tracing by an F, nevertheless distinct portions of p may be occupied by distinct F's. And it seems natural to challenge both the modal and the temporal Haecceitist in the way we challenged the sceptics about (MR) and (K); it is for them to explain in what the distinctions they draw consist. Nevertheless, even if Haecceitism about identity through time is mistaken (we have still to see if it is), it may not be that Haecceitism about transworld identity is mistaken, and so far, for both transtemporal and transworld identity, we have only given some examples to prompt the intuition that a demand for grounds is always justified. It is time to consider some harder cases for our view about both kinds of identity. In this section, we will discuss an alleged example of ungrounded facts about the transworld relation, due to Robert Adams, and in the next, an apparent example of ungrounded facts about the transtemporal relation, due to Kripke.[10]

In Adams's example, we consider a world w in which there are two qualitatively indiscernible iron globes which have always and will always exist; that is all there is to w. But neither globe is essentially immortal, there are no restrictions on the times at which either globe could cease to exist, and the existence of either is in no way tied to the existence of the other. Thus there are worlds u and v just like w, except that in u one of the globes ceases to exist at a time t (before time ends, if it does) while in v it is the other globe which ceases to exist then (the assumption that the globes are indiscernibles in w is not essential, but simplifies the story).[11] In

9. Note that in (C), the continuity condition has to hold for every interval $[t, t']$ such that x exists at t and y at t'. Note also that (C) nowhere quantifies over "time-slices" or "instant-stages" of ordinary continuants. Rather, it embodies Wiggins' conception of transtemporal identity as spatio-temporal coincidence under a concept.

10. The Adams example is from Adams [1979, pp. 22–3]. For Kripke's example, I rely on Shoemaker's account in Shoemaker [1979, pp. 327–8].

11. To sidestep the issue of whether or not the intraworld Identity of Indiscernibles is true, the example may be changed in the following way: in the two-globe

this set-up, according to Adams, the facts about transworld identity are primitive, i.e., ungrounded, because any feature we might appeal to as sufficing for the identity of the globe in *u* with one of the globes in *w* also holds between the *u*-globe and the other globe in *w*; *mutatis mutandis* for the *v*-globe. And the numerical difference between the *u*-globe and the *v*-globe is consequently also ungrounded, for no intrinsic feature differentiates either of these globes in its world from the other in its.

But these conclusions are unwarranted, and are at odds with the natural way of thinking about the globes, on which we can explain the facts about transworld identity in terms of identity through time. That is, we think of *w* as a course of events and of *u* and *v* as courses of events "branching" from *w* at the time *t* when one globe ceases to exist in *u* and the other in *v*. Thus the transworld identities are explained by transtemporal identities across the branch-point at *t*. Before *t*, the very same course of events constitutes *w*, *u*, and *v*, and if we trace back in *u* from some point after *t* into *w*, and trace back in *v* from some point after *t* into *w*, we arrive at different globes; so the transworld difference between the *u*-globe and *v*-globe is explained by the intraworld numerical difference of the globes in *w* together with the branching conception of the worlds. This conception thus eliminates the appearance of ungroundedness in the facts about transworld identity in Adams' example.

We would like to generalize the branching conception of these three worlds to meet a certain objection to the theory of individual essence we have advanced. Although in giving the essence of an individual object we have not rendered the account trivial by appealing to the identity of that object itself, we have allowed nonqualitative properties to enter into essences; for instance, it is part of the essence of the actual oak tree in New College Cloisters to have originated from the acorn *c*, and no other. As a result of this, it may seem that we have not really shown that facts about trans-

world *w*, let the globes be differentiated by contingent properties. Then even if it is necessary that if the globes co-exist then they are differentiated by some property, it is consistent to postulate worlds *u* and *v*, *u* with one globe and *v* with the other, such that the postulated changes in contingent properties in the *u*-globe from *w* to *u*, and in the *v*-globe from *w* to *v*, yield the required cross-world indiscernibility of globes between *u* and *v*. In this situation, the interpretation of our understanding of the distinction between *u* and *v* advanced below is still applicable.

world identity are grounded, since the transworld identity conditions of the objects which enter into the essences of other objects may themselves be ungrounded. But the branching conception provides an assurance, for a wide range of categories of object, that this is not so, since, provided the objects which enter into the essences of other objects in some sense themselves "come from" yet other objects, in a way which eventually leads to a temporal regress, we will at some stage in this regress be able to explain all relevant transworld identities as transtemporal identities holding across a branch point, just as the identities in Adams' example were explained. The generalization of the branching conception we want is this: if *u* and *v* are worlds which at any time have some existent object in common, then *u* and *v* have some initial segment of their courses of history in common. In the light of this thesis, we see that the function of pairs of principles such as (K) and (PSI) is to enable all facts about transworld identity between *u* and *v* for objects which come into existence after the branch point, to be completely fixed by the content of the initial segment which *u* and *v* have in common (which may extend infinitely backwards in time).[12]

We can use Adams's example to refine the branching conception further. For instance if the globes in *w* are contingent existents, then there is a world *w′* which is just like *w* except that in it only one of the *w*-globes exists. It is then impossible for *w* and *w′* to have an initial segment in common, since at any time there are two globes in *w* and only one in *w′*, and so it would follow, by the generalized branching conception, that there is no world in which only one of those globes exists, contradicting our initial specification that the globes are not necessarily sometimes co-existent. To deal with this difficulty, we need the notion of a *separable* course of events in the history of a world *w*, a notion which will enable us to count amongst the worlds branching from *w*, worlds which consist in or extend a separable course of events in *w*. Causal isolation would be one criterion of separability, and assuming that the *w*-globes do not causally interact, our world *w′* will also be a world

12. The branching conception has been discussed in a number of places by Hintikka, although he would not agree with my view of its range of applicability. See, e.g., Hintikka [1975, Ch. 2]. My view raises the problem of transworld identity for times themselves, which I address briefly at pp. 84–5 of Forbes [1981].

branching from *w*. But other alleged worlds will be excluded, even when we have the notion of separability. For instance, it might be claimed that there is a world *u* which is just like *w* except that the iron which constitutes the two globes in *w* constitutes three globes in *u*, globes which, like the *w*-globes, have always existed. Here we reach the limits of intelligibility the branching conception imposes, as we come up once more against an ungrounded transworld identity, that between the quantities of iron in *w* and in *u*. On the view which we are presently defending, there is no such world as *u*.

We said above that pairs of essentialist principles function to allow the content of a common initial segment to fix transworld identities amongst later objects. Thus, any particular transworld identity will be grounded either by a transtemporal identity involving those objects themselves, as in Adams's case, or else by facts about ancestry and kind and transtemporal identities amongst entities at some earlier stage in an ancestral tree. So the essence of an object *x* involves those other objects through which we make the first step in tracing back to resolve a question about transworld identity for *x*. However, it is conceivable that an object lacks any such essence, for an object may be in a certain sense "simple". A simple object would be one which in no sense "comes from" any other objects, and if we say in Adams's example that the globes come from the quantities of iron which constitute them, then those quantities would themselves be examples of simple objects, objects without individual essences.[13] But it would be a mistake to infer from this that facts about transworld identity for simple objects may be ungrounded. In Adams' example, the facts about transworld identity for the quantities of iron are also fixed by transtemporal identities, and we rejected the coherence of the hypothesis of a world in which those quantities make up three globes rather than two, if it is stipulated that there have always been these three globes, as opposed to their having come about *via*, say, one of the *w*-globes dividing. Hence the theory of individual essence we are propounding applies only to categories of object whose members may be said to "come from" other objects in some fairly natural sense, as is exemplified by biological development or set-

13. We could appeal to the molecules or atoms which those quantities of iron are composed of, but unless atomism is *a priori* false, the conceptual problem of simple objects will eventually rearise.

theoretic containment. But our claim that identity is an intrinsically grounded relation is not restricted to objects of these categories, as is manifested by its applicability to simples.[14]

A problem about identity through time

The branching conception of possible worlds allows some cases of transworld identity to be directly grounded in transtemporal identities, but we have done no more to defend the view that *this* type of identity is an intrinsically grounded relation than, again, to give some examples of alleged ungrounded transtemporal identities and to point to their peculiarity. This strategy would be easy to outflank if there were quite straightforward examples of ungrounded identity through time and, according to Kripke, there are indeed such examples.[15] Consider a perfectly homogeneous sphere at a fixed location rotating with constant angular velocity through an interval of time $[t_1, t_2]$, and compare the following sequences of half-spheres. The first sequence s_1 consists of a half-sphere for each instant i in $[t_1,t_2]$, the half-sphere which at i occupies the region r occupied at t_1 by the eastern portion of the sphere. Since the sphere is rotating, no half-sphere will appear in this sequence more than once, unless the interval is long enough to allow the sphere to complete a revolution (assume not). The second sequence s_2 is the sequence of half-spheres which would have occupied the region r if the sphere had halted at t_1 and remained stationary through t_2, i.e., it is the constant sequence of a single half-sphere.

Since the sphere is in fact rotating, the half-sphere in r at t_1 is distinct from the half-sphere in r at t_2, but this transtemporal difference appears to be ungrounded. To see why, suppose we try

14. Thus someone who thinks that bare individuals can be abstracted from objects with properties need not contradict our thesis that identity is intrinsically grounded if he makes no *de re* judgements about such individuals which require for their truth that there be worlds in which the same bare individual exists only at times after these worlds diverge.

15. In the discussion which follows, the reader should distinguish two questions. One is whether Kripke has given a counterexample to the continuity account of transtemporal identity, and the other is whether he has given an example of a bare identity of a sort some analogue of which could arise in the modal case (if we were ingenious enough to think of it) to cause trouble for the principle upon which our defence of essentialism has been premised. The arguments in the text are intended to justify answering the second question in the negative, though they are admittedly less conclusive with respect to the first, in this context, less important, question. I am grateful to Christopher Peacocke for criticism of an earlier version of this section; the suggestion about atomism *ad fin* is based on a speculation of his.

to use the spatio-temporal continuity analysis of transtemporal identity, (C) on page 147, to explain the numerical difference of these two half-spheres. Then we find that they are actually *identified* by this account, since the sum of the regions occupied by the half-spheres at the associated instants i in the sequence s_1 is of course a continuous region: it is just the region r itself, and this region is continuously occupied by a halfsphere of the appropriate sort (fixed by the dimensions of r). At t_2, the half-sphere which was in r at t_1 is in some other region of space, and, of course, tracing that half-sphere also yields a spatio-temporally continuous path, the path determined by the constant sequence s_2, i.e., the path traced through the interval by the half-sphere which was in r at t_1. But what the account of transtemporal identity in (C) fails to do is to give us a reason to count this path as the path of a single half-sphere rather than the path consisting in just the region r, which we know would be the path of a single half-sphere only if the sphere had been stationary.[16]

A possible reaction to this case is to look for further features to ground transtemporal identity, features which get the case right, and to use these features to strengthen our criterion for transtemporal identity. It seems that the question of which path is the path of a single half-sphere depends upon the angular velocity or, more generally, upon the motion properties, of the half-spheres in each sequence. For instance, if at every i in $[t_1, t_2]$, each half-sphere in a sequence constructed like s_1 is at rest, then we know that the region determined by that sequence (r itself) is indeed the path "followed" by a single sphere during $[t_1, t_2]$; so perhaps we can add

16. Hirsch [1971] gives a set of rules for tracing the careers of objects through time under sortal concepts along a path P, one of which is the 'No Choice Requirement': there is no path P' such that F is instanced on every point of P', and P' is spatio-temporally continuous, and P and P' partly coincide and partly diverge (p. 36). But he wishes to allow that in cases where we do have a choice we may make one non-arbitrarily in accordance with the criterion of coherence with identity judgements by the other rules. However, Kripke's example does not require an extension of an already partly determined notion of identity: it is fundamental to that notion. So the No Choice Requirement could not be used to solve it. A similar remark applies to the suggestion that there is no fact of the matter about what the correct identity judgements are in the example. Another possible reaction is to query the genuineness of the sortals the example uses, 'half-sphere in region r at t'' etc., but this reaction also seems to me to lack credibility: we really have no difficulty in conceiving of the objects in terms of which the example is formulated, and in grasping the idea of a spatial route followed by such an object through an interval of time.

something to the continuity criterion which speaks of motion properties. But Kripke could fairly object that appeal to motion properties to ground facts about transtemporal identity is circular, since concepts of motion are defined in ways which require the application of transtemporal identity; the simplest example is that of the linear velocity of an object at a time, which is the limit of a sequence of average velocities, each average velocity being the distance travelled by that object during a certain interval, divided by the magnitude of that interval. Indeed, to speak of two groups of facts here, those about the transtemporal identity relationships amongst the half-spheres and those about the motion properties of the half-spheres, appears mistaken: there are just two different ways of formulating the same facts.[17]

Another possible reaction to the example is to say that our judgements of transtemporal identity are guided by how we think the half-sphere would behave, were the other half somehow taken away, yielding circumstances in which the continuity criterion by itself would give the right answer. But however this idea is worked up it seems to reverse the facts about what grounds what: we do not think that the truth of the counterfactual grounds the identity facts in the actual case; rather, the counterfactual about where that half would be is true because of the actual identity facts.

It would not be right to think that the example turns on some special feature of rotation, since the same problems arise if we consider a segment *g* of a homogeneous rigid rod moving through space on a straight path, and choose an interval of time at each moment of which some segment or other of the rod occupies the region which was occupied by *g* at the beginning of the interval. But by comparing this case with Kripke's, certain common features emerge. First, in each case the problematic objects are singled out by sortals which refer to regions of spacetime, which we do because the objects in question (half-spheres, rod segments) are not wholly circumscribed by physically proper boundaries;

17. Shoemaker [1979] tries to get round this point by defining motion concepts for spatio-temporally continuous sequences of instantaneous thing-stages. But to arrive at our genuine concepts of motion he has to be able to distinguish those sequences which make up continuants from those which do not, and to do this, as he himself recognizes, he has to rely on a notion of causal connection which is itself defined in a way which presumes upon the notion of transtemporal identity (see pp. 329–30 and 336–7). So the detour through sequences of thing-stages does not seem to help.

e.g., there is no physical mark of the distinction between one rod segment and another. It may be that such objects form a conceptually special category, for which there are non-*ad hoc* reasons to complicate the account of their transtemporal identity. Secondly, in the specifications of each example, we stipulate particular motion properties for some *other* object of which the problematic objects (half-spheres, rod segments) are *part*; and this suggests that if we may assume that the motion of the whole applies to the parts as well, we can derive consequences for the transtemporal identity relationships amongst the parts without circularity.

Unfortunately, the idea that the motion properties for wholes ground transtemporal identity for parts cannot be non-circularly implemented. Consider the case of the rod: from the fact that the rod has moved in such-and-such a way during a certain interval (the movement specified by the displacements of its end-points), nothing follows about which rod-segments singled out at one time are identical with which segments singled out at a later time, unless we know the facts about the relative spatial distances amongst the rod segments throughout the interval, facts which presuppose transtemporal identity for the rod segments. By specifying that the rod is rigid, of course, we fix what those facts are, but this specification just uses the notion we are trying to elucidate: it specifies, e.g., that the distance between *this* segment and *that* one is the same *throughout* the relevant period of motion.

Nevertheless, there is an urge to resist accepting that Kripke has given a case which refutes the thesis that identities and non-identities must be intrinsically grounded, since this would mean that the identities amongst half-spheres in his case are of the same sort as those in the examples of bare identities given earlier, for instance, the alleged identity between Oldman and one or other of the Newmen. But it seems clear that the true identity judgement about the half-spheres is capable of being supported in a way in which the judgements in our paradigm examples of bare identities are not: we simply do not find the former mysterious in the way we do the latter. What we should try to do, therefore, is to pin down exactly what the differentiating feature of the two cases amounts to, with a view to deriving some account of the intrinsic grounds of the identity facts in Kripke's case from that feature.

One who holds that the facts about the spatial routes traced by the parts through space during the relevant interval are themselves

bare facts posits a certain analogy between the two cases, for he could describe the case of persons as one in which the same self is in some kind of quasi-motion through the space of bodies, and as a matter of bare fact, moves from Oldman to Newman-1. Part of the reason why we are inclined to reject such an analogy is that ascription of Oldman's identity to Newman-1 has no consequences of either an actual or a dispositional nature in which the correctness of that ascription, as opposed to the other one, could manifest itself; nor do we understand what difference in initial conditions might bring about migration to Newman I *rather than* to Newman-2. But in the motion cases, different claims about the transtemporal identity conditions of the parts equate with different motion properties, and these differences can certainly be expected to have at least dispositional manifestations, the exact nature of which will depend on the laws of nature: consider, for instance, how we would expect an object to behave in a Newtonian universe were it to collide with a segment of the sphere, one which is rotating with a given angular velocity, rather than stationary. We also understand how initial conditions could differ, in terms of forces acting on bodies (in this case, torques), so that in one case we eventually get a stationary sphere, and in the other, a rotating one; admittedly, this does not apply to the Newtonian possibility of two worlds differing actualistically only with respect to the value of the constant angular velocity of a certain sphere, the sole occupant of each world, but in this case, whose prescientific intelligibility might be doubted, there are still the dispositional differences.

Might such dispositional facts be appealed to as the grounds of the identity facts? Normally, one would reject such an appeal, on the grounds that dispositional facts about an object must themselves have a categorical basis in the nature of the object or its current properties. But in the present example, where it is beginning to look as if there are no other candidates, there is a case to be made for grounding the identity facts in the dispositional ones. The undisputed datum is that we understand the distinction between a situation in which one account of transtemporal identity amongst the half-spheres is the right one and a situation in which another is the right one. Moreover, the terms in which this understanding is given are, *ipso facto*, those which will specify the required grounds. But the idea that there are no terms other than identity itself in which the understanding is given is difficult to

comprehend, for these identity questions are not open to being settled just by observation, and this makes it quite mysterious how we could come to have such an understanding of the difference between the two situations: to say that there is nothing in which our understanding consists seems little different, in this instance, from saying that our understanding is empty of content. So when pressed as to what it means to say that the identity facts are these rather than those, we may well turn to dispositional differences as constitutive of the distinction. The details of such differences are of course *a posteriori*, but the claim would just be that what it is for the identity facts to be one way rather than another is for the facts about what would happen were certain interactions involving the relevant object to occur, to be one way rather than another, according to whatever the laws of nature happen to be.

It may be objected to this that one cannot allow a range of actual facts about ordinary objects to be grounded in modal facts about them, and that anyway, some of the modal facts in question, laws of nature, are formulated for entities – persistents – of the very kind for which the difficulties with which we are grappling arise. But both these points can be met by pointing out that the objects for which the difficulties arise are rather special, in that they are homogeneous: this is required to ensure that the region of space occupied by a given part at the beginning of an interval in a case of motion is continuously occupied by an object of exactly the same dimensions as that which was there initially. And we might hold that for this special category of object (only), transtemporal identity is grounded in modal facts in addition to the continuity considerations embodied in (C). The view that homogeneity gives rise to a *conceptually* special case is at least worthy of consideration.

There is also another way in which Kripke's example might be conceptually special. Perhaps the relevance of the dispositional facts is limited to the question of how it can be told whether or not a homogeneous sphere is rotating, yet the facts about the identity conditions of the half-spheres can still be grounded: grounded in facts about the identity of other objects. The obvious response is that the same problems will arise for the other objects, but this response assumes that we never reach a level on which it does not make sense to say that the objects in question have parts; for the problem only arises when we consider parts of things. However, "simple" objects, or "atoms", are by definition objects which do

not have parts; this means, e.g., that the description 'the portion of atom *a* in region *r*' does not denote, if region *r* is a subregion of the region occupied by *a*. For atoms, at any rate, criterion (C) is a complete account of transtemporal identity; but then we can ground the transtemporal identity of the half-spheres in that of the atoms which make them up. To this it will be protested that the matter of the sphere need not be composed of atoms (is not, if this is implied by homogeneity): but this is just to say that the case is conceptually special, to the extent that our ordinary concept of matter is that of something composed of atoms. *A priori* atomism has rather sunk from view with the percolation of science into common knowledge, but the infinite divisibility of matter is not a hypothesis with which thinkers have been instantly comfortable, merely waiting for science to confirm or disconfirm it. This, then, would be another area of scope for maneuver with Kripke's case.

In conclusion, it should be pointed out that there is no pressing need for us to pursue attempts to undermine the *prima facie* Haecceitist moral of the example very much further, for the theses we are advancing depend only on the correctness of the doctrine that facts about trans*world* identity and non-identity must be intrinsically grounded. We have seen how the transtemporal facts in the case of the sphere manifest themselves dispositionally, and how a difference in causal antecedents may also be relevant to our grasp of the difference between the case of rotation and the case of non-rotation. This puts the example in a different class from the earlier examples of bare transtemporal identities, so that a counterexample to the sufficient conditions of transworld identity for sets and organisms we have endorsed, to be convincing, would have to have features like those of Kripke's case. But there is a difficulty in principle here: causal influences do not cross possible worlds, and dispositional facts are already modal facts. That is, there is nothing available in terms of which the identity in an alleged counterexample to our transworld sufficient conditions could manifest itself, or could come about. Such an identity would have to have the mysterious ineffability characteristic of alleged examples of genuinely bare identities, and we have seen no reason to take any such case seriously. Thus the doctrine on which we have based our defence of the necessity of origin stands.

Chapter 7
Fuzzy Essences and Degrees of Possibility

Two paradoxes

If we were to confine our attention solely to the cases of sets and their members, and organisms and their propagules, we would be encouraged to generalize from Membership Rigidity (*v.* page 107) and the Necessity of Origin, (K) on page 132, that it is essential to any non-simple object to come from those entities which it in fact comes from, or which it comes from in some world. However, this would be incorrect even for the case of certain organisms, ones which come from cells which do not function like propagules.[1] The slime mould is a tiny slug which is formed from the fusion of many single and largely indistinguishable amoebae [Ede 1978, pp. 9–14]. Each amoeba exists as a separate, independent individual for a while, reproducing by ordinary mitosis, but when enough are gathered together in one place, they assemble themselves together into a single organism which is not just a mere collection of amoebae, but rather a functionally differentiated creature which leads a life of its own. The trouble with the proposed generalizations of (ME) and (K) is that they would imply that each individual amoeba is essential to whichever slime mould it becomes a part of, but there is simply no intuition that any such relationship obtains. In advance of philosophical argument, most people would be willing to allow that a given slime mould could have been formed from a slightly different collection of amoebae; on the other hand, there would be much less agreement that a given slime mould could have been formed from a completely, or even very, different collection; and those who have these intuitions must therefore say that no one

1. 'Propagule' is explained on page 131.

constituent amoeba of a slime mould is essential to it, and yet some kind of essentiality of origin attaches to a sufficiently large proportion of these amoebae.

Artefacts yield a more familiar example of the same phenomenon. A fairly complex artefact, such as a watch, is made from a variety of components according to a particular design, and it is not very plausible to insist that each of a given watch's parts is essential to it, or that it could not have had a slightly different design. On the other hand, it does seem plausible to say that it could not have differed considerably in design or in the parts which make it up: then it would not be *this* watch any more. So what we are encountering here is a certain *vagueness* in the individual essences of entities which are made up of parts and constructed according to particular specifications. To specify the essences of such entities, we need to find some way of representing the thought that if an entity of this sort is made up (without leftovers) of parts from a given set, then as we consider sets of parts which have less and less in common with the given set, it becomes *less and less possible* for the entity to have been constructed from the set under consideration. In effect, then, we must find a way of introducing *degrees of possibility.*

There is more at stake here than merely a question about the scope of essentialist principles like (MR) and (K). In our remarks above, we have endorsed what we might call a tolerance principle about the haecceity or *thisness* of an artefact (of course, our use of the term 'haecceity' does not indicate agreement with Haecceitism). A general formulation of tolerance with respect to the parts of which an artefact is made is this:

(T) Necessarily, any artefact could have originated from a slightly different collection of parts from any one collection from which it could have originated.

The intuitive justification for the form of (T) is as follows. First, although we agree (let us assume) that, in fact, the same artefact could have been made from slightly different parts, we do not believe that there is some special property of actual artefacts or the actual world which makes this so: even if things had been different, and artefacts different from the actual ones had existed, there would still have been this tolerance. Hence the initial 'necessarily' in (T). Secondly, the formalized version of (T) will contain a condi-

tional with antecedent and consequent each governed by $\Diamond$, since the effect of (T) is to say that if some make-up is a possibility for some artefact, then some very slightly different make-up is also a possibility for it: if α is a possible artefact, then the schematic form of an instantiation of (T) is

$$\Box[\Diamond F\alpha \rightarrow \Diamond G\alpha].$$

This expresses the idea that the ground of truth of (T) lies wholly in the smallness of the quantity of change being contemplated; of course, this is only strictly true under the simplifying assumption that the 'importance' of a given part to an artefact is not weighted.

But however natural (T) appears to be, it is easy to see that it is in some tension with our doctrine that facts about identity must be intrinsically grounded, the doctrine upon which our defence of the essentialist theses of the previous chapters was based. For it is possible to use (T) to provide an apparent proof that there can be both bare transworld identities and bare nonidentities.

The argument for bare identity is due to Chisholm [1968], so we call it Chisholm's Paradox (it is this argument to which Quine is referring in our quotation on page 51, where he favors transtemporal over transworld identity because anything can be changed to anything by easy stages through some connecting series of possible worlds). Let $\langle w_1,\ldots,w_n\rangle$ be a sequence of worlds and let $\langle \alpha_1,\ldots,\alpha_n\rangle$ be a sequence of artefacts such that each α_i exists in w_i each α_i is constructed according to the same specifications and no α_i changes its parts through time (for the sake of simplicity, these last two conditions will be in force until further notice). Next, suppose that but for a very few components, each α_i is made from the same parts as α_{i+1}, yet the members of the pairs $\langle \alpha_i, \alpha_{i+1}\rangle$ differ from each other in such a way that as i increases so the number of parts α_i has in common with α_1 decreases, until we reach α_n, which has no parts in common with α_1. This set-up is a model of certain possibilities allowed by the tolerance principle: w_2 may be taken to be a world which realizes the possibility that α_1 is made of such and such parts, those which make up α_2; that is, $\alpha_1 = \alpha_2$. But then w_3, which by (T) may be taken to realize the corresponding possibility for α_2, thereby realizes a possibility for α_1, and so on, until we reach the conclusion that w_n realizes a possibility for α_1. But α_n is made of completely different parts from α_1, so this gives us our example of an identity which is a bare identity.[2]

$\Diamond\varphi_1(\alpha)$
$\Diamond\varphi_1(\alpha) \longrightarrow \Diamond\varphi_2(\alpha)$
$\vdots \qquad\qquad \vdots$
$\Diamond\varphi_{n-1}(\alpha) \longrightarrow \Diamond\varphi_n(\alpha)$
$\therefore\ \Diamond\varphi_n(\alpha)$

We can also give an exposition of the paradox without appeal to possible worlds. For each w_i, let φ_i be a predicate which says with rigid designators what parts α_i is made of in w_i, and let us replace 'α_1' with 'α' and treat w_1 as the actual world. Then the inference displayed in the margin is a classically sound argument, for its conditional premises are true by the tolerance principle, the minor premise is true since $\varphi_1(\alpha)$ is actually true, and the only rule of inference employed is *modus ponens*. So α could have been constructed from parts none of which feature in its actual construction. This gives us bare identity and thus a contradiction with our defence of (MR) and (K). Note also that, in resolving this paradox, we will be defending *de re* modality against Quine's criticism that you can change anything to anything by easy stages through some connecting series of worlds, for we will show that our doctrines about identity are consistent with the phenomenon of tolerance in the thisness or haecceities of certain sorts of things.

The argument for bare facts about transworld differences employs the same resources as Chisholm's Paradox.[3] Our intuition of tolerance in the haecceities of artefacts went along with an intuition that one and the same artefact could not be made from entirely different sets of parts in different worlds. According to this latter intuition, α_1 and α_n are indeed distinct things, so let this be granted. However, by the tolerance principle, there is a sequence of worlds σ_1 like the first half of $\langle w_1,\ldots,w_n\rangle$ and a sequence σ_2 like the second half, only in the reverse order, beginning with w_n, each sequence terminating in a world just like a particular world, say w_k, from the middle of the original sequence, such that in the last world of σ_2, α_n is just like α_k in w_k, and in the last world of σ_1, α_1 is just like α_k in w_k. Since α_1 and α_n are distinct, so are these worlds, but the only difference between them is in the identity of the artefacts they contain, and the difference between those artefacts is itself an ungrounded difference. So this example has the schematic form of the Eiffel Tower example (page 125) which we gave as an illustration of an unacceptable drawing of distinctions; hence, if the argument just given is a good one, we were wrong to regard

2. Bare, that is, relative to the simplifying assumptions of the case. A stronger case is obtained by allowing a small change in design from world to world as well.

3. This argument is developed in [Salmon 1979], where it is called 'The Four Worlds Paradox'.

that distinction as unacceptable. Following Salmon [1979] we call this paradox the Four Worlds Paradox, the four worlds being w_1, w_n, the last world of σ_1, and the last world of σ_n.

For a formulation of the paradox with modal operators, let us introduce the name 'β', whose reference is fixed by the description 'the artefact which would have resulted if…', completing the description by filling in the details of α_n's construction in w_n ('β' is a 'descriptive name'; see Evans [1979]). We have just agreed that β would not be α, but we can also give two arguments like Chisholm's Paradox, one which concludes $\Diamond\varphi_k(\alpha)$ and the other $\Diamond\varphi_k(\beta)$. Recalling that we are just now holding design constant, the truth of these two statements together with that of $\Box(\alpha \neq \beta)$, delivered by the necessity of identity, is inconsistent with the requirement that there must be something in virtue of which transworld differences obtain when they do.

Sorites paradoxes

The two modal paradoxes are Sorites paradoxes, that is, paradoxes of vagueness. This is especially easy to see in the case of Chisholm's Paradox, which is exactly like familiar Sorites paradoxes such as the Paradox of the Tall Man. Corresponding to the tolerance principle (T) for original constitution, we have a tolerance principle for height classification: someone only marginally taller than a short man is himself short. To be absolutely precise, one tolerance principle concerns the application conditions of a single predicate, 'is tall', whereas the other tolerance principle is really a family of principles, one for each artefact, and the role of the various men in the Tall Man is played, for each artefact, by the various actual and possible sets of parts from which an artefact of that design could be constructed. If α is an artefact, then the predicates whose application conditions are tolerant are predicates for the possible constitutions of α, that is, the predicates φ_i.

The analogue to the Four Worlds Paradox is obtained by starting with a man five feet in height and applying the tolerance principle for 'short' to conclude that a man of five foot six is short, and then by taking a six foot tall man and concluding from the analogous tolerance principle for 'tall' that a man of five foot six is tall. Here we have a bare difference in height classification: there is no difference between such men in which their difference in height status consists, and, in particular, there is no difference in their height. This kind of bare difference is indisputably ludicrous.

The classification of the modal paradoxes as Sorites paradoxes makes it desirable that the method of resolving them be an instance of a general strategy for resolving Sorites paradoxes. This immediately eliminates some proposed solutions. For instance, it is tempting to think that what is wrong with the argument of Chisholm's Paradox is that, as things actually are, the later worlds in the sequence do not represent possibilities for α_1, but if things had been as they are in w_4, say, then some of those later worlds would have represented possibilities for α_1. Hence it might be suggested that we can resolve the paradox by introducing an accessibility relation, on which some later worlds in the sequence which are accessible to w_4 are not accessible to w_1. But even without examining how the details of this proposal would be worked out, we can see that such a solution is quite *ad hoc*, and does not address the underlying source of the paradox, unless there is a sense in which tall men are not "accessible" to short men but are to men of medium height. Moreover, the accessibility solution applied to the Four Worlds Paradox entails that the last world of σ_1 is accessible to w_1, but the last world of σ_2 is not. It therefore requires us to distinguish between those worlds, a distinction which has the same problematic status as the one between their contained artefacts. So someone motivated to seek a solution to the Four Worlds Paradox because he does not wish to draw distinctions which mark no differences, could not be content with such a treatment of it.[4]

A short way with the modal paradoxes is simply to deny that the

4. The same difficulty afflicts an accessibility solution of Chisholm's Paradox. Suppose we say that w_n is not accessible to w_1 and hence that $\Diamond\phi_n(\alpha)$ is false at w_1, since there is then (we assume) no world accessible from w_1 in which α satisfies ϕ_n. Nevertheless, it is clearly possible, relative to w_1, that *something* be ϕ_n even if it is not α. More particularly, it is possible, relative to w_1, that things be exactly as they are in w_n but for the identity of whatever satisfies ϕ_n, and thus there is a world v, just like w_n but for the mere identity of a single object, a world which, unlike w_n, is accessible from w_1. But when we allow copies of worlds to multiply like this, we manufacture bare facts about the identity relation of the very kind we set out to remove by seeking a solution to the paradoxes. In [1981] Salmon presents a version of the accessibility solution which recognizes our point that Chisholm's Paradox is a paradox of vagueness by classifying the worlds in the Chisholm sequence in three ways, as either determinately possible (determinately accessible) relative to w_1, or as determinately impossible, or as neither. But this is no improvement on a two-valued solution which does not recognize the intermediate cases, since it is still saddled with the distinction between w_n and v.

tolerance principle (T) is true.[5] However, this solution again fails the test for being non-*ad hoc*, since its analogue for standard paradoxes is inadequate. Michael Dummett and Crispin Wright have convincingly argued that the tolerance in the application conditions of such predicates as 'is tall' is a consequence of the point of using them, which is to effect classifications of objects just on the basis of how those objects look. To give up the tolerance principles here would be to eliminate predicates with such a use from our language, since 'sharpening' such predicates would change their use radically. Predicates which are applied just on the basis of how things look cannot have strictly delimited ranges of application, because, as Wright puts it, 'if the conditions under which a predicate applies are to be generally memorable, [that predicate] cannot be unseated by changes too slight to be remembered' ([1975,

5. In Anil Gupta's book [1980] half a chapter is devoted to the modal paradoxes. Gupta argues (p. 103) that they are not paradoxes of vagueness, since genuine examples of the latter can be blocked by sharpening the tolerance of vague predicates in a way which does not block the modal paradoxes. But his reasoning appears to be fallacious. He points out that if we just stipulate that a man is not bald iff he has 106 hairs or more on his head, then Sorites reasoning will not show that every man is bald. This is correct. But next, he claims that no analogous stipulation will block Chisholm's Paradox and, to demonstrate this, chooses the stipulation that necessarily an artefact can have at most one different part at its origination. However, although it is correct that *this* stipulation does not block Chisholm's Paradox, Gupta has failed to compare like with like, for this stipulation is not analogous to the one about the number of hairs. On the latter stipulative solution, there are two possible states of the head, bald and nonbald, and the stipulation assigns each possible quantity of hair to one or other state, no quantity being assigned to both. But Gupta's stipulation about artefacts imposes no limit on the possible original constitutions for a fixed artefact α, since we can arrive at any constitution through a sufficiently long sequence of worlds. A stipulation in the modal case with the same effect as the one about baldness will assign each possible set of parts from which it is possible to construct an artefact according to α's design, to one of two sets, the members of the first set being those which are possible for α, the members of the second those which are not. Then one of the conditionals in the premises of Chisholm's Paradox will be false. Thus, if one chooses a stipulation for the modal case which really is analogous to Gupta's one about baldness, a perfect parallel is preserved. Having mistakenly concluded that Chisholm's Paradox is not like a standard Sorites paradox, Gupta then goes on to offer a solution to it (pp. 104–7). His idea is that the truth of a transworld identity judgement of the form '*a* is the same F as *b*' depends upon the world with respect to which the judgement is made; in Gupta's framework, metalanguage identity judgements are relativized to the worlds of the model, so that which judgements are true will depend on the point of view of the world at which they are evaluated. Insofar as I understand this suggestion, I read it as an indirect way of expressing some kind of counterpart-theoretic notion, with worldbound individuals and a nontransitive counterpart relation.

p. 337; see also Dummett [1975]). In other words, sharp observational predicates would be unlearnable if the phenomena to which they apply form a sensible continuum, as do colors, sizes, and so on. Another of Wright's examples involves predicates for stages of human life, such as 'infant', 'adolescent', etc. One who is an infant at time *t* is still an infant a few seconds later, but then no one ever reaches adulthood. Here the explanation of the tolerance is that with different stage classifications go explanatory distinctions and differences of moral and social status which a sufficiently small degree of development is too slight to support. Hence, in Wright's irresistible illustration, if we are forced to draw a sharp line, as we are in the matter of electoral qualifications, we do so 'with a sense of artificiality and absurdity'. And although it can hardly be used as an uncontested example, the predicate 'person' or 'bearer of a right to life' is surely another case, definitely applying to teenagers and definitely not applying to zygotes, and tolerant because small degrees of biological and psychological development cannot constitute the difference between a case in which they do and a case in which they do not apply, while large degrees of development do constitute such differences.[6]

6. Wertheimer [1971] tries to defend the conservative position about abortion from this suggestion. He writes (p. 81): 'The conservative points…to the similarities between each set of successive stages of fetal development…if this were the whole conservative argument…it would be open to the liberal's *reductio*…which says that if you go back as far as the zygote, the sperm and egg must also be persons. But in fact the conservative can stop at the zygote; fertilization does seem to be a non-arbitrary point marking the inception of a particular object.' But by completely parallel reasoning, we could start with a 10-year-old (by analogy, highly-developed fetus) who is uncontroversially a child (by analogy, person) and appeal to similarities between successive stages of development to conclude that the 70-year-old (by analogy, zygote) is a child, but resist the conclusion ("the liberal's *reductio*") that the corpse is a child, on the grounds that brain death does seem to be a non-arbitrary point marking the ceasing to exist of a particular object. Wertheimer could reject this analogy on the grounds that 'human being' is not a phase sortal, and so not analogous to 'child'. But what is the relevance of this difference to the claim that the analogy is a bad one? So far as I can see, it would only be relevant if it is supposed that ultimate sortals must have sharp ranges of application. Yet this supposition, if it is not justified on quite independent grounds, just begs the question, since the only candidate for a sharp beginning to the range of application is conception, and that a zygote is a human being is what Wertheimer is trying to prove; so he could not appeal to such a doctrine about ultimate sortals without circularity. Wertheimer goes on to say, 'It needs to be stressed here that we are talking about life and death on a colossal scale…so the situation contrasts sharply with that in which a society selects a date on which to confer certain legal rights'.

The distinction between what is possible and what is impossible for an object is as large a distinction as that between the tall and the short, one primary colour and another, or persons and non-persons, and therefore cannot turn upon a small degree of change in the respect relevant to making the difference. For artefacts, which respects are relevant is certainly open to dispute; one account might be that these respects are simply constitution and design, but it is arguable that we should add function to the list, and the period of time during which the artefact exists, and in the rather special case of works of art, perhaps also the identity of the creator. But we shall not try to argue for the correctness of any single account at this point, since the problem which faces us is to develop an apparatus which permits the appropriate respects, whatever they may be, to contribute to the individuality of an artefact in such a way that it can survive small amounts of change in these respects. For there is a substantial difference between artefacts and, say, sets, manifested by there being nothing corresponding to intraworld extensionality, (□E) on page 102, for artefacts. Sameness of parts is not sufficient for identity of artefacts at a world, since the very same parts may turn up at different times as the parts of artefacts with different designs and functions.

Another reason to insist upon the legitimacy of the idea of tolerant haecceities for artefacts appeals to intraworld transtemporal phenomena. One who would allow an artefact to survive replacement of a part within a world must allow transworld tolerance in original constitution, again on pain of laying down a sharp boundary on inappropriate conceptual terrain. Consider a sequence of worlds in which the time at which a particular part of an artefact is replaced by a certain new part is moved further and further back until we have a world in which the artificer is choosing which of the two parts to put in place in his original construction (we hold constant the stretch of time occupied by the lifespan of the artefact): it would be unmotivated to draw the line at this last world, admitting all the others. Someone might try to motivate such a line by insisting that until the artefact's construction is complete, it does not

---

But there is no contrast in any respect which tends to show that Wertheimer's reasoning does not fallaciously exploit the vagueness of a concept, as standard Sorites paradoxes do: an insane tyrant could turn the possession of any property which is in fact fuzzy into a matter of life and death.

exist, but then he has to say that when a part is removed, the artefact goes out of existence, since otherwise we would have a bare non-identity across time: just before the last part is put in place, a certain artefact does not exist, while once in place it does, so if that part is then removed again, we have intrinsically indistinguishable entities, and if one is not identical to the complete artefact, neither is the other. Thus someone who admits temporal but not modal tolerance, and tries to motivate his position, has to say that in having a part replaced an artefact goes out of existence and then comes back into existence; according to this legislation, it is strictly false to say of any artefact that it is (now) missing a part. Such a view is rather pointless, since the difficulties which arise from modal tolerance have their temporal analogues (see note 19), and if they can be handled in the temporal case, it seems unreasonable to refuse to extend the solution to the modal one.

We therefore conclude that the tolerance principles underlying the modal paradoxes are as inviolable as those underlying any Sorites paradox, and turn to the problem of extending the most reasonable solution of the standard paradoxes to these modal ones.[7]

The semantics of vagueness

How should we resolve standard Sorites paradoxes? It is very plausible that such paradoxes arise from the application of a semantic apparatus appropriate only for sharp predicates to languages or portions of natural language containing vague predicates. This view is in conscious opposition to the idea that vagueness arises from deficiency of meaning or is a source of incoherence;[8] rather, vague concepts are held to be legitimate and unproblematic as they stand, so long as we associate only the appropriate semantics with them.

The crucial notion of this semantics is that of the degree to which an object falls under a concept, or the degree to which a predicate applies to an object, and there is a familiar tradition of

7. Christopher Peacocke pointed out to me that the line of reasoning here does not yield the same objection to sharpening haecceity predicates as was made to the proposal to sharpen observational or stage predicates, namely, that such sharpening would nullify the point of having the predicates in the language. The conclusion he draws from this is that not every predicate of degree is one to which Wright-like considerations apply.

8. For deficiency of meaning see Fine [1975]; for source of incoherence see Dummett [1975] (for Wright, this is also a possible moral of the paradoxes).

semantics for vagueness using this notion, which finds perhaps its most sophisticated formulation in J. A. Goguen's [1969] logic for inexact concepts. We have already used the notion of degree above – indeed, it is hard to avoid its use in setting up examples of situations to which we are going to apply vague predicates – and we can formally introduce the notion from the use of vague predicates in the comparative form. If one of two men is taller than the other and the first but not the second is definitely tall, then the first is tall to a higher degree than the other, and so satisfies 'tall' to a higher degree than the other. Since satisfaction of predicates transforms into truth of sentences, it follows that the claim that the first man is tall has a higher degree of truth (is more true) than the claim that the second is. It is hard to find a well-motivated objection to any of these transitions, although it must be borne in mind that the resulting notion of the degree to which a person is tall is non-observational (cannot be told just by looking), unlike the question of his general height status (see further Peacocke [1981, p. 125]).

The suggestion is, then, that the familiar two-valued semantics be modified by including between its two values of absolute truth and absolute falsehood a range of intermediate degrees of truth, each of which is a possible semantic value for a sentence containing a vague expression. Since ordinary cases of vagueness often arise out of sensible continua, it seems reasonable to allow the degrees of truth to form a continuum; to begin with, then, the closed interval [0, 1] on the real line is a useful model of the set of degrees of truth, with 0 playing the role of absolute falsity and 1 the role of absolute truth. A model for a countable propositional language will therefore consist in an assignment of exactly one degree of truth from [0, 1] to each sentence letter, and the truth-value of any sentence can be computed as soon as we generalize the truth-tables for the connectives to the new degree-theoretic framework.

Noting that in two-valued logic a disjunction takes the better of the two values of its disjuncts, and a conjunction the worse of its conjuncts, we obtain the following clauses for degree-theoretic semantics:

(i) $\mathrm{Val}[A \vee B] = \mathrm{Max}\{\mathrm{Val}[A], \mathrm{Val}[B]\}$
(ii) $\mathrm{Val}[A \,\&\, B] = \mathrm{Min}\{\mathrm{Val}[A], \mathrm{Val}[B]\}$.

In two-valued logic, the value of a negation is the complement in

the two-membered set {0, 1} of the degree to which the negated sentence falls short of absolute truth. So for negation, we put

(iii) Val [~A] = 1 – Val[A].

Clauses (i), (ii), and (iii) together give us the usual interdefinabilities of & and ∨ for the truth-value interval [0, 1]. But the natural clause for → does not preserve its classical definability by the other connectives. Intuitively, we want the conditional to be material in a generalized sense, that is, it should be true if the consequent is at least as true as the antecedent, but we also want it to take values in the other cases which *reflect the gap* in degree of truth between antecedent and consequent. If the antecedent is only marginally more true than the consequent, the conditional should be only marginally less than wholly true, while if the antecedent is much more true than the consequent, the conditional should be considerably less than wholly true, with the limiting case being that of classical falsehood. The simplest clause which bestows these features on → is:

(iv) Val [$A \rightarrow B$] = 1 – (Val [A] – Val [B]) if Val[A] > Val [B]
= 1 otherwise.

For this system of propositional logic, we define a formula to be valid iff its value is 1 on any assignment of degrees of truth to its sentential letters, and we say that an argument is valid iff there is no assignment of degrees of truth to its sentence letters such that the value of the conclusion falls below that of the lowest-valued premise. More precisely, we say:

$$\Sigma \models A \text{ iff } \mathrm{Val}[A] \geqslant \bigwedge\{\mathrm{Val}[\sigma]: \sigma \in \Sigma\}$$

where '$\bigwedge\{\mathrm{Val}[\sigma]: \sigma \in \Sigma\}$' denotes the greatest lower bound of the values of the members of Σ, relative to the standard order of the reals (there may be no such thing as "the" lowest valued premise if there are infinitely many premises).

Clauses (i)–(iv) suffice for a resolution of standard Sorites paradoxes. Let $\langle a_1,\ldots,a_n\rangle$ be a sequence of men of increasing height such that the statement that a_1 is short is wholly true and the statement that a_n is short is wholly false, although there is only a marginal difference in height between adjacent men in the sequence. The tolerance of 'short' implies, with respect to the two-valued framework, that each conditional of the form:

a_i is short $\rightarrow$ a_{i+1} is short

a_1 is short
a_1 is short $\rightarrow$ a_2 is short
$\vdots$ $\vdots$
a_{n-1} is short $\rightarrow$ a_n is short

$\therefore$ a_n is short

is true. Hence the argument displayed in the margin is classically sound. But this is inconsistent with the fact that a_n is, say, six foot six. However, on the degree-theoretic framework, we see the argument in a different light. The problem is that *modus ponens* is an unreliable rule of inference in this framework, in a way that &-Elimination, for instance, is not: if Val$[A \rightarrow B] = 1$ then applications of *modus ponens* are unproblematic, but in our argument none of the conditionals is wholly true. In each, the degree of truth of the antecedent is marginally higher than the degree of truth of the consequent because each a_i is marginally shorter than the corresponding a_{i+1} (note again that even the comparative facts about degrees of truth need not be accessible to simple looking and seeing, since a marginal difference in height need not be observationally detectable). By clause (iv), therefore, each conditional is very slightly less than wholly true, and *modus ponens* is being used to detach consequents whose degrees of truth are dropping steadily towards 0. The paradoxical argument therefore illustrates a possibility implicit in the semantics, that from an absolute truth we may reason through a chain of conditionals each of which is almost wholly true and yet end up with a complete falsehood.

This is an elegant and appealing diagnosis of paradoxes of vagueness; it is because the conditionals are almost wholly true that the argument seems to us to be irresistible, and so, besides being neutralized, its persuasive force is explained. The only serious objections to this approach to vagueness involve what Fine has called 'penumbral connections',[9] which, if they obtain, are inconsistent with the fact that on clauses (i)–(iv), the degree of truth of a compound formula is always the same function of the degrees of truth of its component subformulae: in place of classical truth-functionality, we have degree-functionality. But Fine [1975, p. 26] has given a putative counterexample to degree-functionality, involving a conjunction whose degree of truth is claimed by Fine to be different from what clause (ii) says it should be. 'Is pink' and 'is red' are contraries (a penumbral connection), and hence, according to Fine, 'α is pink and α is red' must be wholly false. But if α is poised exactly midway between paradigm pink and para-

9. Many more equally serious objections have appeared in the literature, some of which I take up in the Postscript to this chapter.

digm red, then each conjunct has a middle degree of truth, approximately 0.5, which, if clause (ii) is to be believed, is passed on to the whole conjunction. Thus, Fine concludes, clause (ii) gives the wrong result in this kind of case.

However, this objection is unconvincing. To say that 'is pink' and 'is red' are contraries, from the degree-theoretic point of view, is not to say that nothing can be both, but rather, to say that nothing can be wholly pink and also wholly red. A thing can of course be red to a certain degree and pink to a concomitant degree. If in the situation of the example one man says 'α is pink' and another 'α is red' and neither is judged to have uttered something wholly false, why should this fate befall the first man if he anticipates and utters the second man's thought as well as his own, using 'and' to avoid an unnatural break in his speech? A reply of this kind can also be made to someone who holds that 'α is red and α is not red' should be wholly false, or that 'α is red or α is not red' should be wholly true. One reason (not Fine's) for ascribing these truth-values should certainly be rejected: someone might think that further investigation of an intermediate case of red would reveal whether or not the thing in question is really red. But this is not our view of vagueness, according to which it is an ineliminable feature of colour concepts, and not the product of some discriminatory limitation to which our sensory apparatus is subject.

If the conditional premises of a Sorites argument are not wholly true, what of the tolerance principles which justify those premises? Such principles are universal quantifications to the effect that if one thing is related thus and so to another then the second has a certain property if the first has it. So to be precise about the truth-values of these principles, we have to extend the degree-theoretic apparatus to quantifiers, i.e., to first-order logic. This will have the additional advantage of enabling us to see how the degree of truth of an atomic sentence such as 'α is short' is determined by the semantic properties of its constituent name and predicate. Our intuition was that a predicate like 'is short' is satisfied by different objects to different degrees, so as to specify its extension, we need to state not just which objects it applies to, but also, for each such object, the degree to which it applies. Following Goguen, we think of such an extension as a function from a set X of objects into the set $\mathcal{J}$ of degrees of truth (such functions are often called 'fuzzy sets', since they can be regarded as determining a set whose mem-

bers belong to it to different degrees – the set of short things would be an example). Note that, on this approach, vagueness resides entirely in concepts. The objects in X are perfectly determinate and the fuzzy sets themselves also have exact identity conditions: sets of this sort are the same iff the same things are members of each to the same degree.

More generally, if F is an n-place atomic predicate then we assign to F a function ξ_F from a set X of n-tuples of objects drawn from a domain D into a set $\mathfrak{J}$ of degrees of truth (X is called the 'universe' of F). Then for atomic sentences we have:

(v) $\mathrm{Val}[F(t_1,\ldots,t_n)] = \xi_F(\langle \mathrm{Ref}(t_1),\ldots,\mathrm{Ref}(t_n)\rangle)$.

Existential and universal quantification can always be thought of as equivalent to infinitary disjunction and conjunction, so the quantifier clauses involve the infinitary analogues of Max and Min, the least upper bound ($\bigvee$) and greatest lower bound ($\bigwedge$) operations:

(vi) $\mathrm{Val}[(\exists v)Av] = \bigvee\{\mathrm{Val}[A(a/v)]$: all a in $D\}$

and

(vii) $\mathrm{Val}[(\forall v)Av] = \bigwedge\{\mathrm{Val}[A(a/v)]$: all a in $D\}$.

We can now see that no tolerance principle is wholly true; for instance, a version of the principle for the Tall Man is

$(\forall x)(\forall y)(\mathrm{Short}(x)$ & y is one centimeter taller than $x \rightarrow \mathrm{Short}(y))$.

Any instance of the conditional matrix with 'a' for 'x' and 'b' for 'y' is either wholly true (because b is not one centimeter taller than a or because 'Short(a)' and 'Short(b)' are both wholly true or wholly false) or else slightly less than wholly true, because 'Short(b)' is slightly less true than 'Short(a)'. Thus, by (vii), the value of the universally quantified sentence is slightly less than wholly true, and can be brought closer and closer to absolute truth by taking smaller and smaller differences in height. Hence our earlier insistence that tolerance principles are true requires qualification when we move out of the two-valued framework: they are merely almost wholly true. But this in itself is sufficient to show that it would be absurd to deny them.

Closgs and counterparts

Our goal is to extend the best resolution of standard Sorites paradoxes to the modal paradoxes, so the next step is to import the degree-theoretic apparatus into the modal logical framework. There are two obstacles to be overcome at this point, one technical and the other philosophical. We take the technical problem first.

(I) When we compare artefacts across worlds, we assess degrees of similarity in at least two respects, constitution and design; in our presentation of the modal paradoxes, we only allowed constitution to vary, but this was an artificial restriction. Suppose we now consider two artefacts β and γ in a world w and ask how similar they are (as they are in w) to some artefact α as it is in the actual world. Perhaps β is close in design to α but not in constitution, while the converse is true for γ. So with each of β and γ we can associate a pair of numbers, measuring degree of similarity to α in each of the two respects. Yet there does not seem to be any reason why these two numbers have to be resolvable into a single number giving overall degree of similarity to α so that β and γ can be compared by that yardstick. But if only pairs of numbers are available, β and γ may be incomparable in respect of overall similarity, and in such a case [0, 1] would be an inadequate model of the set of degrees, since it is totally ordered by $\leqslant$.

However, this technical difficulty can be overcome. If there is no fact of the matter about which of x or y possesses the greater or lesser degree of similarity to z because the degrees of similarity which they do possess are incomparable in respect of which is the greater or lesser, then the degrees of truth of two atomic sentences of the form 't_1 is similar to t_2' may also be incomparable in respect of which is the greater or lesser, and thus [0, 1] fails to model the degrees of truth, since any two numbers in [0, 1] are, of course, comparable in respect of which is the greater or lesser. But it is easy to find mathematical objects for which this condition does not hold. Suppose degree of similarity is given by a pair of numbers each of which is in [0, 1], and each of which measures a single aspect of similarity (imagine only two aspects are relevant in the examples under discussion). Then we might use $[0, 1] \times [0, 1]$, the set of pairs of reals from [0, 1], as a model of the set of degrees of truth, since we can define 'less than or equal to' as:

$$\langle a,b\rangle \leqslant \langle c,d\rangle \text{ iff } a \leqslant c \text{ and } b \leqslant d$$

where the symbol $\leqslant$ on the right-hand-side stands for the normal

order on the reals. This is an example of a *component-wise extension* of a relation or operation: if the relation or operation is defined for single objects, then we define its application to n-tuples of these objects in terms of its application to the objects which occupy the jth positions of each n-tuple, $1 \leqslant j \leqslant n$. And with this object as our model of the set of degrees of truth, we do obtain incomparable degrees; for example, (2, 3) and (3, 2) are incomparable in terms of which is the "lesser", given our definition which extends this notion to pairs of reals. With such a model of the degrees of truth, there is no need to change any of the connective clauses (i)-(vii) above in a fundamental way, since the arithmetical operations they employ can also be extended component-wise to $[0, 1] \times [0, 1]$. To extend clause (iv) for the conditional, define A_i (B_i)to be the i'th component of Val[A] (Val[B]), and $c(A_i, B_i)$ to be $1 - (A_i - B_i)$ if $A_i > B_i$, 1 otherwise. Then

(iv)′ Val $[A \rightarrow B] = \langle c(A_1, B_1), c(A_2, B_2)\rangle$.[10]

Obviously, the restriction to two respects of similarity in the above discussion is inessential: for any n, we can admit n respects of similarity and choose $[0, 1] \times \ldots \times [0, 1]$ (n times) as our model of the degrees of truth, extending arithmetical notions component-wise. In particular, we say

$$\langle a_1,\ldots,a_n\rangle \leqslant \langle b_1,\ldots,b_n\rangle \text{ iff } a_i \leqslant b_i,\ 1 \leqslant i \leqslant n.$$

It turns out, however, that even such sets as finite products of the unit interval have structural properties which are unnecessary to model degrees of truth and to permit the definition of reasonable clauses for the connectives. In algebraic investigations of this question, Goguen has shown that, for a logic of vagueness, the minimum acceptable structural requirement on the set of degrees is that it have the order type of a complete lattice-ordered semi-group ('closg') in which the lattice maximum is identity for the group operation $*$. Fortunately, to understand what is to come, readers will be pleased to learn that they need only keep in mind two examples of closgs, $[0, 1]$ as above and $[0, 1] \times [0, 1]$, or perhaps more generally, $[0, 1] \times \ldots \times [0, 1]$. It is therefore unnecessary to pursue formal questions here any further, or to explain the terminology involved in Goguen's acronym.

10. Here I correct an error in the first edition pointed out by Tim Williamson.

(II) So much for the technical difficulty in extending degree-theory to modal logic. The philosophical difficulty concerns the coherence of the notion of the degree to which an object satisfies a composition predicate at a world. In the standard semantics for S5, transworld heirlines for objects are given by real crossworld identities in the model. So how could it be a matter of degree whether or not an object α satisfies a composition predicate ϕ at a world *w*? These predicates are not *themselves* predicates of degree: α either has *p* as a part at *w* or it does not. If it is to be true that α could have had the composition ϕ only to a certain degree, then the only other places that vagueness might be located is in the interpretation of the modal operator corresponding to 'could have' or else in the interpretation of the term α. Locating the vagueness in the semantics of the modal operators is certainly one option (see Salmon [1981, Appendix I; 1986]) but an alternative, which I argue to be superior in the Postscript to this chapter, is to adopt a counterpart-theoretic semantics on which, in evaluating a formula containing a term *t* at a world, we consider how things are with the counterparts of the referent of *t* rather than the referent of *t* itself. In counterpart theory, transworld heirlines are given not by real crossworld identities, but rather by some other transworld relation which – by contrast with identity – it may make good sense to regard as a relation of degree. Using counterpart theory, we can inject tolerance into the general concept of being a possibility or necessity for an object α.[11]

Although we have already dismissed a number of objections to counterpart theory, our discussion in Chapter 4 left two issues unresolved. One was the problem of defining counterparthood in a way different from Lewis's definition of overall similarity, so that certain intuitively true modal judgements are ascribed the right truth-value. The other was the problem of dealing with the necessity of identity. The difficulty we mentioned is that if we define similarity so as to obtain pleasing consequences, we introduce an element of the ungrounded if the definition is stipulative at any point, and so we fail to meet Quine's challenge to *elucidate* a relation which can be used to make sense of *de re* modality; or else we

11. I am grateful to Nathan Salmon for observations about the corresponding discussion in the first edition of the book, which have led to what I hope is a better exposition of the rationale for introducing counterpart theory as a solution to the current difficulties.

end up with a completely elucidated relation which is structurally isomorphic to a transworld identity relation, and the difference between counterpart-theoretic and standard semantics collapses.

We are now in a position to see how this dilemma is to be avoided. In connection with artefacts, degree of counterparthood is fixed by degree of similarity in particular respects, including composition and design, but *not* including the subsequent history of an artefact, involving who owns it and the path it traces through space and time. It is thus possible for an artefact to have had a very different history from the one it actually has, while some other artefact has a career rather like that of the given artefact. So our counterpart relation, which admits of degrees, is completely grounded in facts about crossworld similarity, but only facts of a particular sort. Furthermore, it is in no way isomorphic to identity. This is not merely because it admits of degrees, but because it may be many-one or one-many between a pair of worlds. This fact raises our worry about the necessity of identity (formulae (20)-(23) on pp. 65–66, (20) reproduced as (1) below) which will certainly fail in the present system. Consider

(1) $a = b \rightarrow \Box(a = b)$

which translates as

(2) $a = b \rightarrow (\forall w)(\forall x)(\forall y)(Cxaw \ \& \ Cybw \rightarrow x = y)$.

Suppose a and b are the same, and that w is a world where there are two artefacts c and d such that c and d are of the same design as a and each has half of a's parts. Other things equal, this makes each a counterpart of a (hence of b) to degree 0.5, and thus the conditional

(3) $(Ccaw \ \& \ Cdbw) \rightarrow c = d$

has degree of truth 0.5, since its antecedent has degree of truth 0.5 and its consequent is wholly false. Hence (1) is invalid on our approach; that is, it is not wholly true on every assignment of degrees.

If (1) is invalid, what becomes of the argument for it from $(\forall x)\Box(x = x)$ and Leibniz's Law ((23)–(25), pp. 66–67)? In fact, it is not the Law which causes the problem, but rather the premise $(\forall x)\Box(x = x)$. Consider the sentence $\Box(a = a)$. In our evaluation clauses for $\Box$ and $\Diamond$, (xi-a)–(xii-b) on pp. 60–61, we use t_i for the

i'th individual constant *token* in an object language expression, and there are two such constant tokens in $\Box(a = a)$, even though they are both of the same type. Thus, by evaluation clause (xii-a), the truth of $\Box(a = a)$ at a world requires that every world w, if c is a counterpart of a at w and d is a counterpart of a at w, then $c = d$ is true at w. We have still to reformulate those clauses to allow for degrees intermediate between 0 and 1, but we can already see, in virtue of the example just given, that $c = d$ may be wholly false although it is not wholly false that c is a counterpart of a at w, nor that b is a counterpart of a at w. The same point is perhaps more obvious if we make it in terms of translation rather than evaluation, since the translation of $\Box(a = a)$ is

$$(\forall w)(\forall x)(\forall y)(Cxaw \,\&\, Cyaw \rightarrow x = y)$$

which, in the light of our example above, is obviously not valid counterpart-theoretically.[12]

Thus while every *de dicto* modal thesis about identity has the same truth-value in the present framework as it has in the classical framework, a difference emerges over the *de re*, not because identity somehow becomes fuzzy, but because *de re* sentences introduce a new fuzzy relation, that of counterparthood, which in turn gives rise to degrees of possibility. What then of Kripke's claims that the necessity of identity is intuitively valid? The examples used by Kripke to invoke the intuition that such formulae as (1) are valid usually involve objects, such as planets and people, to which the notion of part has no very natural application. So we can respond to Kripke that when one considers entities of other categories, there may no longer be an intuition in favour of the Necessity of Identity. For instance, let α and β be clocks of identical design on opposite walls of a room and imagine a possible world in which there is only one clock in the room, made out of half of α's parts (strictly, counterparts of these) and half of β's, but with the same design as the actual clocks. Such a state of affairs is evidently

12. If we require for the truth of $\Box(a = a)$ only that at every world w, every counterpart of a at w is identical to itself, then what Kripke claims to be an instance of Leibniz's Law, $a = b \rightarrow (\Box(a = a) \leftrightarrow \Box(a = b))$, would turn out not to be, since by the same principle its translation would be

$a = b \rightarrow [(\forall w)(\forall x)(Cxaw \rightarrow x = x) \leftrightarrow (\forall w)(\forall x)(\forall y)(Cxaw \& Cybw \rightarrow x = y)]$.

However, I prefer to retain for the modal language the standard syntactic notion of being a substitution-instance of a schema.

genuinely possible, in virtue of which the present approach dictates that 'These two actual clocks could have been a single clock' is not wholly false. However, it is simply untrue that there is a firm pretheoretic intuition that this result is unacceptable.

There is one further objection to the use of counterpart theory which we ought to consider. In Chapter 3, following Hazen, we accused Kripke and Plantinga of confusing object and metalanguage in objecting that Counterpart Theory misrepresents the contents of modal judgements. Nathan Salmon has attempted to give a more sophisticated version of this objection which escapes Hazen's refutation of it.[13] Salmon's idea is to put the counterpart relation into the modal object language and then to compare the counterpart-theoretic truth-conditions of 'it is possible that *a* is *F*' with 'it is possible that some counterpart of *a* is *F*'. But there is no improvement of the Kripke/Plantinga objection to be extracted from this line of thought. Salmon claims that "intuitively", the second modal sentence,

(4) $\Diamond(\exists x)[Cxa \,\&\, Fx]$

is weaker than the first,

(5) $\Diamond Fa$

yet, according to Counterpart Theory, (4) entails (5); therefore Counterpart Theory misrepresents the content of (5).[14] But (5) is a logical consequence of (4) only when the two-place predicate *C* is treated as a logical constant of a certain unfamiliar and highly technical sort. So Hazen's reply to Kripke and Plantinga, that pretheoretic intuitions are not in question, applies here too. Hence all philosophical obstacles in the way of using counterpart theory to resolve the modal paradoxes may be overcome.

13. Salmon [1981, p. 235]. The discussion which follows has benefited from helpful remarks Salmon made in response to some earlier, ineffective, criticisms of his position which I made.

14. The translation of (4) is $(\exists w)(\exists z)(Czaw \,\&\, (\exists x)(Exw \,\&\, Cxaw \,\&\, Fxw))$ and the translation of (5) is $(\exists w)(\exists x)(Cxaw \,\&\, Fxw)$. (4) entails (5) because of the condition in Counterpart Theory that each thing is its own unique counterpart both at the world where it exists and at the worlds where it has no existent counterpart.

Counterpart theory with degrees of possibility

We now give a brief but rigorous formal description of counterpart theory with degrees, by defining a degree-model **M** for the language L_c of counterpart theory, which, in combination with a translation scheme for mapping L_m formulae into L_c formulae, will enable us to exhibit precisely the invalidity of the modal arguments which constitute the paradoxes. An equivalent way of proceeding would be to give another model theory for L_m, one in which formulae take various degrees of truth in addition to 1 and 0, and such a theory can be read off from the definition of degree model for L_c; but translations are slightly easier to work with.

The only vague predicate of L_c will be the counterpart predicate, so its extension will be given as a fuzzy set, that is, as a function into the set $\mathcal{J}$ of degrees of truth, which will in turn be an arbitrary closg. However, in order to treat all predicates uniformly we will also specify the extensions of the exact predicates as functions. An n-tuple of objects either definitely is or definitely is not in the extension of an exact n-place predicate, and thus the extension of such a predicate can be given by a function which maps its members to the value $\underline{1}$, the maximum of $\mathcal{J}$, and its non-members to the value $\underline{0}$, the minimum of $\mathcal{J}$. The language L_c contains two sorts of terms, including the constant $w^\star$ of sort 1, all the constants of L_m, which are of sort 2 in L_c, and for each n-place predicate of L_m an $n + 1$-place predicate whose last place is reserved for a term of a sort 1. So the existence predicate $E\_$ of L_m is correlated with $E\_, \_$, a predicate of sort (2,1). In addition, L_c contains a three-place predicate $C\_, \_, \_$ of category $\langle 2,2,1\rangle$, which is read as '_ is a counterpart of _ at _'. The two-sorted language is for ease of readability while, as before, the three-place counterpart predicate is needed to obtain a correct logic of existence.

A degree-model M for L is a two-sorted 9-tuple

$$(W, D, \mathcal{J}, Q, R, I, H, w^\star, val)$$

where W is a set of entities of the first sort (worlds) and D is a set of individuals. $\mathcal{J}$ is a closg (see previous section) whose elements are k-tuples of real numbers from the interval [0, 1], for whatever fixed k is the number of distinct criteria by which counterparthood is assessed. The lattice ordering is defined componentwise and the group operation $*$ by

(viii) $\langle a_1,\ldots,a_k\rangle * \langle b_1,\ldots,b_k\rangle = \langle (a_1 \times b_1),\ldots,(a_k \times b_k)\rangle$.

Q is a distinguished function from $D \times W$ into the subset $\{\underline{0}, \underline{1}\}$ of $\mathcal{J}$, where $\underline{0}$ is the k-tuple $(0,\ldots,0)$ and $\underline{1}$ the k-tuple $(1,\ldots,1)$, absolute falsity and absolute truth respectively. Q interprets the two-place existence predicate of L_c and is subject to the constraint that

(ix) for all w, w' in W, if $w \neq w'$ then if $Q(x, w) = \underline{1}$ then $Q(x, w') = \underline{0}$.

So we are going to restrict ourselves to worldbound individuals, things which exist in at most one world. However, this restriction is inessential and we will later mention a reason for relaxing it in a more complex semantics. R is a function from $D \times D \times W$, the universe of the counterpart relation, into $\mathcal{J}$, and meets a number of conditions. First, reflexivity in variables of the second sort:

(x) for all $x \in D$ and $w \in W$, if $Q(x, w) = \underline{1}$ then $R(x, x, w) = \underline{1}$.

But symmetry in individual variables with respect to degree is plausible only when design is held constant; if complex artefacts can be counterparts of simple ones, proportion of parts in common may not be the same. And, obviously, no version of transitivity with respect to degrees is desirable. But we do impose two other conditions:

(xi) for all $x \in D$ and $w \in W$, if all $y \in D$ are such that if $Q(y, w) = \underline{1}$ then $R(y, x, w) = \underline{0}$, then for all $y \in D$, $R(y, x, w) = \underline{1}$ iff $y = x$.

This says that any object with no existing counterpart at w is its own sole counterpart there, and, it will be recalled, is the condition motivated by the analogy with standard possible worlds semantics, in which an atomic predicate may be satisfied at a world by an object which does not exist at that world; thus on counterpart-theoretic semantics as well as Kripke semantics, the Falsehood Principle is imposed on or withheld from atomic predicates on a case by case basis. We will also insist that counterparthood be properly a crossworld relation when it holds between distinct things:

(xii) for all $x, y \in D$, $w \in W$, if both $Q(x, w)$ and $Q(y, w) = \underline{1}$ then $R(x, y, w) > \underline{0}$ iff $x = y$.

I interprets the identity symbol of L_c, which is of category $\langle 2,2 \rangle$ and is a function from $D \times D$ into $\{\underline{0}, \underline{1}\}$ such that $I(a, b) = \underline{1}$ iff

$a = b$. H is a set of characteristic functions f_i^{n+2}, one for each non-logical n + 1-place predicate F_i of L_c. $w^\star$ is a designated member of W, and, lastly, *val* is a function which assigns members of D to constants of sort 2 under the constraint that

(xiii) $val(c) = x$ only if $Q(x, w^\star)$

However, it is also possible to have names for non-actuals in L_c.[15]

The earlier connective clauses need to be modified slightly for more general truth-value sets, which, for instance, need not be complemented. To interpret negation and implication, Goguen [1975, p. 356] defines the functions Neg and Imp thus:

(xiv) $\mathrm{Neg}(\langle a_1,\ldots,a_k\rangle) = \langle(1 - a_1),\ldots,(1 - a_k)\rangle$
(xv) $\mathrm{Imp}(a, b) = \mathrm{lub}\{x: (x*a) \leqslant b\}$.[16]

15. It may be thought that there is a difficulty in counterpart theory with names for non-actuals, such as 'β' on page 163, on the grounds that for a modal sentence containing this name to have a determinate truth condition, 'β' must refer to a particular one of the many worldbound individuals with constitution ϕ_n, and it is unclear how it is to be made to do so. But the condition for determinacy of reference is just that it be possible that there is an object such that necessarily 'β' refers to it alone, i.e.

$\Diamond(\exists x)\Box(\forall y)(R(\text{'}\beta\text{'}, y) \leftrightarrow x = y)$.

In counterpart theory, this is the condition:

$(\exists w)(\exists x)(Exw \,\&\, (\forall u)(\forall z)(\forall y)(Czxu \,\&\, Eyu \rightarrow\ [R(\text{'}\beta\text{'}, y,u) \leftrightarrow y = z]))$.

When counterparthood is identity-like because criteria for it are as strict as those for transworld identity for sets (two-valued counterpart theory) this condition is clearly satisfied, and thus the reference of 'β' is determinate. This is already enough to show that the worldbound nature of the individuals in counterpart theory is not an obstacle to the introduction of names for non-actuals. When we are in the degree-theoretic framework, the condition remains true only by allowing R to be a relation of degree such that if y and z are distinct and both counterparts of x at w', then the degree to which 'β' refers to y at w' is 1 minus the degree to which z is a counterpart of x at w'; this is a strange but not impossible consequence, granted that we are talking about reference to non-actuals. Only if some world where there is an artefact of the fixed design with constitution φ were actual would 'β' be a name of an actual. Note that by our views about the counterpart relation, any sentence of the form $\Diamond\psi(\beta)$ will have a fixed degree of truth regardless of which of the worldbound individuals satisfying ϕ_n 'β' is taken to denote, so for evaluation purposes we can just pick one at random. Note also that the modal determinacy condition implies that possibly there is an x such that actually, 'β' refers to x; since this is true, the Falsehood Convention could not be applied to 'refers'.

16. With respect to note 17 below, it is important to note that if subtraction always makes sense, then a componentwise generalization of clause (iv) on page 170 also satisfies Goguen's conditions on an implication function.

For disjunction and conjunction we take the operations of join and meet defined componentwise in $\mathcal{J}$ by Max and Min respectively, while the quantifier clauses are still (vi) and (vii) on page 173, and the definitions of validity for formulae and arguments also are unchanged. This completes the account of degree-theoretic model theory for counterpart theory.

Consequences

To begin, let us see how a semantics of degrees in counterpart theory defuses the modal paradoxes. Chisholm's Paradox is straightforwardly dealt with, for when we consider its modal operator formulation on page 162, we see that its conditional premisses, which are all of the form

$$\Diamond\varphi_i(\alpha) \rightarrow \Diamond\varphi_{i+1}(\alpha)$$

have L_c-translations of the form:

(A) $(\exists u)(\exists x)[Cx\alpha u \;\&\; \varphi_i(x, u)] \rightarrow (\exists v)(\exists y)[Cy\alpha v \;\&\; \varphi_{i+1}(y, v)]$

and that no instance of (A) is wholly true. Rather, in each such instance, the consequent is slightly less true than the antecedent because anything with constitution φ_{i+1} is slightly less similar to α than anything with constitution φ_i, and so is a counterpart of α at a world, if at all, to a slightly lesser degree than something with constitution φ_i. Our clauses for the connectives ensure that this small gap in degree of counterparthood translates itself up through the structure of the formula to yield an expression slightly less than wholly true: suppose the degree of truth of the antecedent of some instance of (A) is k, in virtue of the highest degree of α-counterparthood open to an object with the relevant constitution at any world u being k (that object has that constitution at u to degree 1, so the conjunction in the antecedent has degree $\mathrm{Min}(k, 1) = k$); then the highest degree of α-counterparthood open at any world to any object with a constitution even further removed from that of α at the actual world will be correspondingly lower, yielding a lower degree of truth for the consequent of the instance of (A); so by clause (iv) for $\rightarrow$, the whole instance gets a degree of truth less than 1. Thus our resolution of Chisholm's Paradox is absolutely parallel to the resolution of the Tall Man.[17]

The Four Worlds Paradox is dealt with similarly, since 'β' is a

17. Some find it natural to formulate the conditional premises of Chisholm's

descriptive name which may be assigned a definite non-actual individual (see note 15). We say that the last world of σ_1 and the last world of σ_2, alleged to differ only with respect to the identity of a particular artefact, are in fact the same world, and those artefacts the same artefact. This artefact is a counterpart at its world of both α_1 and α_n, a counterpart of each to the same degree, even though α_1 and α_n are not counterparts of each other at all. In modal language, we can say that the "Four" Worlds Paradox involves a world which realizes a state of affairs which is "semi-possible", the state of α and β being the same. In deriving the paradox, we appealed to the "fact" that α and β are necessarily distinct, but here we were relying on a version of the necessity of identity,

(6) $(\forall x)\Box(\forall y)(x = y \rightarrow \Box(x = y))$

which is, of course, invalid in the present system, since

(7) $(\forall x)(Exw^{\star} \rightarrow (\forall y)(\forall w)(\forall z)(Cyxw \,\&\, Ezw \,\&\, y = z)) \rightarrow$
$(\forall u)(\forall s)(\forall t)(Csyu \,\&\, Ctzu \rightarrow s = t)))$

is its L_c translation, and the validity of (7) would require that the counterpart relation be transitive in its individual variables in this

---

Paradox as counterfactuals of the form:

$\Diamond\varphi_i(\alpha) \,\Box\!\!\rightarrow\, \Diamond\varphi_{i+1}(\alpha)$

and although this makes the analogy with the Tall Man more remote, essentially the same points hold. First, to handle translation of counterfactuals into L_c, we need to define a new operation 'Rel′' thus. If there are term-occurrences $t_1 \ldots t_n$ in a formula ψ outside the scope of modal operators in ψ, then

$\mathrm{Rel}'(\psi, w) = (\exists u)(\exists v_1)\ldots(\exists v_n)(\mathrm{C}v_1t_1u \,\&\ldots\&\, \mathrm{C}v_nt_nu \,\&\, \mathrm{Rel}(\psi(v_i/t_i), w))$;

otherwise, $\mathrm{Rel}'(\psi, w) = \mathrm{Rel}(\psi, w)$. We now expand L_c by adding the three-place relation symbol $S(u, w, v)$ for comparative similarity ('u is more similar to w than is v'), a symbol of sort (1,1,1), and a degree model is concomitantly understood to be a 10-tuple whose new component is a function $S: W \times W \times W \mapsto \{0,1\}$. The Lewis-Stalnaker analysis of $\Box\!\!\rightarrow$ then motivates the following:

$\mathrm{Rel}(A \,\Box\!\!\rightarrow\, B) = (\exists u)(\mathrm{Rel}'(A \,\&\, B, u)) \,\&\, (\forall v)\,(\mathrm{Rel}'(A \,\&\, \sim B\,, v) \rightarrow \mathrm{S}(u, w, v)))$.

Curiously, this renders every counterfactual in the quasi-Sorites argument wholly false, excepting only the very first. The problem lies with Goguen's clause for Imp, which makes every conditional with a wholly false consequent itself wholly false, provided just that the antecedent has some degree of truth, regardless of how little. So his clause, in this case, does not reflect the gap. However, for the purposes of handling our argument, we can use instead a generalized version of (iv). Then this counterfactual version of Chisholm's Paradox will not lead us to the paradoxical conclusion, since we will resist the inference from 'Possibly, *P*' and 'If it had been that *P*, it would have been possible that *Q*' to 'Possibly *Q*' for the familiar reason, that the degree of possibility of *Q* is dropping off.

sense, that if y is a counterpart of x at u to degree m and z a counterpart of y at v to degree n then z is a counterpart of x at v to a degree not less than Min(m, n). But the relevance of relations of degree to paradoxes of vagueness is precisely that they fail to be transitive. Thus, by rejecting (6), we block the paradoxical argument. But this move is not *ad hoc*. Rather, it is a consequence of our view that the modal paradoxes are Sorites paradoxes, and of our applying a general technique for dealing with such paradoxes to the modal case.

We began this discussion with two aims in view; one was to investigate what kind of essentialism is possible with respect to entities like the slime mould and artefacts, and the other was to demonstrate that the paradoxes pose no threat to our doctrine that facts about identity must be intrinsically grounded. In connection with the first aim, we are now in a position to see that artefacts may be ascribed fuzzy essences. The fuzzy essence of a given artefact may be regarded as a set of pairs, each pair consisting of an object and a degree, the second member of the pair specifying the degree to which the first member is a counterpart of the given artefact. So a fuzzy essence is a perfectly determinate object, like a fuzzy set. In modal operator discourse, two theses are justified by ascription of such fuzzy essences, analogous to the pairs ⟨(MR), (CE)⟩ and ⟨(K), (PSI)⟩. First, at least if each of an artefact's parts is equally important to its functioning, we can say that it is essential to an artefact to have most of the parts it actually has or could have had. The logical form of this thesis, which we label (Z), is awkward to represent because of more general difficulties with vague quantifiers like 'Most', but if we write '(Mx: φx)[ψx]' to mean 'Most of the things which are φ are ψ' then one possible regimentation is:

(Z) $\quad \Box(\forall x)\Box_1(Ex \rightarrow \Box(Ex \rightarrow [(\mathrm{M}y{:}Pyx)[\mathsf{A}_1(Pyx)]]))$

where 'Pyx' is to be read as 'y is a part of x'. Under reasonable assumptions about the interpretation of 'Most', we can even say that (Z) is wholly true. Suppose that b is a counterpart of x to degree k at the fixed world w_1, and that c is a counterpart of x at some world v to degree j. Is it really possible to choose v so that the degree of truth of 'most of the parts of c at v are parts of b at w_1' is so low that subtracted from j, it yields something less than k? In comparing the parts of c at v and the parts of b at w_1, we are again comparing objects some of which are counterparts, and we should

expect that, on the intended interpretation, these degrees of counterparthood will be reflected by the degree of counterparthood between *b* and *c*, and so by the difference between the degrees of counterparthood of *b* to *x* and of *c* to *x*; which latter degrees, of course, will also reflect the *proportion* of parts of *b* and *c* which are counterparts to some positive degree. To arrive at a reasonable and consistent assignment of degrees of counterparthood overall is evidently quite complicated, but it does not seem that, on any such assignment, we must obtain a degree of truth for the subformula of (Z) beginning with □ which is less than *k*.

Secondly, to complete the account of fuzzy essences, we want to formulate a crossworld condition with the metalanguage effect of saying that objects which are similar across worlds in design (and anything else deemed to be relevant) and which have most of their parts in common, are counterparts. Again, for a given pair of objects, degree of similarity, which is relevant to the antecedent, will be matched by degree of counterparthood, which is relevant to the consequent, so this principle is also wholly true. Using '*S*' for the similarity relation involved, the principle we want is:

$$(\text{D}) \quad \Box_1(\forall x)\Box_2(\forall y)[\{\mathsf{A}_1^x\mathsf{A}_2^y(Sxy)\ \&\ \mathsf{A}_1(\text{M}z{:}Pzx)[\mathsf{A}_2(Pzy)]\ \&\ \mathsf{A}_2(\text{M}z{:}Pzy)[\mathsf{A}_1(Pzx)]\} \rightarrow x = y].$$

In the counterpart-theoretic metalanguage, (D) becomes the claim that for any world *u* and object *x* existing at *u*, and for any world *v* and object *y* existing at *v*, and for any *z* which is a counterpart of *x* at *v*, if *x* as it is at *u* is similar to *y* as it is at *v* and most of the parts of *x* at *u* are parts of *y* at *v* and most of the parts of *y* at *v* are parts of *x* at *u*, then $z = y$.[18]

In defusing the paradoxes, we are providing a defence of our view that facts about identity must be intrinsically grounded, but it is evident that the essentialist principles (Z) and (D) are not conse-

18. For ease of comprehension, I have not given a counterpart-theoretic interpretation to 'most' here; but, of course, to say that most of *a*'s parts at *u* are parts of *b* at *v* is to say that most of *a*'s parts at *u* have counterparts at *v* which are parts of *b* at *v*. An unfortunate consequence of Goguen's clause Imp is that if *x* has counterparts to low degree which also exist at *v* then (D) will be wholly false, since one of these can be chosen for *z* while the thing in *v* with most of *x*'s parts at *u* can be chosen for *y*. However, on the revised version of Imp (see notes 16 and 17 above) (D) would then be almost wholly true. We may also prefer to decree that similarity in design and sharing only a few parts is not sufficient for any positive degree of counterparthood.

quences of this view, since we are no longer employing a framework in which transworld identity features. But the counterpart-theoretic framework may be regarded as a generalization of the standard framework (the necessity of identity is an example of a principle which fails when the standard framework is generalized, but it is a not unexpected consequence of a generalization that principles which once held do so no longer) and so we should expect (Z) and (D) to be consequences of a generalized version of the thesis about identity. The generalization is just that transworld heirlines must be intrinsically grounded, or, in Adams's terminology, that facts about realization of thisnesses must be intrinsically grounded. Clearly, our resolution of the paradoxes is in conformity with this thesis, since our whole effort has been towards providing a solution which allows the counterpart relation to hold between objects to just the degree which is demanded by the relevant respects of and degrees of similarity between those objects.

It is left to the reader to establish, by arguments analogous to those used for the oak tree and its acorn, that this requirement of intrinsic grounding cannot reasonably be met by someone who denies (Z) and (D). For these arguments to go through, the branching conception has to be maintained for counterpart theory, but this involves no more than a manageable complication. For instance, in the paradoxical situations we were concerned to model, there was no problem about the transworld heirlines of artefact parts; the difficulty was with what to say about parts put together in different ways or in different groups. Thus the most accurate model reflecting the structure of the situation would be one in which all the worlds have a common initial segment, and the same object occurs in the domains of distinct worlds if it exists at a time in those worlds before they branch. It is possible to work out a detailed semantics for a modal language in which the worlds are ascribed further inner structure, i.e., a sequence of times and a domain of existents at those times, although, in the simplest version, it would not be possible to accommodate intuitions about tolerance in intraworld persistence conditions. However, the counterpart relation itself would only hold to a degree intermediate between 1 and 0 for objects existing after their worlds have branched; and if it held to maximum degree between two such objects, it would be feasible to allow another exception to the world-bound nature of individuals in counterpart theory, for we could

treat identity as a sub-relation of counterparthood, the relation of counterparthood to maximum degree.

The kind of generalization in the modal cases forced by the tolerance principle has its temporal analogue, since the possibility of manipulation of the parts of an artefact through time can give rise to branching, and puzzles about personal identity turn on similar phenomena.[19] For instance, if certain sorts of psychological continuity and connectedness are held to be sufficient for personal identity, then Oldman would seem to be identical both to Newman-1 and to Newman-2, who are themselves distinct. But here personal identity involves transtemporal identity, and the moral of the paradoxical conclusions is just that when criteria said to be sufficient for identity admit of degree, an appropriate semantics for a language, in this case a tensed language, will employ a relation of degree for the interpretation of sentences which are *de re* with respect to the relevant operators, in this case, tense operators. In tensed discourse, the sentence 'Oldman will be two people' will have some degree of truth; indeed, if each Newman is a counterpart of Oldman to maximum degree, which in this example is a not unreasonable assignment, then this sentence will be wholly true, for there will be distinct, contemporaneous counterparts of Oldman. It may be thought that it is objectionable to extend the counterpart-theoretic apparatus to the temporal case, since it requires an ontology of "timebound" individuals, i.e., an ontology of instantaneous things rather than continuants; so the extension apparently implies that there are "really" no such things as continuants, only time-slices which make up continuants as determined by the counterpart relation. But whatever this might mean, it is no consequence of our view, since we do not take the metalanguage

19. See the case of Methuselah in Parfit [1971]. Those who think that, in temporal cases involving ordinary material artefacts where there is branching, the line of continuity is always the one which traces identity, should consider the following entertaining example, due to David Kaplan. Suppose a museum (in California, of course) hires a philosopher to go to Greece, obtain the Ship of Theseus, dismantle it, pack it and dispatch it back to the museum. Suppose also that the philosopher follows these instructions, but that as he removes each plank from the ship he replaces it with a new plank with the same shape, so that when the original planks are all crated he still has a ship in dry dock. The museum receives the planks, reassembles the ship and is about to exhibit it when it receives a phone call from the philosopher, who announces that he has the real ship, and demands large sums of money for keeping quiet about the fraud the museum is about to perpetrate. Assuming the museum director is not an especially dogmatic continuity theorist, will he or she pay up?

sentences of counterpart theory to have the literal meaning which they apparently have; that is, their true meaning is fixed by reverse-translation into the object language. As in the case of the standard framework, the model-theoretic apparatus associated with counterpart theory is simply a convenient device for fixing which arguments are valid, the datum that the paradoxical arguments are invalid being given in advance, by pretheoretic intuition about their object-language formalizations.

Finally, it may be asked why, having set up the counterpart-theoretic generalization of the standard framework, we do not apply it to entities of every category whose members come from other entities in some suitable sense. A short reply to this is that for, say, sets and the products of biological growth, there is no tolerance principle analogous to (T). But why is there no such principle for sets, when application of our apparatus blocks the drawing of the unwanted conclusion that a given set could have had completely different members? In connection with transtemporal identity, it is not too hard to see why there should be a difference between sets and artefacts, for an artefact is a physical thing which traces a spatio-temporal route, and smallish slowish changes in parts or design are consistent with preservation of enough functional organization relative to the normal state of the artefact for the route to be regarded as that of a single object. With sets, nothing corresponds to an artefact's having the same structure as it would have had without replacement of part or change in design (these remarks are formulated to allow for dismantling and reassembly). Thus, within a world, there is nothing more to the set than its members, but this is not true of artefacts. There is a comparable difference between the oak tree and a slime mould, since the identity of the oak tree's propagule acorn is not something which can change as time passes. These contrasts are perhaps less marked when we consider modality rather than time, but there is some plausibility in the thought that our conception of the thisness of an individual is fixed by our conception of how it persists through time, or equivalently, by the content we ascribe to *de re* tensed sentences about it, and in grasping *de re* modal sentences we simply project that part of the content which embodies our conception of the thisness of the individual to the modal case. Thus, if there is no tolerance in the transtemporal heirline of an entity, our conception of its thisness will not permit tolerance in the transworld heirline of

 that entity.

CHAPTER 7: FUZZY ESSENCES

Chapter 8
Substances, Properties, and Events

Substances as things

IN PREVIOUS chapters, we have been concerned with the theory of individual essence for categories of things which we are quite accustomed to thinking of as objects, items whose entitative status is not controversial. The view expounded in those chapters is that individual essences of certain sorts are required by the nature of the identity relation (or some analogue of it), a relation which holds between objects. Thus, to the extent that a certain conception of object seems doubtful, we should expect to find a corresponding weakness in the case for ascribing essences to the putative objects falling under that conception. In this chapter, we will consider three alleged kinds of object, substances, properties, and events, and examine the justification for some essentialist theses about them which others have advanced. This choice of topics is not arbitrary, for we employ modal operators in discourse apparently about such entities, and it is of interest to see how the understanding of modality which we have developed applies to entities whose nature is somewhat opaque by comparison with material objects and sets. We choose the particular topics of substances, properties and events because there is already a sophisticated literature on them, providing a range of views with which those to be presented here may be contrasted.

Apart from Kripke's ideas about origin, probably the most widely discussed variety of essentialism is one due to Hilary Putnam, which he motivates by an argument quite unlike anything we have come across so far. This argument might be called the Doppelgänger Argument, and is used by Putnam to attribute essential properties to natural kinds of things, such as species or material

substances;[1] for instance, Putnam holds that it is essential to water to be H_2O and that it is essential to tigers to be mammals.[2] So we begin our discussion by first expounding and then evaluating the Doppelgänger Argument.

The argument invites us to perform a thought experiment. We are to imagine a place called Twin Earth, where everything appears to be as it is on Earth; upon investigation, however, some differences manifest themselves, and the essentialist claims are motived by intuitions about how we would revise 'first impression' judgements when the results of such investigations are made known to us. Thus, suppose, as explorers, we set out from Earth and arrive on Twin Earth, where we quench our thirst with the stuff which flows in Twin Earth's rivers and streams and falls from the clouds in Twin Earth's sky, and arm ourselves for protection against the carnivorous large orange and black striped cat-like creatures which hunt in Twin Earth's jungles. So we might report back to Earth that there is an abundant supply of water on Twin Earth, and that on Twin Earth the tiger is not yet an endangered species.

Then our scientists find out that what we have been drinking does not have the same molecular structure as water on Earth; instead of being H_2O, it has some complicated structure XYZ. And after killing one of them we discover that the creatures from which we are defending ourselves are not mammals, like tigers on Earth, but rather, very complex machines that happen to look like tigers and to be programmed with tiger-like behavior patterns. At this point, Putnam plausibly claims, we would cease to think of what quenches our thirst on Twin Earth as water, and of the carnivorous predators as tigers. When speaking loosely, we might refer to 'Twin Earth water' and 'Twin Earth tigers', but we would agree that Twin Earth water is not the same substance as water, and that the android tiger is not the animal species *Panthera tigris*. These attitudes, Putnam concludes, betray the view that it is essential to water to be H_2O and essential to tigers to be mammals.

The most straightforward interpretation of Putnam's argument

1. It would be misleading to suggest that Putnam's main concern is to defend essentialist theses about substances. In fact, he is more interested in defending semantic theses about substance terms. For further discussion, see Salmon [1981, Chs. 4 and 5].

2. We will speak throughout of its being essential to a substance to have such-and-such a property although Putnam writes instead that it is necessary. For the purposes of our discussion, the difference is immaterial.

is that it concerns the transworld identity conditions of such entities as substances and species: the transworld heirline of a substance is to be fixed by considerations of molecular structure, or better, abstracting from current physical theory, by considerations of fundamental explanatory nature; while the transworld heirline of a species is to be fixed by consideration either of relative position in taxonomic or evolutionary trees themselves satisfying some kind of crossworld similarity constraint, or by considerations of genetic and physico-chemical crossworld similarity.[3] Understanding Putnam's argument in this way helps to dispel some kinds of puzzlement its conclusion is apt to produce. For instance, the features on the basis of which the word 'water' is ordinarily applied on Earth are duplicated on Twin Earth; so it may be asked how we can withhold it there without changing the meaning of the word. But on the proposed interpretation of Putnam, we can explain what is wrong with the idea that the meaning of the word must change by an analogy from the theory of reference to material objects. Suppose the explorers have left their friend Smith behind on Earth; however, after a while on Twin Earth, they come across an individual exactly like Smith, that is, an individual who possesses all the features which prompt application of the name 'Smith' on Earth (compare Twin Earth water and 'water'). We can suppose that this individual on Twin Earth even appears to recognize the explorers as old friends, and so they believe that he is Smith, and has transported himself to Twin Earth in some other craft. Then it transpires that this individual too is an android.

In this case, it is indisputable that, despite the great surface similarity between the android and the explorers' friend, these are different individuals; it is just wrong to say that the Twin Earth android is Smith, or, more exactly, it is wrong for anyone, such as one of the explorers, in whose language 'Smith' is a name of the Earthman. Similarly, anyone who speaks a language in which

3. It is important to note that Putnam is not making an essentialist claim about *members* of the species *Panthera tigris* or about particular *samples* of water. It we treat 'water' and 'H_2O' as predicates of quantities of stuff, then Putnam's claim about water may be regimented

(i) $\Box(\forall x)(Wx \rightarrow H_2O(x))$.

The essentialist thesis about individual quantities of stuff with which (i) should not be confused is

(ii) $\Box(\forall x)\Box(Wx \rightarrow \Box(Ex \rightarrow H_2O(x))$.

'water' is a name of the substance found on Earth in rivers, streams, etc., will confuse distinct individuals (substances) if he applies 'water' to the water-like substance found on Twin Earth. Exactly parallel remarks may be made about species. It may be true that our ordinary words for animals are less successful in marking out natural groupings than are our substance words, but this is irrelevant to the underlying point Putnam wants to make about the transworld identity conditions of such groups.[4] There are of course disanalogies between substances and persons as well as analogies; for instance, Smith cannot be on both planets at once, while water can be, but this does not undermine our demonstration that superficial similarity of substances in the two places is hardly sufficient to justify application of the same substance word.

Despite the plausibility of the claims Putnam makes about what we should say about Twin Earth, there is a major doubt about whether the essentialist conclusions he wishes to draw really follow. The problem is that an essentialist thesis about substances should be backed by an argument which speaks of the identity conditions of substances from possible world to possible world, but the story Putnam tells concerns merely a journey within the physical space of a single possible world: the explorers go from planet to planet only. Thus the moral confirmed by the story itself is just that the same substance is present in different places (within a world) only if its instances at those places all have the same fundamental physical nature. And this moral is compatible with the existence of a possible world all of whose samples of water have molecular structure XYZ. Putnam does not seem concerned by this gap between his story and his conclusion. He writes [1978, p. 70]:

> Suppose, now, that I discover the microstructure of water…that water is H_2O. At this point, I will be able to say that the stuff on Twin Earth that I earlier mistook for water is not really water. *In the same way*, if you describe, not another planet in the actual universe, but another possible universe in which there is stuff with the chemical formula XYZ which passes the 'operational test' for water, we shall have to say that that stuff isn't water, but merely XYZ.

4. In [1981] Dupré assembles much evidence that our ordinary words for kinds of animals and plants do not pick out the natural kinds recognized by science. But this is irrelevant to essentialist these about kinds, which would just have to be expressed using technical terms for species.

In this extract, we emphasize the words 'In the same way', since this phrase appears merely to mark a *lacuna* where an argument is needed, rather than to signal an unproblematic extension of the given considerations to an analogous case.

Someone might try to fill the lacuna by appealing to the parallel with ordinary objects, like persons, where most, perhaps all, of the features of a person which prompt our recognition of him are contingent features, while a feature we do not have access to in ordinary circumstances, his origin, is essential. It might be said that just as the zygote is to the organism, so physical nature is to the substance. But this cannot just be said. We need to see the argument that someone who denies such a thesis is committed to bare truths or extrinsicness in connection with transworld identity, and also an argument that even for such esoteric entities as substances, bare truth or extrinsicness should not be tolerated. So it looks as if the Doppelgänger Argument is a red herring, and that a successful argument for Putnam's conclusion about the essential properties of substances would have to be pursued by the same methods as were used for sets, organisms, and artefacts in previous chapters.

Identical substances are necessarily identical

There are other arguments in the literature for the essentiality of fundamental physical nature which do not, or need not, rely on Doppelgänger-type thought experiments; these other arguments might be termed 'semantic'. The simplest is due to Kripke, according to whom 'Heat is the motion of molecules', 'Water is H_2O', and other examples, 'are all in some sense of "identity statement", identity statements' [Kripke 1971, p. 143]. Then if 'Water is H_2O' is a true identity statement, 'Necessarily, water is H_2O' is a truth delivered by the necessity of identity. And even if we move to a counterpart-theoretic framework, in which the necessity of identity is not valid, the evidence of the Twin Earth thought experiment suggests that this will not lead to the kind of counterexample to the essentialist claim which the sceptic wants. However, three points may be made about Kripke's argument, which appear to undermine it.

(i) If 'Water is H_2O' is an identity sentence, then 'water' and 'H_2O' must be terms picking out entities. So far, we have referred to these entities as substances, but exactly what kind of entity is the substance water (or, for that matter, the species *tiger*)? Certainly not the Goodmanian sum of actual samples of water, since it is the individual essence of that sum of water to be composed of

exactly the water it is in fact composed of, while it seems true to say that there could have been more or less water than there actually is. So it seems that the substance (species) must be an abstract entity of some sort and, if this is so, then it is no longer clear that 'Water is H_2O' is even true. For, on this construal, all that science will have discovered is that actual samples of water have the structure H_2O, and it is controversial that actual coextensiveness is sufficient for identity of the abstract objects, which are perhaps the properties of being water and being H_2O; whether scientists should be said to have discovered an identity of properties or merely an identity of the sets of things which actually have those properties is a philosophical question on which a theory of the essences of substances should remain neutral, if possible; but Kripke's argument presumes a particular answer to that question.

(ii) Another presumption of the argument is that it is definitely correct to say that ice and steam are forms of water, water in the solid and gaseous states respectively, rather than that ice, water and steam are three forms of H_2O. For if we say that latter, then 'Water is H_2O' is not an identity sentence (if it is true) but, rather, is elliptical for 'Water is H_2O in the liquid state'. Again, it is unclear why a theory of essences should have to rule that one form of speech is the right one, and that we do not have the 'is' of predication in 'Water is H_2O'.

(iii) If 'Water is H_2O' is an identity sentence, then its necessitation follows only if the terms flanking the 'is' of identity are rigid designators. However, 'H_2O' seems to have semantically significant structure; perhaps it abbreviates a definite description such as 'the substance pure samples of which have molecules composed of two atoms of hydrogen and one of oxygen in such-and-such a configuration'. Is this description a rigid designator? That is, does it denote the same substance at every possible world, or even just at every world where it denotes at all? A sceptic about the necessity of water's being H_2O can afford for the sake of the argument to concede that 'hydrogen' and 'oxygen' are rigid designators: he may still ask why it should be thought that putting two hydrogen atoms and one oxygen atom together in that configuration in any world always yields a molecule of the same substance. In other words, on the hypothesis that 'H_2O' abbreviates such a definite description, assuming that it is a rigid designator is equivalent to assuming that what it is to be the same substance in different worlds is to have the

same molecular structure, or explanatorily fundamental nature, in those worlds. Obviously, this is the doctrine which is to be established, so a simple argument from the necessity of identity is quite circular.[5]

Crossworld equivalence relations

We can avoid the difficulty of specifying the entity to which 'water' refers if we treat the word as a predicate. Putnam emphasizes the fact that 'water' is introduced by ostension of samples, and we can regard such ostensive sentences as attributing a property, that of being water, to objects, the quantities of stuff which are the samples. Suppose two such quantities are both water; then although

5. Similar points are made in Salmon [1981, Ch. 6]. In Almog [1981] it is argued that the sense of 'water' has an indexical element: what 'water' refers to is determined by a feature of the context in which the word is introduced into our language; it refers to the substance which, in that context, had certain phenomenal features (the context is a composite of place, time and possible world). But Almog is unable to tell a consistent story from this premise, for he wants to say both that 'water' is directly referential, that is, it introduces its extension into propositions expressed using it, where the extension at a world is (p. 352) 'an appropriate body of liquid' ('The only sensible way is to acknowledge that…[substance terms]…bring into the propositions the extensions themselves' (p. 354)) and that these terms bring their contents into the propositions, where the content is an intension, i.e., a function from worlds to bodies of liquid (p. 353). And he also professes metaphysical agnosticism about what the reference of 'water' is (pp. 354–5). Part of the trouble here stems from a problem about the underlying theory of indexicals, due to Kaplan [1989b], in which the content of an indexical in a context of use is at once an intension, a function from worlds to extensions (a constant function) and the extension of the indexical in that context. One must make up one's mind which of the two it is to be. Moreover, the spirit of Kaplan's theory demands that the content be the extension, but this is not a possible choice for 'water' if it is to be true that there could have been more or less water than there actually is. But quite independently of how these details are to be worked out, it is not very plausible that substance terms are indexicals, i.e., words whose semantic values or contents vary from context to context. It may seem as if this is right for the Twin Earth case, but only when the redundant extra feature of Twin-Earth-English speakers is added to the example; then their use of 'water' and our use seems like your use of 'I' and mine. Without this feature, the example is just like one in which real gold is found in one part of this planet and fool's gold in another; someone who uses 'gold' for both substances is using the word ambiguously, and if there are speakers in the area where fool's gold is to be found who call it 'gold' we should say that either they speak our language and use the word ambiguously, or, more likely (assuming they have no contact with us of the kind which makes it plausible to regard them as members of our linguistic community) that they speak another language in which, by sheer coincidence, they use the same word as we do for a substance which looks rather like the substance to which we apply that word. It would be quite unmotivated to speak of indexicality in this case, which is not significantly different from Putnam's.

we might say that they are two quantities of the same substance, this need not commit us to an ontology of substances, for the statement that this stuff is a sample of the same substance as that stuff can be regimented as an atomic sentence of the form Rtt', to the effect that this sample is cosubstantial with that one. This move undercuts any argument for its being essential to water to be H_2O which turns on treating substances as entities and applying our earlier considerations concerning transworld identity to them (any defence of the essentialist thesis for which it was crucial that substances be treated as entities would *ipso facto* be unsatisfactory). Our point about the Twin Earth thought experiment may then be expressed by the objection that although the experiment establishes that the cross-spatial and cross-temporal (intraworld) application conditions of the relation of being cosubstantial with involve the fundamental physical nature of the samples, it does not follow that the same is true of the crossworld application conditions.

However, to say that it does not follow prompts the more general question of how the intraworld application conditions of a two-place relation in modal language are affected if that language is equipped with the doubly-indexed 'actually' operators so that the relation can be used to make crossworld comparisons. If '*Cab*' means that *a* is a sample of the same substance as *b* then

$$\Diamond_1 \mathsf{A}^a \mathsf{A}_1^b Cab$$

makes such a comparison. In possible worlds terms, it says that there is some *w* such that *a* as it is in the actual world is cosubstantial with *b* as it is in *w*. For '*b*' we could then put some descriptive name 'β' of a counterfactual sample of a substance much like water except for having the structure XYZ. With this apparatus, we might then try to fill in the lacuna in Putnam's argument as follows. The Doppelgänger arguments show that, within a single possible world, application of the relation 'is cosubstantial with' is constrained by facts about fundamental physical nature. To ignore this constraint would be to change the meaning of that relation. But adding extra modal operators to modal language cannot affect the sense of a relation symbol already in the language, and so we cannot ignore the constraint in crossworld applications of the relation. Hence

$$\Diamond_1 \mathsf{A}^a \mathsf{A}_1^\beta Ca\beta$$

must be false.

To evaluate this argument, we should consider some other cases of equivalence relations. First, take the relation 'is sameshaped with'. It is clear that the intraworld and crossworld application conditions of this relation are the same. A related fact is that, in cases where the relation holds, there is a level at which it does not make sense to ask in virtue of what it holds: such cases afford examples of bare truths. Even if we agree to say that two objects are sameshaped in virtue of their both being square, we cannot say what the first's being square, or the second's being square, consists in: it would simply double-count the same fact to speak here of sizes of angles and lengths of sides, in the way that it would double-count the same fact to say that a man's being a bachelor *consists in* his being male and not yet married (so the relationship of analytic equivalence is not what we intend by 'consisting in'). Contrast this with 'is of the same nationality as'. One could say that *a* and *b* are of the same nationality in virtue of their both being citizens of the United Kingdom, just as two shapes may be sameshaped in virtue of their both being square. But, in this case, one can also say in virtue of what fact *a*, or *b*, is a citizen of the UK. This may be in virtue of having been born there, or in virtue of having undergone a naturalization process, or in virtue of the obtaining of any other condition which the UK authorities happen to deem to be sufficient for citizenship. Moreover, in making a crossworld application of 'is of the same nationality as' as in 'if *b* had been thus-and-so then *b* would have had the same nationality as *a* actually has', we may imagine a world in which some condition suffices for being British which does not actually suffice, and we can do this without changing the meaning of the relation.

How does 'is cosubstantial with' compare to these cases? As before, we can say that *a* and *b* are cosubstantial in virtue of both being samples of water, but it also seems correct to say that, for the intraworld case, each is a sample of water in virtue of its having a certain physical nature. We expect the superficial and easily detectable differences between pieces of stuff to reflect fundamental differences which explain the superficial ones, and it is the fundamental differences which have the final say in classification; so someone who refuses to classify samples in this way may fairly be said not to understand what a substance is. However, before concluding that 'is cosubstantial with' is like 'is the same shape as', and that the 'same explanatory nature' criterion should be pro-

jected across worlds, we should ask why this criterion is appropriate for intraworld applications. The conception of substance that goes along with the criterion is made for our conception of the physical universe as a law-governed causal system, for one consequence of claiming that *a* and *b* are samples of the same substance is that, *ceteris paribus*, they may be interchanged in a fixed type of environment without affecting the outcome of processes unfolding in that environment; it is thus that experimental results confirm or disconfirm an application of 'cosubstantial'. This consequence derives from our understanding of the causal powers of underlying physical natures – there is a common *explanans* in causal explanations of the behaviour of two samples of the same substance in the same conditions – together with the reliability of the laws governing causal connections.

If these remarks about the intraworld application of 'is cosubstantial with' are correct, it follows that in making a crossworld application of this relation the rationale for the criterion of intraworld application is lost; for, of course, distinct possible worlds cannot constitute a single causal system. Here there is an analogy with 'is the same *F* as', where *F* is a sortal for material objects, and a contrast with 'is the same set as'; for none of the standard criteria for intraworld identity of *F*'s, e.g. spatio-temporal continuity, can be applied across worlds, while there is no problem in applying 'has the same members as' across worlds.

This conclusion may seem inconsistent with our earlier advocacy of the branching conception of possible worlds, a conception which effects a reduction of some instances of crossworld identity to intraworld identity. When we are dealing with samples of the same substance there seems to be nothing corresponding to the possibility of an object's coming into existence in a world after that world has branched from the actual world, so it might be suggested that the branching conception reduces all applications of 'is cosubstantial with' to intraworld applications. But this is incorrect, since a world might branch from the actual world before there is any sample of the relevant substance in the actual world. Furthermore, the branching conception does not assert that every possible world branches off from the actual world, only that every world with an object in common with the actual world branches from it; there are other possible worlds (in some sense 'qualitative') which do not represent possibilities for any actual object, or for any non-

actual object whose identity can be fixed by predicates and names of actuals (such as the artefact β from the Four Worlds Paradox), and it cannot just be assumed that 'is cosubstantial with' has no application between such a world and the actual one.

Hence we have failed to discover any compelling ground for holding that the relation of cosubstantiality gives rise to essentialist truths of the kind it has been credited with the power to generate; the possibility remains open that, like 'is of the same nationality', a particular substance-predicate may be applied in distinct worlds on the basis of distinct criteria. In the case of 'same nationality', what justifies us in claiming that *b* is British in a world where there are new ways of coming to be British is that the nation of which *b* is a citizen in that world is indeed the UK, and that 'has British citizenship' is applicable to *b* in virtue of the consequences of the status he has with respect to the UK in that world – he has the same rights and privileges as UK citizens actually have, for instance. One reason why the necessity of water's being H_2O seems plausible, perhaps, is that it is not easy to think of a counterexample in which there are features corresponding to these in the citizenship case.[6] But that is hardly an argument for the essentialist claim, about which this discussion justifies some agnosticism.[7]

6. Unlike the nationality case, there is a problem in agreeing upon what would count as the same consequences. But suppose that matter is non-atomic and yet some substance exhibits all the more obvious characteristics of water, although its composition is not proportionally two to one of anything. There may be no compelling reason to apply 'water' to portions of such stuff, but there does not seem to be any compelling reason not to, either.

7. How do these considerations apply to the question of the essential properties of species? In [1976] McGinn argues that the necessity of origin applies to species and endorses Dummett's remark [1973, p. 144] that 'even if creatures exactly like men arose from dragons' teeth, they would not be men, because not children of Adam'. In our terminology, the claim is that the crossworld extension of 'is a member of the same species as' contains $\langle a, u, b, v\rangle$ only if the species of *a*'s ancestors in *u* are the same as the species of *b*'s ancestors in *v*. How strong is the case for this claim? The intraworld application conditions of 'is cospecific with' do not refer to ancestry (although this is relevant to the generic classification of species) and so there is even less support for the essentialist thesis here than there was for the one about substances, where intraworld application conditions are answerable to physical nature. The main criterion for intraworld application of 'is of the same species as' is that of reproductive behaviour: in a population of fauna within a particular geographical region, the boundaries of the species represent barriers across which mating, or at least the production of viable offspring, does not naturally take place. Evidently, this criterion is straightforwardly applied only to organisms which reproduce sexually, and only to populations in the same geological period and same geograph-

Properties

It is possible to think of properties as entities of a certain sort; for instance, if the substance water and the property of being water are the same thing, then if we construe the discussion above to have been about the transworld identity conditions of substances we may think of it as a special case of the more general problem of the transworld identity conditions of properties. This problem would also have arisen earlier if instead of formalizing (17) on page 89, 'my car could have been the same colour as yours actually is', using the doubly-indexed actually operators, we had regimented it as:

(1) $(\exists x)(x$ is a colour & applies to$(x$, your car) & $\Diamond$(applies to$(x$, my car))).

If (1) is intelligible, and if transworld identity must be intrinsically grounded for all categories of entities, then it must make sense to ask in virtue of what the colour property which applies to your car in the actual world is the same property as the colour property which applies to my car in a world which verifies (1); and the same can be said for any property of a given type, a type which can be expressed by a predicate (other than a sortal for ordinary things) which can replace 'colour' in our sentence, or one like it, e.g. 'size', 'shape', 'make', 'design', 'age' and so on.

Our view is that such transworld identities of properties reduce to the holding of other transworld relations between ordinary individuals. Thus to say that a certain shape property in u is the same as a certain shape property in v is to say that the things in u to

---

ical region. However, when these conditions are satisfied, the decision whether or not to discern one species or two in a given population is still answerable to our view of the correct shape of the evolutionary tree: it is just that reproductive isolation in such circumstances is the best evidence for what that shape should be. When the conditions are not satisfied, there is an element of arbitrariness in taxonomic classification (see Maynard Smith [1975, pp. 209ff.]). Where populations in different periods are concerned, taxonomists tend to classify using similarity measures derived from decisions taken in cases where the three conditions are satisfied. However, if fossils suggested an earlier species very similar to a current one, but there was good reason to believe that the earlier species had a very different descent, a distinction between the species would be maintained. Again, however, this crosstemporal constraint is underpinned by the idea of a single evolutionary history encompassing all life on Earth, a conception which has no crossworld application. So if Dummett's creatures were exactly like men, this kind of reason for not counting them as members of the species *Homo sapiens* lapses. Perhaps counting such possible creatures as being of the same species as actual men implies commitment to counterfactuals about mating behaviour and the viability of offspring, but there is no reason why such counterfactuals could not be true.

which the first shape property applies are sameshaped with the things in *v* to which the second shape property applies; so the reduction is effected by a crossworld equivalence relation which holds between things of the kind to which the property applies within a world. But as our discussion of substances reveals, it is not always straightforward to decide what the crossworld criteria of application of the equivalence relation should be; for instance, consider the problem of applying 'is sameshaped with' across worlds with radically different physico-geometric properties. However, the intraworld criteria provide guidance up to a point, for it seems correct to say that if there is nothing in virtue of which the relation holds within a world, that is, if its holding in a particular case is a bare fact, then its holding across worlds in a particular case will also be a bare fact. Consider, for instance, the example of length. Someone might suggest that, within a world, two objects are of the same length iff, were they laid alongside one another, the corner points of the adjacent edges would be in contact. But apart from the fact that in the present context it is unhelpful to try to ground an apparently non-modal fact in a modal fact (the one expressed by the counterfactual), such counterfactual analyses do not work unless the concept they are analysing is dispositional; for it may well be that although two objects are of the same length, circumstances are such that if either were or had been moved, its length would change or would have changed. Similar objections can always be found to proposals of this kind, which shows that being of a given length is not the same as being disposed to affect measurement instruments in a certain way, nor indeed to behave in any other particular fashion.

However, not all properties are like length properties. Some are explicitly dispositional, like the property of fragility, which is the disposition to fragment if subject to an impact of a certain force.[8] But there is no mystery about the crossworld application condi-

8. Fragility is not just the disposition to fragment when involved in the initial stages of a process which in normal circumstances would result in an impact being suffered. Rather, the impact must occur. But even then the outcome of the impact is not all that matters, Suppose that an object would shatter if subject to a certain impact but for God's intention to disrupt the causal efficacy of the impact so that no shattering occurs although the impact does. Then even if God intends to do this every time, such an object should still be said to be fragile, since the normal process which leads from impact to shattering is being subject to interference from outside.

tions of 'is disposed to fragment like'; what is required is that the same counterfactuals be true of each object at its own world. It is not required, nor is it sufficient, that the objects have the same physical nature. For a physical nature which realizes the dispositional property of fragility at one world may not do so at another; and the same counterfactuals might be true at different worlds of objects with different physical natures across those worlds.

Another interesting class of properties are those which Locke called 'secondary qualities', which he identified with powers in objects to affect us in certain ways.[9] The property of being yellow, for example, could be identified with the disposition to produce in normal observers and in standard observational circumstances visual experiences with a certain characteristic phenomenal quality, what one might call 'phenomenal yellowness'.[10] Thus applica-

9. See Locke's *Essay Concerning Human Understanding*, 2.vii, 14–17. Unfortunately, Locke says the same thing about primary qualities (*loc. cit.*, 8), which are supposed to differ from secondary qualities in virtue of producing ideas in our minds which resemble the primary qualities themselves. But if primary qualities are also dispositions, it is hard to see what 'resemble' could mean here. We may interpret Locke charitably, as intending the dispositional/non-dispositional contrast.

10. See Peacocke [1983] for a theory of 'phenomenal colour' which elegantly resolves the quandary of the relative primacy of '*F*' and 'looks *F*', for *F* a colour word. The idea that 'yellow' is so much as coextensive with 'produces experiences with such and such features in normal observers in normal circumstances' has been contested in Averill [1982], where the author points out that a surface made up of tiny red and green dots looks yellow to normal observers under normal lighting conditions, but is really red and green; that a black and white disk can appear to have red, yellow, and blue bands when it is spinning; and that a small green patch on a large yellow surface looks blue to normal observers. This may suggest that there is a notion of what the colour of an object really is which our dispositional account does not explain, but other possible morals are that there is a non-circular reason for saying that, in these cases, circumstances are not normal, or, contrary to Averill's implied assumption about the cases, that there is no clear matter of fact about the colours of the objects, because it is unclear which circumstances are normal, and an element of decision, guided by other cases, is involved. This last option appears attractive. For instance, the facts we are faced with in the case of the red and green dotted surface is that it produces sensations with the phenomenal-yellow quality in some conditions and that it produces sensations with phenomenal red and phenomenal green dotwise intermingled in others (close-range observation). Neither type of circumstance has an obvious claim to be regarded as the "normal" one, but, if we have to choose, it is hardly inexplicable that we should prefer the circumstance of close-range observation (similarly for the preference for observation of the disk when stationary and of the small patch when not embedded in a rather special background); in particular, there is no reason to suspect that this choice cannot be justified without presuming upon a fact about what colour the object really is.

tions of 'is of the same colour as' within the actual world are relative to a conception of normality which may not be appropriate for another world in which objects are nevertheless coloured; what matters are *actual* standards of normality. A more radical account is that, e.g., 'yellow' *means* 'produces phenomenally yellow sensations in normal observers in standard circumstances', so that for an object to be yellow at a world *w* it is required to affect observers in a certain way at *w*. Kripke [1972, p. 354] has objected to such an account that

> 'yellow' does not mean 'tends to produce such and such a sensation'; if we had different neural structures, if atmospheric conditions had been different, if we had been blind, and so on, then yellow objects would have done no such thing.[11]

However, a defender of the radical view could account for the apparent plausibility of this objection by appeal to a scope distinction. What is true is that, concerning things which are yellow, if we had had different neural structures, these things would not have produced phenomenally yellow sensations in us, which means that in those circumstances they would not have been yellow. And if we had evolved without the capacity for visual sensations, nothing would have been coloured. So, according to this view, there is no particularly intimate connection between colours and wavelengths of light, other than the contingent connections which scientists have discovered actually to obtain.[12]

11. Kripke appears to hold that a non-circular specification of 'normal circumstances' is impossible (*loc. cit.*). See the previous note for a reply to this.

12. Kripke's view, also that of Davies and Humberstone [1981], is that colour words stand for properties which physicists can mark out in more fundamental terms, presumably terms involving the wavelength of emitted or reflected light. Thus, if we had been different in certain ways so that red objects had looked yellow to us and we had applied 'yellow' to them instead of to the objects which in those circumstances would have looked yellow to us had we not been thus different, we would be picking out with that use of 'yellow' a different property from the property we actually use it to express. The Kripke-Davies-Humberstone view does justice to Locke's idea that colours are powers in objects to affect us in certain ways, since that view identifies the colour property with the physical ground of the disposition. The radical view does justice to Locke's idea that, in some sense, colours are not really 'in' the objects themselves. Here the thought is that primary and secondary qualities differ in that the identity of secondary qualities at a world depends upon facts about the make-up and environment of sentient beings at that world.

Events

We shall take events to be dated, unrepeatable occurrences occupying definite intervals of time. Clearly, the time of an event contributes to its intraworld identity condition. A very rough and intuitive account can be set out by adding to the time of an event two further components, the objects *involved in* the event, and the types of changes in properties which these objects undergo for the duration of the event; this will be elaborated and qualified in due course. We might include the location of an event in its intraworld identity condition, but such a component would be redundant, since the same objects cannot be in different places at the same times; so the objects and times fix the places, whereas place and time does not fix object (a statue and the bronze of which it is made are different objects, although they are in the same place). On this intuitive view, then, an event consists in a triple of (i) a set of objects; (ii) types of changes of properties for each object in the set; and (iii) an interval of time. Obviously, not any combination of items falling under (i), (ii), and (iii) constitutes so much as a possible event; while we say that a triple constitutes an event, rather than is identical to it, to leave it open that one and the same event may be constituted by different triples in different worlds. Finally, it might be preferable to treat types of changes of properties as fixing a type for the event, like sorts for objects; types, like sorts, will be specified by expressions of a special category' the type of an event being, very crudely, the 'basic, change which is exemplified when the event occurs, in a sense of 'basic' yet to be explained. So an event of type τ would be an exemplification of τ by a set of objects and an interval of time.

In a famous footnote to a discussion of counterfactuals about events, David Lewis gives a clear statement of an intuition about events which appears to show that *de re* locutions are quite unproblematic in connection with them [Lewis 1973b, fn. 9]. Lewis was discussing whether or not a certain event, the death of Socrates, would have occurred if things had been different in certain ways, and he points out that to ask whether that event would have occurred is not the same thing as asking whether there would have been a unique death of Socrates. For we can imagine that everything is as it actually is up to and through the death of Socrates, and then Socrates is resurrected to die a second time. In such possible circumstances, it is plausible that his first death is the same event as his actual death, but it is not a unique death. Equally,

Socrates might have died in a totally different way from the way in which he actually died, and plausibly the event which was his actual death does not occur in such circumstances. Here Lewis is apparently relying on an intuitive conception of *the event itself*, apart from such descriptions of it as 'the death of Socrates'; the event has a thisness under which it can be projected across worlds, a thisness which may either be primitive or else intrinsically grounded. If the latter, then there must be favoured descriptions of the event, which will fasten onto its individual essence.

Opposed to Lewis's view is a position which we will call *de re* scepticism, which says that the conception to which Lewis hopefully appeals simply does not have any content: there is no such thing as *the event itself*. Since we are being neutral on the question of whether or not Lewis's conception is completely analysable, it follows that *de re* scepticism is stronger than the denial of the thesis that events have individual essences. Such a sceptic could make sense of *de re* modal discourse involving quantification over events by employing some relatively stipulative or extrinsically grounded criteria for transworld heirlines, a position which would be consonant with a degree of scorn towards the alleged entitative status of events themselves.

To proceed further, our rough and ready sketch of the nature of events needs to be elaborated; fortunately, we can avoid extra work here by simply availing ourselves of the most sophisticated and complete development in the current literature of the intuitive ideas about events employed above, due to L. B. Lombard [1979, 1981, 1982]. The central idea of Lombard's theory of the nature of events is that of a *quality space*; events are movements of objects at times in such spaces. A quality space is a space of static properties, properties possession of which does not imply change (compare being five feet in height with growing). Quality spaces are *closed* under certain kinds of physically possible changes of static properties, and the individuation of quality spaces determines the individuation of events. Let us confine ourselves, as Lombard does, to events which are changes in a single physical object, such as the death of Socrates, as opposed to, say, a mass suicide. So, for instance, if an object changes shape during an interval of time while it is also changing location, we will say that two events are occurring, since the object is moving simultaneously in two different quality spaces. Static shape qualities and static location prop-

erties are assigned to different quality spaces since changes of shape and changes of position are unconnected, in the sense that an object's position can change without its shape changing, and conversely. By the closure condition, all shape properties are in one space and all location properties in another.

An object may change in virtue of having a part which changes; in Lombard's view, there is exactly one event occurring in such situations, which involves both the larger object and its part. An object is then said to be *minimally* involved in an event *e* iff it is the smallest object a change in which is identical to *e*. When a change in an object is not to be construed as a change in other objects out of which the first is composed, Lombard says that the object is *atomic*; an *atomic quality space* is a quality space with only properties which it is physically possible that an atomic object has, and an *atomic event* is a movement of a single atomic object in a single atomic quality space. This gives us a basis for a classification of events; each event is either an atomic event, a temporal sequence of atomic events, an event composed of simultaneous atomic events, or a temporal sequence of such composite events. From these materials, Lombard derives the notion of a *canonical description* of an event. An atomic event's canonical description is a singular term of the form

$$[s, \varphi, t]$$

where *s* is a name of the subject of the event, *t* is the time interval the event occupies, and φ is an expression for the gaining and losing of the static properties gained and lost by *s* in its movement in the quality space; the singular term schema might be read 'the φ-ing of *s* during *t*'. A non-atomic event's canonical description is constructed out of the descriptions of the event's atomic constituents, so as to describe how those events are bound together into the non-atomic event. So the death of Socrates is a non-atomic event whose canonical description will involve an account of the physical effect on his body of the poison he swallowed.

Lombard's essentialism

Canonical descriptions seem to be the favoured descriptions the essentialist is looking for under which an event is projected into counterfactual situations, or pronounced to be absent from such situations; not that those who use *de re* locutions in connection with events must be able to come up with such a description, but

its availability legitimizes the use. If canonical descriptions are favoured in this sense, then it must be essential to events to occur when they do and to have the minimal subject which in fact they do have. Lombard argues for both theses. *Prima facie*, it may seem obvious that even the minimally involved subject of an event is accidental to it, but this appearance rests upon a scope confusion. Since β might have been the particle with the positive charge instead of α, it is possible that the impinging of the particle with the positive charge on the photographic plate involves β as minimal subject rather than α. But the obvious truth here bestows no plausibility at all on the claim that the actual impinging of the charged particle might have had β as minimal subject. However, Lombard [1981, pp. 142–5] constructs an ingenious case for the anti-essentialist about minimal subject (or for the outright *de re* sceptic) in which there is no scope confusion, before trying to show that, even in this hard case, intuition is on the side of the essentialist.

Suppose that in u the Ship of Theseus is constructed from planks $p_1,\ldots,p_n$ and that in u the Ship of Theseus sinks in a certain storm in a certain place and time. Let v be a world resembling u as much as possible compatible with the following conditions obtaining in v: the Ship of Theseus undergoes a gradual replanking process until it comes to be made of planks $q_1,\ldots,q_n$, while planks $p_1,\ldots,p_n$ are assembled into another ship, the pseudo Ship of Theseus, and on the fateful day in v it is this second ship which sinks. The sceptic may then argue: in u, the sinking of the Ship of Theseus, e_1, is the 'sum' of the sinkings of planks $p_1,\ldots,p_n$, which sum we call e_2 (so $e_2 = \text{sum}[e_{p_1},\ldots,e_{p_n}]$). It is hard to deny that e_2 occurs in v. But then by transitivity of identity, e_1 occurs in v and is the sinking of the pseudo Ship, since it is this ship's going down which, in v, is identical to the event $\text{sum}[e_{p_1},\ldots,e_{p_n}]$.

It would be somewhat wooden to dispute the example on the grounds of its assumptions about transtemporal and transworld identity for ships (questions about the transworld heirlines of events may presume upon the transworld heirlines of the relevant non-events). Instead, Lombard attacks the premise that e_1 is identical to $\text{sum}[e_{p_1},\ldots,e_{p_n}]$, by claiming that e_1 has a modal property which e_2 ($= \text{sum}_i[e_{p_i}]$) lacks: it is essential to e_2 to have each e_{p_i} as a part, but since the Ship of Theseus could have been made of planks $p_1,\ldots,p_{n-1}, q_n$, e_1 might have had e_{q_n} as a part instead of e_{p_n}. Or, more simply, we can just say that e_1 could have occurred with-

out e_{p_n} occurring, while e_2 could not have so occurred. However, like all arguments for distinctions established by possible discernibility, this objection may be met by disputing the underlying transworld identity judgement, in this case that e_1 occurs in some worlds where the Ship of Theseus has q_n in place of p_n. But someone who denies transworld identity between e_1 in u and the sinking of the Ship of Theseus in any world where it has q_n must offend our stricture that transworld distinctness cannot be imputed where there is nothing intrinsic in which the difference consists. For, with respect to some worlds, there will be no difference at all between the sinking of the Ship of Theseus in those worlds and its sinking in *u:* the sinking will occur in the same manner, at the same time, in the same place, and will involve the same ship *ex hypothesi*. That is, the natures of the events, at the very least, are the same in the different worlds, so there is nothing in which a numerical difference between any pair of them could consist. In this fairly minor way, then, our thesis that facts about identity must be intrinsically grounded is already relevant.

The anti-essentialist may try a different manoeuvre, which Lombard does not consider in this context, which involves putting forward a sufficient condition for transworld event identity in terms of causes, that if *e* has exactly the causes in *u* which *f* has in *v*, then *e* and *f* are the same event (since he is an anti-essentialist, he will deny the converse, but the objection could also be made by an essentialist with an alternative account to Lombard's). According to this condition, the following is a counterexample to Lombard's view. Let *e* be an actual event involving an object *x* at rest at place *p* at time *t* and moving away from *p* thereafter, the causes of this movement being events which are impacts of other objects, these events contiguous in space and time to *e*. Since it is contingent that *x* is at *p* at *t*, we can choose a world *w* with another object *y* at *p* at *t* such that *w* is as similar as possible to the actual world up to and including *t*. Then the very same impacts of other objects which caused *e* in the actual world also occur in *w*, but in *w* they bring about a movement of *y* (a movement of the same type, even if not the same movement). By the objector's sufficient condition for event identity, therefore, *y*'s movement in *w* is *e*. So *e* could have minimally involved *y* instead of *x*.

The strangeness of this sufficient condition appears to be traceable to an intuition about what the intrinsic features of an event

relevant to its crossworld individuation are. In the case of material objects and sets, such intrinsic features were invariably in some sense 'internal' to the thing, either its members, or its constitution and design, or its origin and kind. Although an event's causes are obviously not causally isolated from it, they are not internal to it in a way analogous with the material object case, while the minimal subject clearly is internal. For this reason, the idea that having the same causes is sufficient for crossworld identity of events may be rejected.

Lombard's second essentialist thesis is that the time of an event is essential to it, where by 'time of an event' we mean the interval exactly occupied by it. Lombard argues for this thesis by confronting the sceptic about it with a *reductio* [1982, pp. 9–13]. If an event can occur at different times within different worlds, then if we take a pair of events e_1 and e_2 occurring in a world u which are intrinsically indistinguishable in u except by the time of their occurrence (Lombard uses the example of a marble's rolling twice along the same route at the same speed) there must be worlds in which e_1 and e_2 switch times, so that e_1 occurs at t_2 and e_2 at t_1. From among those worlds, choose v as similar as possible to u except for the switched times of occurrence; then it seems that the only substantial difference between u and v will be in the identity of the events which occur at t_1 and t_2 in the two worlds which involve that marble's moving, since the intrinsic indistinguishability of the events implies that the antecedents of one do not have to be intrinsically distinguishable from those of the other. Thus, although Lombard does not put it in this terminology, we see that rejection of essentialism about time leads to ungrounded differences between entities and resulting inadmissible distinctions amongst worlds much like that which we arrived at by the Four Worlds Paradox.[13]

13. Lombard's own view is that what is wrong with taking u and v to be numerically distinct worlds is that it contravenes a supervenience principle for events which he states thus: (E) Possible worlds cannot be alike with respect to the truth and falsity of propositions concerning objects, the properties which those objects have, and the times at which those objects have those properties, and yet be unalike with respect to the truth and falsity of propositions concerning events. Lombard thinks that this principle has all the plausibility of an analogous supervenience principle for sets: (S) Possible worlds cannot be alike with respect to the truth and falsity of propositions concerning the existence of objects, and yet be unalike with respect to the truth and falsity of proposit1ons concerning the existence of sets. Each principle, he says, expresses the doctrine that there is nothing 'hidden' about sets or events, nothing more to know about

Lombard's strict essentialism about the time of an event could be relaxed by allowing a small amount of change, so that the same thunderstorm could have started a little later, but not a day later. Such a modification would require application of the degree-theoretic and counterpart-theoretic apparatus to modal discourse about events, but this does not constitute a very great change to Lombard's approach. A more radical worry concerns whether or not Lombard's theory does capture genuine aspects of our concept of event. Crucial to the construction of the theory is Lombard's use of the *de re*/*de dicto* distinction to reply to certain objections, but while the content of the distinction is well understood in the case of quantification over ordinary objects, a sceptic about essentialism about events may feel that too much is already conceded if a distinction is allowed between its being possible for the death of Socrates to have occurred earlier and the death of Socrates' potential to have occurred earlier. Such a *de re* sceptic may say with some plausibility that although there are worlds which are distinguished in terms of where, when and how Socrates dies at them, there is no such distinction as the one Lombard aims to capture.

For the *de re* sceptic, Lombard's events are philosophers, fabrications; Lombard has done no more than isolate three features of events (we assume the type of an event will also be said to be essential to it) and attribute to the event the transworld identity conditions of the set of those features. This enables us to make sense of the *de re*/*de dicto* distinction as Lombard wishes us to, but only because his criteria have stipulatively introduced an entity with transworld being; only then can such a sentence as, say, 'the death of Socrates might not have been the death of Socrates', be given an interpretation; pretheoretically, it is uninterpretable or obviously false. The source of such *de re* scepticism is not hard to uncover. It lies in the fact that ordinary event sentences of the sort

them once we know their natures (in the case of sets, their members) [1982, pp. 13–14]. But principle (S) does not have this effect, since it is compatible with the same set having different members in different worlds with the same domain of non-sets; furthermore, it entails a restricted version of Set Existence, that if all the members of x at some world u exist at v and v has the same domain of non-sets as u, then x exists at v. But Lombard explicitly rejects the idea that the existence of a set 'follows simply from the existence of the objects which are, in fact, its members' (*op. cit*. p. 14). Hence Lombard really needs (MR) and (CE) to express his view about sets, while the supervenience principle (E) appears to be a *consequence* of the idea that the intrinsic features of an event which give its individual essence are its subject, time, and type, rather than a principle which *leads* to this view.

that occur in non-philosophical discourse (this excludes e.g. 'every event is identical to itself') invariably permit paraphrase by other sentences in which the appearance of reference to an entity, the event of such-and-such happening, is eliminated. Thus, instead of saying that the assassination of Kennedy was the work of a conspiracy, we can say that there was a conspiracy to assassinate Kennedy; instead of saying that there were three attempts to scale Everest by the southwest face before the first successful one, we can say that three times it was unsuccessfully tried to scale Everest, where 'three times' is a temporal operator with a rather obvious evaluation clause in standard tense-logical semantics; and so on.

A full defence of such *de re* scepticism would require the production of a translation procedure, like the one which has been given here for translating a wide range of possible worlds sentences back into modal language; and, at present, there is little likelihood of such a scheme, since it is unclear what the procedures are which we employ to come up with paraphrases of particular event sentences, and unclear how rich the base language would have to be to permit the translation of all the event sentences we would want to be able to interpret. However, there is a well-known case to which the *de re* sceptic can point as illustrating what he has in mind here, the case of what Quine [1963, Ch. 1] calls 'virtual set theory'. Many statements which use terms of the form 'the set of *F*'s', i.e., statements which would naturally be formalized with set abstracts, can in fact be formulated equivalently without any apparent reference to sets. For instance, the statement 'the set of *F*'s is a subset of the set of *G*'s', i.e.,

(a) $\{x: Fx\} \subseteq \{x: Gx\}$

makes inessential reference to sets, for we could simply say "all *F*'s are *G*'s"

(b) $(\forall z)(Fz \rightarrow Gz)$

instead. There is a principle at work here which is generalizable as follows. The language of sets contains a special predicate $\in$ for set membership and a special variable-binding operator, the set abstraction operator $\{\_: \_\}$. To permit the reverse translation of set-theoretic sentences, then, we need to be able to eliminate both the special predicate (which, since it makes sense only when applied to sets, would not be wanted in the base language) and occurrences

of set abstracts. And, up to a point, this can be effected by the following rule:

(c) $r \in \{s: \varphi\} = \varphi[r/s]$

where r and s are variables, r and s not bound in φ, and $\varphi[r/s]$ is φ with r substituted throughout for s. Recalling that (a) abbreviates

(d) $(\forall z)(z \in \{x: Fx\} \rightarrow z \in \{x: Gx\})$

it is easy to see that (c) yields (b) for (a). It is also easy to check the translatability of other simple expressions, and thus of more complex formulae built out of them; for instance, "the set of *F*'s is the same as the set of *G*'s" becomes "all and only *F*'s are *G*'s".

The limits of this translation procedure are reached when we consider set-theoretic statements in which set abstracts occur on the left hand side of $\in$. It is certainly possible to go on eliminating abstracts, by using

(e) $\{x: Fx\} \in \{x: Gx\} = (\exists z)(z = \{x: Fx\} \ \& \ z \in \{x: Gx\})$

and then applications of (c) and the biconditional treatment of set identity. But (e) introduces an existential quantifier, which says that something is identical to the set of *F*'s; therefore (e) cannot be employed by one who seeks to show an ontology of sets to be eliminable.[14] However, so long as we restrict our theory of sets to those assertions for which an adequate reverse translation based on (c) is available, then we may reasonably be said not really to have an ontology of sets: thus Quine's use of the adjective 'virtual'. And provided our sets are virtual entities only, ones which we introduce by introducing certain means of expression subject to (c), we are not compelled by the arguments of earlier chapters to settle for any one particular account of their essences: there is no fact of the matter about the modal properties of these virtual entities. This is not to impugn the theory of essences which we have advanced for sets, of course, since the 'iterative' conception of set for which the essentialist theory was developed goes well beyond what is consistent with sets being virtual.

In the case of events, it is conceivable that some of the technical terms of Lombard's theory are analogous to $\in$ in preventing refor-

14. Here I am assuming the standard interpretation of $\exists$. If $\exists$ is given a "semi-substitutional" interpretation [Parsons 1971], this would have to be qualified.

mulation of sentences apparently involving an ontology of events. But this would have to be shown in some detail: given the variety of non-event-invoking means of expression illustrated three paragraphs back, there is initial plausibility in the thought that, even for such terms, clauses analogous to (c) will be available; indeed, Lombard's canonical form for terms referring to events might be a key component of the required reverse-translation scheme. If this is so, then the kind of distinctions between this very event's occurring in a world, on one hand, and an event just like it in certain respects occurring in a world, on the other, will not be a distinction he is entitled to, since all real distinctions between worlds will be manifested at the level of the paraphrasing sentences. From this point of view, what Lombard has done is just to erect a *de re*/*de dicto* distinction upon differences manifested at this level, and while this enables us to assign determinate truth-conditions to sentences regimented as *de re* modal sentences quantifying over events, there is no fact of the matter at dispute between Lombard and someone who chooses to erect the *de re*/*de dicto* distinction in a different way, a way which results in disagreement with Lombard's assignment of truth-conditions to those regimented sentences.[15] If what has gone before is at all correct, there is thus the greatest possible difference between events, on the one hand, and material things and non-virtual sets, on the other.

15. Every non-*de re* sceptic would agree that Socrates' actual death occurs in a world *w* just like the actual world until after the time of his death, whereupon Socrates is resurrected to die again, since by the branching conception there is no real application of transworld identity in making this judgement. But if a *de re* sceptic about events appeals to anti-realism about them to justify his scepticism, and then gives a 'best candidate' criterion for their transworld heirlines, a criterion which would allow an event to change its subject, time, and type from world to world, it is hard to see that Lombard has presented any consideration which convicts such a sceptic of making a mistake. Evidently, the claim that there is the greatest possible difference between events and ordinary objects in this respect rests on the assumption that anti-realism about events is a position which is open to us, while anti-realism about common-or-garden objects is not.

Chapter 9
The Justification of Modal Concepts

Non-cognitivism

IN THE foregoing chapters, we have investigated a variety of modal theses whose formulations all involve application of the concepts of broadly logical possibility and necessity, but we have at no point queried the legitimacy of these concepts themselves (as opposed to certain interpretations of them, e.g., the quantifier readings of $\Box$ and $\Diamond$). Since philosophers have in fact expressed scepticism about whether there is any well-defined content to be attributed to these concepts, e.g., Quine [1966, 13], it would be unsatisfactory to conclude our discussion without addressing this issue, especially as it is not unreasonable to expect that our attempts to establish claims which employ these concepts should have increased our understanding of their nature, which in turn should help us expedite a defence of them. Nevertheless, it should be emphasized from the outset that this chapter is of a much more provisional nature than any of the preceding ones; specifically, at some points we will be able to do no more than describe problems for the approach we will take, since to tackle those problems adequately would require another book.

A justification of concepts of broadly logical modality should have two components, metaphysical and epistemological respectively. The first, metaphysical, component is a theory of what it is which determines the modal status of truths and falsehoods; a biologist can say what it is which makes 'Elizabeth II is descended from George VI' true, but it needs a philosopher to explain what feature of reality makes it necessary that Elizabeth II, if she exists, is descended from George VI. The second, epistemological, component comprises an account of how we come to know what is

necessary and what is contingent, and by requiring this second component, we impose a constraint on the first: no metaphysical account which renders it impossible to give a plausible epistemological theory is to be countenanced.

A philosopher who holds that in some respectable sense there are features of reality which make modal judgements true or false may be termed an *objectivist* about modality, provided it is understood that there is nothing in the use of this label which implies that an objectivist cannot appeal to facts about psychology; both "internal" and "external" features of reality are *prima facie* suitable for an objective grounding of modal truth. The position which affords the proper contrast with objectivism is that of the non-cognitivist, according to whom the content of modal propositions is such as to render the notions of truth and falsity not genuinely applicable to them. The best-known non-cognitivist theory of a modal concept is Hume's theory of physical or causal necessity. According to Hume, if a subject undergoes repeated experience of C-type events being followed by E-type events then C-type events acquire the capacity to induce certain states of mind in the subject, and the feeling of this state of mind arising on perception of a given C-type event prompts the subject to hold that the particular E-type event which he perceives to follow, does so as a matter of causal necessity. On one construal, Hume's idea is *not* that the belief that a particular *e* follows a particular *c* of necessity can be *analysed* as the belief that *e*'s following *c* is accompanied by the arising of the relevant state of mind, for that would be an objectivist view – it would be a matter of fact whether or not that state of mind did arise. Rather, his doctrine is that the content of the first belief which distinguishes it from the belief merely that *e* follows *c* is a component which expresses (without asserting) the fact that the subject is in the appropriate state of mind; in Dummett's terminology, when we say that *c* causally necessitates *e*, we are making a quasi-assertion rather than a genuine assertion.[1] Of course, in speaking of the repeated experience types 'giving rise to' a capacity in future instances to 'induce' states of mind, double use is being made of the concept being explained; but there is no circularity

1. See Dummett's distinction between assertion and quasi-assertion, [1973, pp. 352–355], where he writes, '…we may well suspect that such [non-cognitivist] theories represent rather cheap attempts to resolve difficult philosophical problems by ruling them out of order.'

here, since the explanation can be applied over again to the terms of the theory itself.

Hume did not much address himself to the question of either strictly or broadly logical necessity, but one can see, at least very roughly, how he might have extended his account of causal necessity to stronger modalities. When we assert 'Necessarily, *A*' we would be said to be asserting *A* and expressing the fact that apprehension of the proposition expressed by *A* induces a certain state of mind in us. The capacity of the proposition to induce the state of mind would be one it has in virtue of some feature of it with the role which corresponds, in the case of causal necessity, to that of the regularity in the subject's experience of C-type events being followed by E-type events; this feature is the one which causes the impressions in us from which we derive the concept of necessity. But what feature of propositions endows them with this power? Hume comes close to addressing this question in the following passage, noted by Stroud [1977, p. 241]:

> Thus as the necessity, which makes two times two equal to four, or three angles of a triangle equal to two right ones, lies only in the act of the understanding, by which we consider or compare these ideas; in like manner, the necessity or power, which unites causes and effects, lies in the determination of the mind to pass from the one to the other.

Hume's view is that we then make a mistake: we project something essentially "inner" onto the external world, and acquire the mistaken belief that the concept of necessity we have applies to propositions in virtue of objective properties of ideas and, as a consequence of this, we mistakenly believe that modal judgements can be true or false.

We can explain this further by reference to the primary/secondary quality distinction. We have already noted that there are two distinct ways of drawing this distinction, either in terms of whether or not a property is a disposition to produce in us sensations with certain features, or in terms of whether or not the property is one which is really "in" the object itself, as opposed to being possessed by the object only in virtue of features of sensations resulting from perception of the object. Up to a point, Hume's view of necessity is like the view of secondary qualities held by someone who draws the primary/secondary distinction in the second way, but the comparison is slightly misleading, for such a person need not be a non-

cognitivist about secondary quality terms; he can say that it is literally true that an American mail-box is blue since it is literally true that perception of such a thing has certain effects on us, while Hume, on the present construal, denies that it is ever literally true that a proposition is necessary, since this does not mean that the act of the understanding consisting in our reflecting on it has certain effects on us. The relationship of the effects to the content is rather like a Humean view of the relationship of the effects of perception of a distasteful scene to the content of a judgement of moral disapproval about the relevant goings-on: the effects are certain psychological states in us, and part of the content of the judgement, the "evaluative" part, expresses but does not assert the fact that we are in such states, being in which motivates us to behave in certain ways by producing desires or aversions in us, the exact nature of which turns on contingent features of our psychological make-up. So a non-cognitivist might try to elucidate the difference in content between A and $\Box A$ in terms of the attitudes towards the proposition A which arise in us upon the obtaining of the psychological states whose presence is expressed or signalled (in the way that a cheer expresses or signals approval) by the use of the modal prefix.[2]

Another well known non-cognitivist account of a modal concept is Wittgenstein's theory of mathematical necessity (here I follow the interpretation of Wright [1980, Chs. 19–23]). Wittgenstein was an extreme conventionalist about mathematical necessity, in that he held that there is no sense in which the premises of a mathematical proof *necessitate* its conclusion: any claim of necessity merely lays down a new convention. His position contrasts with that of the moderate conventionalist, who holds that once certain general conventions governing the logical constants are accepted (rules of inference) together with certain other conventions about non-logical words (basic axioms) the conclusion of any proof is necessitated by its premises in virtue of each step's being an application of some already accepted general convention; thus any necessity which is not immediately a matter of convention ultimately reduces to conventions. The difficulty for the moderate conventionalist is to explain in what sense the correctness of a particular step in a proof is a "consequence" of an antecedently

2. See Stroud [1977, p. 821] and Peacocke [1982] for discussion of contemporary versions of Hume's non-cognitivism about causal necessity.

accepted convention: if this is a further convention, we have an apparently unending regress of conventions, while if it is not, we have a relationship of necessary consequence which is not subsumed under any convention. The difficulty for Wittgenstein's view, on the other hand, is to square it with obvious facts about mathematical practice, especially the feeling of any competent person following a proof that he is compelled to accept each new line as a consequence of the earlier ones.

The view which Wright extracts from Wittgenstein is that each mathematical theorem is like a rule of a game: the rules themselves cannot be said to be true or false, and as rules are added (in the hope of producing a better game), new constraints are thereby imposed on what is admitted as correct play. But we add new rules to a game as a matter of decision, so by analogy Wittgenstein held that there is an element of decision in accepting a new proof in mathematics. Explicit assertions of necessity he interprets as prescriptive: to say that it is necessary that p is to urge a certain policy, namely, that of not admitting any description such that were some situation to be accurately described in that way, its obtaining would refute some non-modal generalization "appropriately" related to p. For example, to say that it is necessary that $7 + 7 = 14$ is to urge, among other things, that we do not ever accept "there are 7 apples in A's basket and 7 apples in B's and 15 in both", since this conflicts with: if there are 7 *F*'s in one container and 7 *G*'s in another, then there are 14 *F*-or-*G*'s in both. Since prescriptions of policy are not true or false, neither is any sentence of the form $\Box p$.[3]

A non-cognitivist theory of apparently fact-stating discourse is usually a "last resort" account, deriving from some general doctrine about meaning. To the degree that one finds the non-cognitivism unbelievable, one will prefer to reject the general doctrine but, of course, the right to find the non-cognitivism unbelievable must be earned by a critique of the semantic theory (or in Wittgenstein's case, semantic scepticism) underpinning it. Perhaps no one is inclined to Hume's semantics nowadays; but addressing Wittgenstein's scepticism is another matter, and is one of the tasks we mentioned earlier as requiring a separate book. At this point, therefore, we will just continue the taxonomy of accounts of modal concepts at variance with the one we will eventually adopt.

3. Wright usefully summarizes his interpretation in [1980, pp. 410–11].

Quasi-psychologism

Since the view we will investigate next does not claim to provide an *analysis* of possibility and necessity, we call it quasi-psychologistic rather than psychologistic; nevertheless, the main idea of the approach is that the modal status of truths and falsehoods is ultimately grounded upon human intellectual abilities. For example, Stroud [1977, p. 245] has written that to explain necessity and possibility, we need to pursue 'questions about the human mind and its capacities'; and these explanations must appeal to 'empirically discoverable, natural facts about us'. One interpretation of these claims, the preferred one, is that they are of a piece with the Wittgensteinian non-cognitivism sketched above. Wittgenstein suggests that it is a matter of decision whether or not to accept the conclusion of a proof, yet it is clear that there is no disagreement between trained subjects on whether or not a given proof should be accepted (*modulo* the constructivist/platonist dispute), which may seem surprising if decisions are called for. But the phenomenon of agreement can be explained by the brute propensity of humans who have been trained in certain ways all to go on in the same way, a propensity itself residing in, or being, an empirically discoverable natural fact about us. However, it is also possible to interpret Stroud's remarks as advancing the view that the necessity of a truth follows semantically from our inability to make sense of the supposition that it is false, an inability which itself admits of an explanation which will not cite features of the subject matter of the proposition in question, so that facts about the mind's capacities are genuinely the basic ones. Thus modal statements will be true or false depending on facts about our abilities.

Wiggins [1980, p. 106] explicitly urges a quasi-psychologistic approach:

> What we have here…is not a reduction or elimination of necessity or possibility…but the following elucidation of possibility or necessity *de re*: (i) x can be φ iff it is possible to conceive of x that it is φ; (ii) x must be φ iff it is not possible to conceive of x that it is not φ.

So, according to Wiggins and the second reading of Stroud, the first step in an analysis of possibility and necessity is to look to what we can and cannot make sense of, or conceive (in some sense of 'can' and 'cannot' other than the broadly logical, one assumes).

It seems that this approach is either circular or else extensionally incorrect: there are some impossibilities which it classifies

wrongly. To see this, let us agree with Wiggins that no android could have been human, and conversely. Suppose now that we are visually presented with a creature which for all we know may be either android or human, but which is actually an android. Then it is epistemically but not logically possible that this creature is human, and so one might say that conceivably it is human; thus, *ad hominem*, Wiggins is wrong to equate what is conceivable and what is possible. However, this argument turns on interpreting 'conceivable' as 'epistemically possible', which may be at once too broad and too narrow an interpretation. It is therefore reasonable to ask that this sense of 'conceivably' be put aside as not what Wiggins means (nor what Stroud means, if we identify what we can conceive of with what we can make intelligible to ourselves).

There is another sense of 'conceivably' which is not so straightforwardly tied to epistemic considerations, and which, again, is not what Wiggins should mean, since the impossible is also conceivable in this sense. Kripke sometimes writes that we can conceive of finding out that such-and-such is the case, even though in fact we know that it is not the case, and can infer from that that it is necessary that it is not; see Kripke [1972, pp. 269, 313–14, 318–19]. Moreover, although 'conceivably' is intensional, it does seem that if we can conceive of finding out that p then we can conceive of p's being the case: someone who holds that p is inconceivable, if he also agrees that p's being found out implies its being true, could at best conceive of a person claiming to have found out that p, or of an investigation which results in such a claim, a claim which then survives our best efforts to refute it; we can say this without endorsing the false general principle that if I is an intensional and F a factive operator, then $IF(p)$ implies $I(p)$ (consider 'Smith hopes that Jones knows that p'), the point being specific to the meanings of 'it is conceivable that' and 'it is discovered that'. Then, since we can certainly conceive of finding out that the android is human, or better (in Wiggins' style, since 'the android is human' is a *de dicto* impossibility), since we can conceive of finding out, concerning the android, that it is human, then we can conceive, of the android, that it is human, even though we know it is an android.

An explanation of how this is possible should be sought. The explanation is that, in this sense, 'It is conceivable that p' means that the hypothesis that p does not by itself contradict any principle

which is constitutive of the content of a concept involved in the proposition that *p*: refusal to rule out *p a priori* is *not* indicative of failure to grasp some of these concepts. (In fact, there are really two explanations here, depending on whether we require that there should in fact be no such contradiction, or just that none should be evident to the subject, but in this context the ambivalence is irrelevant.) Thus, in the example, we can conceive, of the android, that it is human, because the android is presented to us perceptually, and it is not required to have a demonstrative thought about it that we think of it as an android, even if in fact we know that it is one. Someone who holds that it is necessary that water is H_2O can allow that it is conceivable that water is not H_2O for a similar reason, for one can have a way of thinking of the substance water which allows for thoughts about water that do not involve thinking of it as H_2O; in view of the way use of the word 'water' is taught, by ostension of samples, most persons' concept of water will be like this. And, in the most familiar example, even after we have discovered that Hesperus and Phosphorous are the same planet, it is still conceivable that they are distinct, because we have two different ways of thinking of that planet which, even when they are conjoined, do not imply the identity.[4] If we say that *p* is *strongly a priori* iff failure to assent to *p* is indicative of failure to grasp some concept involved in the thought that *p* or failure to perform some simple logical inference, then the present senses of 'conceivably, *p*' are comprised in 'it is not strongly *a priori* that not-*p*'.

There is one special case where we might expect a close relationship between conceivability and broadly logical possibility, the case of sets. Wiggins' acceptance of essentialism about set-membership but scepticism about the necessity of origin arises out of his attempt to associate conceivability and possibility, for it is plausible that when we think of the set $X = \{a, b\}$, we do indeed think of it as the entity whose members are exactly *a* and *b*, and it is therefore inconceivable that it should lack these members or have any others. But in this sense of 'conceivable', as we have just seen, the other component of Wiggins' essentialism, according to which

4. To say that a way of thinking of a planet is associated with a particular name of it is not to subscribe to a description theory of names, if that is a theory about what it is for an object to be the referent of a name. What makes an entity *x* the intentional object of a way of thinking will have to do with the relations in which the thinker stood to *x* in acquiring that way of thinking of it. See Evans [1982, pp. 14–22].

it is of the essence of a thing to belong to the kind of which it actually belongs, fails to follow: we can conceive, of the android, that it is human. And it is plausible that attempts further to refine conceivability to get rid of this consequence will bring with it other essentialist claims about which Wiggins is sceptical, such as the necessity of origin.

The main problem for Wiggins' approach is to come up with some further sense of 'conceivably' which does capture broadly logical possibility, a sense which can be characterized, if not in the fundamental terms of the theory of psychological capacities, at least in terms which are sufficiently far removed from those which a non-psychologistic objectivist might use to explain broadly logical modalities. But in the prevailing absence of any detailed quasi-psychologistic theory, it is very hard to see where such a characterization of 'conceivably' is to come from. The etymology of the word demands that at least consistency of concepts be imposed, but we have seen that a further restriction is required, and there is apparently no other psychological capacity which will exclude just what needs to be excluded. For instance, the faculty of pictorial imagination is powerless to exclude the conceivable impossibilities: we can certainly picture discovering, of the creature in front of us, that it is human. Thus the whole approach in terms of conceivability looks unpromising.

The theory of content

The explanation of the second sense of 'conceivably' in the previous section would be regarded as useless by many philosophers, on account of its appeal to principles which are supposed to be constitutive of the content of the concepts appearing in the hypotheses asserted to be conceivable. The problem is with the notion of the 'content' of a concept, a notion said by Quine, for instance, to be itself without content. Since we will attempt to explain necessity in terms of the content of concepts, we must at this point address, or at least note, Quine's views. An *a priori* principle constitutive of the content of a concept, say the concept of being an *F*, is intended to contrast with *a posteriori* beliefs about *F*'s, which are beliefs whose possession requires prior mastery of the concept. However, Quine has argued influentially that this alleged distinction between *a priori* truths about the concept of *F*-hood and empirical truths about *F*'s is not one which survives careful scrutiny.

Chapter 9: The Justification Of Modal Concepts

Anyone with a broadly empiricist outlook who tries to maintain

the distinction will attempt to do so by taking the *a priori* principles to be principles about how application of the concept should be constrained by experience (perhaps relative to the application of other concepts), but, according to Quine, although this is what such propositions should be like, there are no propositions which express conceptual truths in this sense. For, if there were, then given a sequence of experiences recalcitrant with respect to our current views about *F*'s, those conceptual truths, if they really are such, should dictate which of the empirical propositions about *F*'s that we currently believe should be abandoned, and which empirical propositions about *F*'s we should come to believe instead; or, at least, they should dictate that certain revisions are open while certain others are not, even if they do not determine a unique candidate. But Quine argues that, in fact, no particular propositions about *F*'s have this role: the impact of any sequence of experiences can be distributed throughout the range of propositions a thinker may be disposed to assert prior to undergoing that sequence, in an endless variety of ways. For example, a scientific theory may consistently be held true regardless of the evidence to the contrary, if the theorist is willing to continue to append *ad hoc* hypotheses and complicate other parts of the theory to explain away the awkward evidence: this procedure could be carried even to the point of abandoning logical principles. Thus, given some proposition purportedly stating an *a priori* principle constitutive of the concept of being an *F*, we could choose a revision of our theory of *F*'s to accommodate some experiences of *F*'s, a revision that involves a rejection of the alleged constitutive principle.[5]

In assessing the plausibility of Quine's views, a task which cannot be pursued very far here, one should separate the relatively uncontroversial idea that experience confronts whole theories rather than single hypotheses on a one-to-one basis, from the much more controversial view that any adjustment of a theory is open to us in advance. Someone who accepts the holism of the first thought is no more committed to denying determinate content to the individual hypotheses which comprise the theory than is someone who holds that individual hypotheses are the primary

5. The classic statement of Quine's position (qualified in later work) is 'Two Dogmas of Empiricism', on which my account of his views is based; see paper (ii) in Quine [1961]. In the comments that follow I am indebted to part I of Dummett [1978, pp. 375–84].

bearers of content committed to denying determinate content to the words which make up the hypothesis. In each case, the content of the part can be identified with its contribution to the content of wholes in which it may occur. The second element of Quine's view decrees that nothing can be isolated for individual hypotheses as comprising such a contribution, but this appears to be problematic. First, it is unclear that sense can be given to the notion of an experience's being *recalcitrant* for a given theory if no adjustment is ruled out in advance; the logical principles which determine what is recalcitrant and what is not seem very different from the working hypotheses of the theory itself. If one can dissolve recalcitrance by treating these principles as if they were mere hypotheses and abandoning them, one begins to lose one's grip on what the goal of inquiry is and why it is pursued at all. Secondly, Quine's no-specific-bearing doctrine seems to contradict the facts about the practice of science, for in most cases scientists do not have much difficulty in determining to which part of a given theory particular experimental evidence is most relevant. Perhaps philosophers of science have only recently begun to develop the sophisticated analyses of the relation of evidential confirmation which would be required for a full answer to Quine on this point; see, e.g., Glymour [1980, pp. 110–23].

The proposal that the content of a hypothesis can be identified with its contribution towards the content of any theory to which it belongs is at best a schema of a position, until we have some identity criterion for contents to articulate the kind of contribution envisaged. An interesting proposal about this has been made by Hartry Field [1977]. Note that for Quine's point about different ways of revising a theory to obtain, we have to be considering two theorists, or a single theorist at different times; for, obviously, one theorist can in one revision revise a theory in only one way. So we might suggest that for a subject S at time t, propositions p and q are the same iff the experiences of S through t either warrant (for S) the holding of both or fail to warrant (for S) the holding of both. But this is clearly far too simple, since it implies that all the propositions S holds at t have the same content: we also wish to consider what S would say about p and q under the supposition that his experience is like *this*, or like *that*, etc. Moreover, the division of propositions into two classes by a given sequence of experiences, those warranted by it and those not, is too coarse. Experience con-

Chapter 9: The Justification Of Modal Concepts

firms or disconfirms hypotheses, and confirmation comes in degrees. Field combines these points into the following criterion of sameness of content (in his terminology, sameness of conceptual role) of p and q for S at t: p and q are the same iff for any proposition r, S's subjective conditional probability for p given r is the same as his subjective conditional probability for q given r. That is,

(1) $p = q$ iff $(\forall r)(\text{PB}(p|r) = \text{PB}(q|r))$

where PB is S's subjective conditional probability function at t; $\text{PB}(x|y)$ is the probability x has for S given y (which need not be the probability S would ascribe to x were he to come to believe y).[6] We might have tried to restrict the range of the variable r to 'observation propositions', that is, to propositions which would merely state how things look or have looked to S, but (1) allows for the conceivable case in which S's subjective conditional probability function is such that distinct non-observational propositions are assigned the same probability come what experiences may, but whose conditional probabilities come apart given some non-observational proposition.

Criterion (1) leads to an elucidation of the relativized notion of a principle constitutive of a concept for a subject S at a time t; this would be a principle which, for S at t, has maximum probability regardless of what proposition r is given, and which is relevant to the account of S's reasons for making the assignment of conditional probabilities he does in the cases of propositions which involve the relevant concept and which, for some r, are not maximally probable given r. Here we have no very radical departure from Quine; for instance, as time passes, it is still open to us to say either that the subject is changing his beliefs, or that he is altering the content of his concepts.

Criterion (1) may appear obviously circular, since in attempting to explain identity for propositions, it quantifies over propositions, so that if the right-hand-side of (1) were applied in an attempt to settle an identity question, the verdict could turn on whether or not univocal substitution for r is being made, which in turn could

6. The subjective conditional probability of 'this die will show a 3' given that it will show either a 3 or a 5 is, for most people, ½. But the actual world might be such that if I were to come to believe that the die had shown a 3 or a 5, then I would acquire additional beliefs as a result of which I would not agree that the chances of it being 3 are 50-50.

depend upon whether or not p and q are the same proposition. However, a similar situation arises with a number of "synthetic" identity criteria for categories of entity: that material objects x and y (of the same sort) are identical iff for any material object z and time t, x is in spatial relation R to z at t iff y is in spatial relation R to z at t; that events e and f are the same iff for any event d, e is a cause (effect) of d iff f is a cause (effect) of d; the functionalist criterion for identity of mental states is also similar in structure. Such criteria merely specify a relational framework within which we individuate the relevant entities in a manner consistent with the criterion; the criteria give the terms in which we specify that in which identity and difference for those entities consists.

To give a non-relative account of the distinction between the conceptual and the empirical, therefore, what we have to do is to specify a relational framework which comprises the thoughts of different subjects, or of the same subject at different times. We can isolate some problems which face us here by considering why Field's criterion cannot just be generalized in the most straightforward way: the obvious objection is that two subjects may well attach the same content to some proposition but disagree about conditional probabilities for it since they disagree about background facts. But there is a natural way to avoid this objection, for if the two subjects do attach the same content to some proposition, we would expect this to be manifested counterfactually: if they were to agree in their background beliefs, they would agree in their assignments of conditional probabilities to that proposition. So the suggestion is that for any propositions p and q such that S believes p and S' believes q, we should say:

(2) $p = q$ iff for any collection of background beliefs B, if S and S' were both to accept B, then for any r, it would be that $\text{PB}(p|r)$ for $S = \text{PB}(q|r)$ for S'.

As it stands, (2) is in need of explanation and refinement. For example, for each proposition p, we have to find some way of circumscribing what facts are background relative to p, so as not to include in the 'background', e.g., S's beliefs about the conditional probability of p, for each r. But the more pressing question is whether there is reason to hold that, in principle, no such counterfactual criterion can succeed. The problem is again one of apparent circularity, since we are presuming on the notion of S and S'

having the same background beliefs B, and also using the variable r to stand for one and the same proposition as the given condition for the assignments both S and S' make. The presumption of same background beliefs certainly imports an extra element of complexity to this criterion over and above what was present in the earlier, intrasubjective criterion of identity at a time, and raises questions we cannot possibly pursue adequately in the present context.[7]

There is a further problem with the present line of inquiry, having to do with the suitability of the notions being employed for substantiating a notion of the content of a concept our grasp of which permits us to arrive at principles constitutive of that concept. As Wright has written [1980, pp. 354–5]:

> We want to attribute to ourselves a capacity reflectively to apprehend impositions and constraints which the manner in which we understand particular expressions places upon us…the capacity is thus…essentially a capacity to discern the character of one's own understanding. [But] Wittgenstein repudiates the view that each of us may regard himself as knowing reflectively what kind of application of an expression conforms to the meaning he attaches to it.

As with his conventionalism about mathematical necessity, there is again the problem of squaring this claim with the subjective phenomena. Thus a philosopher, in investigating whether a causal judgement is always equivalent to some related counterfactual, may consider a case where the relevant counterfactual is true, e.g., 'if his sister had not had a child, Smith would not have become an uncle', and conclude that there is no equivalence, since the application of 'causes' which conforms to the meaning he attaches to it is that his sister's giving birth did not cause Smith to become an uncle. Wittgenstein would claim that we just find ourselves with a brute propensity to say one thing rather than the other, and that the hypothesis of a capacity to apprehend the content of one's concept of causation and employ that apprehension in the testing of philosophical analyses, does no work. But we do have the practice of testing an intuition against a range of cases in a search for consistency or for an answer to a given hard case, which it is natural to describe as trying more accurately to apprehend the content of the

7. In the last paragraph I am indebted to an unpublished manuscript of Peacocke's, in which the crude counterfactual identity condition given above is refined to meet many of the objections to it.

concept in question. And we are familiar enough with how children acquire conceptual sophistication, e.g., how a sequence of question-and-answer sessions in the presence of observable phenomena can lead a child to realize that '*x* arrives at the destination before *y*' is insufficient for the truth of '*x* travelled faster than *y*'. With these phenomena in mind, the entities whose identity condition are given by Field-style criteria seem appropriate objects of reflective apprehension, since what we apprehend is a difference in our reactions to a real or imagined situation, according to which proposition we are entertaining as assertible.

Is our drawing one conclusion rather than another anything more than the manifestation of a brute propensity? Appeal to brute propensities can always be made to explain any behaviour whatsoever, but if we are not to be denied at the outset the right to ascribe some mental life, e.g. beliefs, desires, and intentions, it is unclear why we cannot ascribe states of understanding to explain the kind of behaviour just described. Perhaps the conventionalist theory of necessity can be made to account for such phenomena as the search for consistency, but that theory was supposed to be ushered in only after the critique of such notions as 'the content of a concept' had done its work and left us looking for a new way to draw the necessary/contingent distinction. Wittgenstein's critique is to the effect that for a word to have a definite content is for there to be a distinction between correct and incorrect application of it, while if meanings are cognitively accessible in a special way from the first person point of view, then we are not even in a position to draw a distinction between a word's having a determinate meaning and its having no meaning at all, so that applications of it are quite arbitrary. A tremendous weight is therefore borne by this contention about an individualistic conception of states of understanding (as sketched in the quotation from Wright). Those philosophers who have seen no paradox in allowing that a subject can believe that he has a particular singular belief, although in fact he has no such belief, have not felt compelled to abandon individualistic conceptions of what it is to hold a belief; see [Evans 1982, pp. 44–6].

Indisputably, much more would have to be said at this point to provide a genuine vindication of the traditional notions.[8] But perhaps enough has been said to establish the following modest ratio-

8. I have discussed one interpretation of Wittgenstein's critique of these notions in Forbes [1984b].

nale for moving on: however difficult the issues raised by Quine and Wittgenstein are, the assaults on the traditional notions are not so immediately compelling that all interest in a justification of modal concepts which employs these notions instantly evaporates. So we will proceed with the development of such a justification.

The source of necessity

The striking feature of the arguments we gave in earlier chapters in defence of such *de re* modal principles as Crossworld Extensionality and the Necessity of Origin is that they are wholly *a priori:* the doctrine which does most of the work is that identity is an intrinsically grounded relation, and this doctrine, if true, is true in virtue of the content of the concept of identity, and is established by *a priori* reflection upon that concept. This suggests that we can explain the necessity posited in the principles as arising out of *a priori* facts about the content of the concepts involved in them. However, to make this more precise, we must explain carefully how the necessity arises, for its having its source in the content of concepts has to be shown to be consistent with certain other phenomena; in particular, with the conceivability of the opposite of something metaphysically necessary (for we already explained such a sense of conceivability in terms of consistency with conceptual content); and with the obtaining of necessary *a posteriori* and contingent *a priori* truths.

This last phenomenon is in fact not one which presents much of a difficulty, provided one agrees that the 'canonical' or most direct method of establishing a necessary *a posteriori* truth is by inference from a singular *a posteriori* truth and a general *a priori* one; for then the source of the necessity in an *a posteriori* truth is still an *a priori* truth. All the familiar examples are like this; for instance, the necessity of Hesperus being Phosphorous is inferred from the hypothesis that Hesperus is Phosphorous, itself based on inference from physical theory and observational evidence, together with the necessity of identity, which is defensible only *a priori*, if at all. We may conjecture that no necessary *a posteriori* truth departs from this pattern, and might expect an account of the source of broadly logical necessity to have such a consequence. So far as the contingent *a priori* is concerned, it has been well- argued by others ([Evans 1979], [Davies and Humberstone 1980]) that the contingency of such statements is in a good sense superficial, and our account of the *a priori* grounds of necessity will be consistent with

superficial contingency in *a priori* truths, since that is consistent with the *a priori* truths still giving rise to non-contingency of a "deeper" sort.

Broadly logical necessity may be *de dicto* as well as *de re*, but the gulf between *de dicto* necessity and conceivability in our second sense is apparently of a different nature from the gulf between the latter and *de re* necessity. *De dicto* necessities are straightforwardly explicable in terms of the content of concepts, for they are simply definitions, or principles constitutive of some concept's content, or logical consequences of some concept's content, or logical consequences of such principles. This is not to say that it cannot be a matter of controversy whether an alleged *de dicto* necessity really is such, for it may be a matter of controversy whether a given principle really is constitutive of a concept's content. This can happen even with logical concepts, where broader considerations about the nature of content, such as those urged by Dummett in his defence of intuitionistic logic [1975b], need to be appealed to in order to judge the putatively constitutive nature of particular principles, in this case, natural deduction rules.

Other disputes concern whether or not an alleged *de dicto* necessity can really be shown to be a logical consequence of content-constitutive principles. For instance, establishing that nothing can be both red and green all over, or even just red all over and green in part, requires an unobvious derivation of the mutual exclusiveness of colour classifications for a fixed surface area: such exclusiveness is not apparent from principles constitutive of the content of individual colour concepts. As is familiar, difficult questions about *de dicto* necessity are especially common in connection with the concepts of space and time. It may seem from these remarks, in fact, that there is no gulf at all between conceivability and *de dicto* possibility, for if conceivability requires logical consistency with constitutive principles, then it must be co-extensive with *de dicto* possibility. But if we recall our distinction within our second sense of conceivability, according to whether we require logical or merely epistemic consistency with constitutive principles, we can see that there is room for a gap between *de dicto* possibility and one of the distinguished senses of conceivability. For if only epistemic consistency is required, a *de dicto* impossibility may be conceivable, when the conflict of the impossible hypothesis with constitutive principles is not perceived.

The position which we will now argue for is that *de re* necessity does not differ from *de dicto* necessity in respect of how it arises: it is still a form of conceptual necessity. However, while a *de dicto* thesis wears its conceptual content on its sleeve, the concepts which are the source of the *de re* necessity are not manifest in the simple form $\Box Fa$. We can bring this out by contrasting the conceivability of 'not-*Fa*' with its impossibility. 'Not-*Fa*' is conceivable because the only concepts principles governing which must be respected (epistemically or logically) are those expressed by the predicate or involved in the way of thinking of the subject associated with the subject term; if the subject term is a perceptual demonstrative, these would be the concepts needed for a specification of the representational content of the perception, what it is "as of". But the truth of $\Box Fa$, if it is true *a posteriori*, is to be explained by the involvement of further concepts.[9]

It would be unilluminating to say that when $\Box Fa$ is true, this is because principles governing certain concepts require that *Fa* be true in every world, not merely because if *Fa* is *a posteriori* this claim would be false, but because 'true in every world' simply repeats $\Box$; rather, we want the correctness of the attributions of necessity to be a consequence of the fact that certain conceptual relationships obtain. Furthermore, at this point we want an exposition free of the apparatus of possible worlds, since it is because the content of the modal concepts are as they are that this apparatus can be applied; so, until we have independently specified the content of the concepts interpreted by the extensional machinery, we shall not speak of the intrinsic grounding of identity or any other transworld relationship.

Let us recall what we said would be the standard form of an essentialist thesis (page 95),

$$(S') \quad \Box(\forall v)\Box(\forall u_1)\ldots\Box(\forall u_n)\Box[(Cv \,\&\, Av,u_1,\ldots,u_n) \rightarrow \Box(Ev \rightarrow Av,u_1,\ldots,u_n)].$$

True instances of this are *a priori* truths. Additionally, the category concept (expressed by the predicate substituted for) *C* and the concepts in the expression (substituted for) $Av,u_1,\ldots,u_n$ seem, in the true instances, to be related as follows: our understanding of

9. The truth of *a priori de re* necessities about individual objects, for instance, $\Box\sim(Fa \,\&\, \sim Fa)$, can be explained simply by their being implied by *de dicto* necessities, or, as in this case, by being instances of a *de dicto* scheme.

what it is to be a thing of category *C* involves, at least in part, thinking of it as a thing with certain properties, or standing in certain relations to some other things, where the relevant properties and relations are specified in the (instance of the) formula $Av,u_1,\ldots,u_n$. For instance, one's understanding of what it is to be a set involves thinking of a set as a gathering together of antecedently given entities (this is the iterative conception of set). One's understanding of what it is to be an artefact involves thinking of an artefact as a functionally unified assemblage of components (or a single component) the form of which is fixed by some design. And one's understanding of what it is to be an organism, a living thing, involves thinking of an organism as an item with a characteristic biological functioning occupying some niche in a generational tree of *self-reproducing* entities.[10] So here we have *a priori* truths constitutive of the concepts of set, artefact, and organism.

It is surely no coincidence that those concepts which appear in the description of what it is to be a thing of a certain category are the characteristic concepts of the *a priori* essentialist theses which are the modal premises of the canonical derivations of *a posteriori de re* necessities concerning things of those categories. However, what we have said so far is still sufficient only to explain *de dicto* necessities. To explain the necessities an essentialist posits, we must link the introduction of *de re* modality to the rigidity the essentialist claims for the properties and relations, concepts of which figure in the content of our understanding of what it is to be of that category. The heart of our proposal about the link is this. Mastery of the *de re* use of modal operators requires more than a disposition to employ them consistently with the interdefinabilities and the *de dicto* truths: there are many configurations of properties and relations in which a given sequence of individuals cannot stand, despite the satisfiability of the configuration by *some* sequence of individuals. The intuitive thought which we have about such cases is that any individuals which stood in that configuration would not be *these* individuals, a thought which implies a

10. In [1979, p. 96] Maynard Smith offers this definition: 'We shall regard as alive any population of entities which has the properties of multiplication, heredity and variation.' Presumably the population has the properties in virtue of properties of the entities themselves, and I am taking it that the capacity to reproduce itself is a main characteristic of a living thing, together with its having been the product of reproduction, and thus having undergone growth from a starting point.

conception of what it is to be *this* individual, or *this* one, etc., a conception of something which does not alter under any counterfactual hypothesis which is itself genuinely possible; we can call this conception the conception of an individual's thisness (not to beg any questions against the Haecceitist here, we can allow that the conception may admit of articulation only by necessary conditions which are not jointly sufficient). The explanation of at least the *de re* necessities which instantiate (S′) suggested by these points, then, is that we form conceptions of thisnesses by invoking the concepts involved in how we think of what it is to be a member of a particular category, which concepts are either monadic or consist in certain existentially quantified conditions: having *some* members, *some* components, *some* starting point in a self-reproductive act of a parent or parents; our method is to fix the content of the thisness of an individual *x* in terms of the identities of the entities which, for the individual *x* in question, satisfy the conditions which are existentially quantified in the specification of our understanding of what it is to be a thing of *x*'s particular category. That is, for any *x*, our idea of what it is to be *x* is that being *x* is being the thing which has the individual nature specified by the properties, relations and *relata* introduced in the manner just described by the category concept for the category to which x in fact belongs.

The natural response to an unreasonable hypothesis of possibility for an object *x*, that in such a state of affairs it would not be *x* which satisfied the conditions, is evidence that we do possess concepts of thisnesses for individuals. The necessity of some *a posteriori* truth about an object *x* may then be explained by its asserting that certain objects stand in certain relations, just the objects and relations which are specific to the individual nature of *x*. This is the simplest case, while more complex *a posteriori* truths will be necessary if they are modal logical consequences of simple *a posteriori* necessities and *de dicto* necessities. Our knowledge of how the specific content of the thisness of a particular individual follows from *a posteriori* facts about it is itself expressed in *a priori de re* principles such as the true instances of (S′).

However, this whole account is clearly tailored to the needs of the theory of essences which has been defended in previous chapters, and may therefore appear suspiciously *ad hoc*. To remove this appearance, we need a reason why a conception of individual this-

Chapter 9: The Justification Of Modal Concepts

ness is required and why it should be derived in a fashion similar to the one we have spelled out.

That some conception of individual thisness is required is no mystery, for if the practice of making *de re* attributions of possibility to objects is coherent at all, there must be a distinction between correct and incorrect attributions. So the question is why this distinction should be drawn in terms of the sort sketched above: why not settle for a boundary marked by conceivability in some sense or other, or by the property of not implying any *de dicto* impossibility? To see our way towards an answer to this question, it is helpful again to revert to the analogy with time. To make the step from *de dicto* uses of tenses to *de re*, we have to master a conception appropriate for the temporal case of what it is to be a particular object, a conception under which objects have determinate pasts and futures, just as the conception for the modal case is one under which they have determinate possibilities (perhaps to varying degrees). The temporal conception we employ is, at any rate in part, that of a thing as an occupier of a continuous route through space, a route which continuously unfolds with the passage of time. But why this conception?

Perhaps we can say a little more here than: this is what we do. Shoemaker [1979, pp. 336–9] has introduced the idea of a 'gerrymandered' object to canvass some alternatives; for instance, a 'klable' may be defined to be the object consisting in the stages of a certain table from noon to midnight and the stages of a certain chair from midnight to noon, and exists so long as both table and chair exist. Then there could be truths about klables, since these would just be elliptical for truths about tables and chairs, but klables would not be on an ontological par with tables and chairs, according to Shoemaker, since the former are logical fictions while the latter are not; for Shoemaker, 'klable' does not pick out entities which are really there, awaiting recognition. Similarly, we would like to be able to hold that a use of *de re* modality in which the objectual quantifiers range over entities whose possibilities are circumscribed only by *de dicto* necessities does not identify ranges of possibilities for real things, in the way that the past and future of a klable is not the past and future of a real thing; in extensional terms, the transworld heirline of an 'entity' for which all is possible but a *de dicto* impossibility is no more the heirline of a real thing than the spatio-temporal path of a klable is the path of a real thing.

And the same should be said about other less liberal attempts to circumscribe *de re* possibilities which conflict with the essentialist theses which have been defended in this book.

Suppose, however, that someone presses the question why there are no klables, and is not satisfied with the answer that klables are not objects, according to our conception of object, since he asks what is so sacred about that conception. It is at this point that philosophers often reach for transcendental arguments, but we shall content ourselves with a more modest response. Our ordinary concept of object does not admit klables. So if it is true that the circumscription of possibilities for entities resultant upon the formation of the conception of what it is to be a particular thing, the conception which we have identified as the notion grasp of which is required to make the step from *de dicto* to *de re* modality, if it is true that that conception stands to a thing's possibilities as the spatio-temporal continuity conception stands to a thing's past and future, then we have a relative justification of these modal concepts, in that putative alternatives would lead to analogously gerrymandered entities. Of course, there may be a whole alternative scheme of Goodmanesque concepts under which the gerrymandered entities would be appropriate, but in our scheme, it is notable that the rule for reidentifying a particular type of gerrymandered entity has a highly non-intrinsic nature. For example, to identify this klable at a later time, the question of exactly what time it is at which the identification is to be made enters essentially, for it is crucial which side of noon the time lies on. Thus the intrinsicness of the rule is important to the distinction between the genuine and the gerrymandered, and in that case the essentialism we have been concerned to defend is maximally appropriate for fixing the boundaries of the possible for real rather than gerrymandered objects, given its relationship to intrinsicness which we have uncovered.[11,12]

11. At the beginning of this chapter, the epistemological component of a justification of modal knowledge was mentioned. An attractive feature of our account of modal knowledge is that it does not render its possession mysterious in virtue of some *sui generis* inaccessibility of the facts; however, such knowledge will inherit the problems surrounding the general notion of *a priori* knowledge. A proper treatment of this notion could be given only in the context of a full theory of knowledge, which will not be attempted here But the problem to be addressed is this. Granted that the point of drawing a distinction between knowledge and merely true belief is to differentiate reliable from unreliable methods of belief acquisition, so that knowledge is acquired only when it is *no*

accident that in the circumstances a belief which is true is acquired, what we want of a method of acquiring *a priori* knowledge is that it reliably extract from mastery of a concept the principles or rules to which that mastery conforms. If concept-mastery is a type of knowledge how, the method of acquiring the appropriate knowledge-that may be no different in its workings from general procedures for recovering principles underlying performances: the *a priori* status of the principles would be a consequence of the subject-matter the method is applied to, its status as knowledge a consequence of the reliability of the method in the circumstances of its application.

12. The theory of *de re* necessity developed here is relevant to the question of which modal logic is the correct logic for broadly logical possibility and necessity, since it throws in doubt the coherence of the idea of a world accessible to some worlds and inaccessible to others. A set of possible worlds is a model of a putative modal reality, and we can say that such a model is admissible provided all *a priori* conceptual truths hold at every world. Someone tempted to speculate about 'alternative conceptual schemes', is therefore speculating about inadmissible models, not inaccessible worlds. Of the various admissible models, only one is the "right" model, and we need *a posteriori* information to determine which it is, e.g., given representatives of organisms, we need to know the actual facts about the biological relationships amongst those organisms to distinguish the right model from one in which the relationship of being a propagule is rigid but there are impossible instances of that relation. Non-admissible models are therefore impossible, speaking in the broadly logical sense, since they contain impossible worlds, though perhaps only *a posteriori* impossible worlds. Now consider the suggestion that some world w in the right model is not accessible from the actual world, but is accessible from some world accessible from the actual world. Such a world w is contingently impossible, relative to $w^\star$. But in what could such impossibility consist? No *a priori* conceptual truth can fail at it, since it is then not a possible world at all (by definition, no such world is in the right, or even any admissible, model). Could some *a posteriori* necessary truth, necessary at $w^\star$, fail at w? Evidently not: the same *a priori* conceptual truths hold at every world, and any *a posteriori* truth τ necessary at the actual world is so by being true at the actual world and by some conceptual truth's entailing that τ's truth makes it necessary. Thus τ holds at any world accessible to the actual world, so the same conceptual truth will make it necessary at such a world over again; hence we never reach a world where some actual impossibility is true. Since a world is accessible to the actual world provided everything true at it is actually possible, failure of transitivity of accessibility therefore never arises. Similar reasoning settles the question of symmetry, which means that S5 emerges as the correct system.

CHAPTER 9: THE JUSTIFICATION OF MODAL CONCEPTS

Bibliography

Adams, Robert. 1974. Theories of Actuality. *Noûs* 8:211–231.

Adams, Robert. 1979. Primitive Thisness and Primitive Identity. *The Journal of Philosophy* 76:5–26.

Adams, Robert. 1981. Actualism and Thisness. *Synthese* 49:3–41.

Aleksandrov, A. D., A. N. Kolmogorov, and M. A. Lavrent'ev, eds. 1963. *Mathematics: Its Content, Methods, and Meaning (3 vols.)*. Cambridge, Mass.: The MIT Press.

Almog, Joseph. 1981. Dthis and Dthat: Indexicality Goes Beyond That. *Philosophical Studies* 39:347–381.

Aristotle. 1928. De Interpretatione. In *The Works of Aristotle Volume 1*, edited by S. D. Ross. Oxford: Oxford University Press.

Averill, Edward. 1982. The Primary-Secondary Quality Distinction. *The Philosophical Review* 91:347–381.

Benacerraf, Paul. 1965. What Numbers Could Not Be. *The Philosophical Review* 74:47–73.

Bencivenga, Ermano. 1976. Set Theory and Free Logic. *The Journal of Philosophical Logic* 5:1–15.

Boolos, George. 1971. The Iterative Conception of Set. *The Journal of Philosophy* 68:215–231.

Chellas, Brian. 1980. *Modal Logic*. Cambridge and New York: Cambridge University Press.

Chisholm, Roderick. 1967. Identity Through Possible Worlds: Some Questions. *Noûs* 1:1–8.

Chisholm, Roderick. 1970. Identity Through Time. In *Language, Belief and Metaphysics*, edited by H. E. Kiefer and M. Munitz.

New York: State University of New York Press.

Davies, Martin. 1978. Weak Necessity and Truth Theories. *The Journal of Philosophical Logic* 7:415–439.

Davies, Martin. 1981. *Meaning, Quantification and Necessity*. London: Routledge.

Davies, Martin. 1983. Actuality and Context-Dependence II. *Analysis* 43:128–133.

Davies, Martin, and Lloyd Humberstone. 1980. Two Notions of Necessity. *Philosophical Studies* 38:1–30.

DeRose, Keith. 1991. Epistemic Possibilities. *The Philosophical Review* 100:581–605.

Deutsch, Harry. 1989. On Direct Reference. In *Themes from Kaplan*, edited by J. Almog, J. Perry and H. Wettstein. Oxford and New York: Oxford University Press.

Devlin, Keith. 1979. *Fundamentals of Contemporary Set Theory*. Berlin and New York: Springer Verlag.

Divers, John. 1995. Modal Fictionalism Cannot Deliver Possible World Semantics. *Analysis* 55:81–89.

Dugundji, J. 1940. Note on a Property of Matrices for Lewis and Langford's Calculi of Propositions. *The Journal of Symbolic Logic* 5:151–.

Dummett, Michael. 1973. *Frege: Philosophy of Language*. London: Duckworth.

Dummett, Michael. 1975a. The Philosophical Basis of Intuitionistic Logic. In Logic Colloquium 1973, edited by H. E. Rose and J. C. Shepherdson. Amsterdam, New York and Oxford: North Holland.

Dummett, Michael. 1975b. Wang's Paradox. *Synthese* 30:301–324.

Dummett, Michael. 1977. *Elements of Intuitionism*. Oxford and New York: Oxford University Press.

Dummett, Michael. 1978. *Truth and Other Enigmas*. London: Duckworth.

Dummett, Michael. 1982. Realism. *Synthese* 52:55–112.

Dupré, John. 1981. Natural Kinds and Biological Taxa. *The Philosophical Review* 90:66–90.

Ede, D. 1978. *An Introduction to Developmental Biology*. Glasgow and London: Blackie.

Edgington, Dorothy. 1996. Vagueness by Degrees. In *Vagueness: A Reader*, edited by R. Keefe and P. Smith. Cambridge, Mass.: The MIT Press.

Enderton, Herbert. 1972. *A Mathematical Introduction to Logic*. New York: Academic Press.

Evans, Gareth. 1979. Reference and Contingency. *The Monist* 62:161–184.

Evans, Gareth. 1982. *The Varieties of Reference*. Oxford and New York: Oxford University Press.

Evans, Gareth. 1985. Does Tense Logic Rest on a Mistake? In *Collected Papers*. Oxford and New York: Oxford University Press.

Field, Hartry. 1977. Logic, Meaning and Conceptual Role. *The Journal of Philosophy* 74:379–409.

Field, Hartry. 1980. *Science Without Numbers*. Oxford and New York: Blackwell.

Fine, Kit. 1975. Vagueness, Truth and Logic. *Synthese* 30:265–300.

Fine, Kit. 1977a. Postscript. In *Worlds, Times and Selves*, edited by A. N. Prior and K. Fine. London: Duckworth.

Fine, Kit. 1977b. Properties, Propositions and Sets. *The Journal of Philosophical Logic* 6:135–191.

Fine, Kit. 1978a. Model Theory for Modal Logic Part I: The *De Re*/*De Dicto* Distinction. *The Journal of Philosophical Logic* 7:125–56.

Fine, Kit. 1978b. Model Theory for Modal Logic Part II: The Elimination of the *De Re*. *The Journal of Philosophical Logic* 7:277–306.

Fine, Kit. 1981a. First-Order Modal Theories I—Sets. *Noûs* 15:177–205.

Fine, Kit. 1981b. Model Theory for Modal Logic Part III: Existence and Predication. *The Journal of Philosophical Logic* 10:293–307.

Fine, Kit. 1985. Plantinga on the Reduction of Possibilist Discourse. In *Alvin Plantinga*, edited by J. Tomberlin. Dordrecht: Reidel.

Forbes, Graeme. 1980a. Origin and Identity. *Philosophical Studies* 37:353–362.

Forbes, Graeme. 1980b. Relative Identity and Anti-Essentialism. In *Proceedings of the 4th International Wittgenstein Symposium*: Holder-Pichler-Tempsky.

Forbes, Graeme. 1981. On the Philosophical Basis of Essentialist Theories. *The Journal of Philosophical Logic* 10:73–99.

Forbes, Graeme. 1982. Canonical Counterpart Theory. *Analysis* 42:33–37.

Forbes, Graeme. 1983. Actuality and Context-Dependence I. *Analysis* 43:123–128.

Forbes, Graeme. 1984a. Scepticism and Semantic Knowledge. *Proceedings of the Aristotelian Society* 84:223–37.

Forbes, Graeme. 1984b. Two Solutions to Chisholm's Paradox. *Philosophical Studies* 41:171–187.

Forbes, Graeme. 1987a. Free and Classical Counterparts: Response to Lewis. *Analysis* 47:147–152.

Forbes, Graeme. 1987b. Places as Possibilities of Location. *Noûs* 21:295–318.

Forbes, Graeme. 1994. *Modern Logic*. Oxford and New York: Oxford University Press.

Gentzen, G. 1969. Investigations into Logical Deduction. In *The Collected Papers of Gerhard Gentzen*, edited by M. Szabo. New York: North Holland.

Goguen, J. A. 1969. The Logic of Inexact Concepts. *Synthese* 19:325–373.

Goldblatt, Robert. 1993. *Mathematics of Modality*. Vol. 43, CSLI Lecture Notes. Stanford: CSLI/University of Chicago Press.

Glymour, Clark. 1980. *Theory and Evidence*. Princeton: Princeton University Press.

Gupta, Anil. 1980. *The Logic of Common Nouns*. New Haven: Yale University Press.

Hazen, Allen. 1976. Expressive Incompleteness in Modal Logic. *The Journal of Philosophical Logic* 5:25–46.

Hazen, Allen. 1979. Counterpart-Theoretic Semantics for Modal

Logic. *The Journal of Philosophy* 76:319–338.

Hintikka, Jaakko. 1975. *The Intentions of Intentionality.* Dordrecht: Reidel.

Hirsch, Eli. 1971. Essence and Identity. In *Identity and Individuation*, edited by M. Munitz. New York: NYU Press.

Hughes, G. E., and M. J. Cresswell. 1968. *An Introduction to Modal Logic*. London: Methuen.

Humberstone, Lloyd. 1981. From Worlds to Possibilities. *The Journal of Philosophical Logic* 10:313–339.

Kaplan, David. 1989a. Afterthoughts. In *Themes from Kaplan*, edited by J. Almog, J. Perry and H. Wettstein. Oxford and New York: Oxford University Press.

Kaplan, David. 1989b. Demonstratives. In *Themes from Kaplan*, edited by J. Almog, J. Perry and H. Wettstein. Oxford and New York: Oxford University Press.

Kearns, J. 1981. Modal Semantics Without Possible Worlds. *The Journal of Symbolic Logic* 46:77–82.

Kripke, Saul. 1963. Semantical Considerations on Modal Logic. In *Reference and Modality*, edited by L. Linsky. Oxford and New York: Oxford University Press.

Kripke, Saul. 1971. Identity and Necessity. In *Identity and Individuation*, edited by M. Munitz. New York: NYU Press.

Kripke, Saul. 1972. Naming and Necessity. In *Semantics of Natural Language*, edited by D. Davidson and G. Harman. Dordrecht: Reidel Publishing Company.

Kripke, Saul. 1980. *Naming and Necessity.* Oxford: Basil Blackwell.

Lemmon, E. J., and D. S. Scott. 1977. *The 'Lemmon Notes': An Introduction to Modal Logic*. Edited by K. Segerberg. Oxford and New York: Basil Blackwell.

Lewis, David. 1968. Counterpart Theory and Quantified Modal Logic. *Journal of Philosophy* 65:113–126.

Lewis, David. 1970. Anselm and Actuality. *Noûs* 4:175–188.

Lewis, David. 1983. *Philosophical Papers Volume 1*. Oxford and New York: Oxford University Press.

Lewis, David. 1986. *On The Plurality of Worlds*. Oxford: Basil Black-

well.

Lombard, L. B. 1979. Events. *The Canadian Journal of Philosophy* 9:425–460.

Lombard, L. B. 1981. Events and Their Subjects. *Pacific Philosophical Quarterly* 62:138–147.

Lombard, L. B. 1982. Events and the Essentiality of Time. *The Canadian Journal of Philosophy* 12:1–17.

Loux, M. J., ed. 1979. *The Possible and the Actual*. Ithaca: Cornell University Press.

Mackie, J. L. 1974. *De* What *Re* is *De Re* Modality? *The Journal of Philosophy* 71:551–561.

Marcus, Ruth. 1962. Interpreting Quantification. *Inquiry* 5:252–259.

Maynard-Smith, John. 1979. *The Theory of Evolution*. London: Penguin.

McGinn, Colin. 1976. On the Necessity of Origin. *The Journal of Philosophy* 73:127–35.

McGinn, Colin. 1981. Modal Reality. In *Reduction, Time and Reality*, edited by R. Healey. Cambridge: Cambridge University Press.

Moravcsik, Julius, ed. 1968. *Aristotle: A Collection of Critical Essays*. London: MacMillan.

Neale, Stephen. 1990. *Descriptions*. Cambridge, Mass.: The MIT Press.

Noonan, Harold. 1994. In Defence of the Letter of Fictionalism. *Analysis* 54:133–139.

Nozick, Robert. 1981. *Philosophical Explanations*. Cambridge, Mass.: Harvard University Press.

Parfit, Derek. 1971. Personal Identity. *The Philosophical Review* 80:3–27.

Parfit, Derek. 1984. *Reasons and Persons*. Oxford and New York: Oxford University Press.

Parsons, Charles. 1971. A Plea for Substitutional Quantification. *The Journal of Philosophy* 68:213–237.

Parsons, Charles. 1977. What is the Iterative Conception of Set? In

Logic, Foundations of Mathematics, and Computability Theory, edited by R. Butts and J. Hintikka. Dordrecht and New York: Reidel.

Peacocke, Christopher. 1978. Necessity and Truth Theories. *The Journal of Philosophical Logic* 7:473–500.

Peacocke, Christopher. 1981. Are Vague Predicates Incoherent? *Synthese* 46:121–141.

Peacocke, Christopher. 1982. Causal Modalities and Realism. In *Reference, Truth and Reality*, edited by M. Platts. London: Routledge and Kegan Paul.

Peacocke, Christopher. 1983. *Sense and Content*. Oxford: Oxford University Press.

Peacocke, Christopher. 1987. Understanding Logical Constants. *Proceedings of the British Academy* 83:153–200.

Plantinga, Alvin. 1974. *The Nature of Necessity*, Clarendon Library of Logic and Philosophy. Oxford: Oxford University Press.

Prawitz, Dag. 1968. *Natural Deduction*. Stockholm: Almqvist & Wiksell.

Prior, Arthur. 1967. *Past, Present and Future*. Oxford and New York: Oxford University Press.

Prior, A. N., and Kit Fine. 1977. *Worlds, Times and Selves*. London: Duckworth.

Putnam, Hilary. 1978. Meaning, Reference and Stereotypes. In *Meaning and Translation*, edited by F. Guenther and M. Guenther-Reutter. London: Duckworth.

Quine, W. V. 1961. Reference and Modality. In *Reference and Modality*, edited by L. Linsky. Oxford and New York: Oxford University Press.

Quine, W. V. 1963. *Set Theory and Its Logic*. Cambridge, Mass.: Harvard Belknap.

Quine, W. V. O. 1966. *The Ways of Paradox*. New York: Random House.

Quine, W. V. 1970. *Philosophy of Logic*. Englewood Cliffs, N.J.: Prentice Hall.

Quine, W. V. 1976. Worlds Away. *The Journal of Philosophy* 73:859–863.

Ramachandran, Murali. 1989. An Alternative Translation Scheme for Counterpart Theory. *Analysis* 49:131–141.

Rescher, Nicholas, and Alisdair Urquart. 1971. *Temporal Logic*. New York: Springer-Verlag.

Rosen, Gideon. 1990. Modal Fictionalism. *Mind* 99:327–354.

Rosen, Gideon. 1993. A Problem For Fictionalism About Possible Worlds. *Analysis* 53:71–81.

Russell, Bertrand. 1918. The Philosophy of Logical Atomism. In *Logic and Knowledge*, edited by R. C. Marsh. London: Allen & Unwin, 1956.

Salmon, Nathan. 1981. *Reference and Essence*. Princeton: Princeton University Press.

Salmon, Nathan. 1986. Modal Paradox: Parts and Counterparts, Points and Counterpoints. In *Midwest Studies in Philosophy XI: Studies in Essentialism*, edited by P. A. French, T. E. Uehling and H. K. Wettstein. Minneapolis: University of Minnesota Press.

Salmon, Nathan. 1989. The Logic of What Might Have Been. *The Philosophical Review* 98:3–34.

Schock, R. 1968. *Logics Without Existence Assumptions*. Stockholm: Almqvist and Wiksell.

Schweizer, Paul. 1993. Quantified Quinean S5. *The Journal of Philosophical Logic* 22:589–605.

Scott, Dana. 1967. Existence and Description in Formal Logic. In *Bertrand Russell: Philosopher of the Century*, edited by R. Schoenman. London and New York: George Allen & Unwin.

Scott, Dana. 1971. On Engendering an Illusion of Understanding. *The Journal of Philosophy* 68:787–807.

Sharvy, Richard. 1968. Why a Class Can't Change Its Members. *Noûs* 2:303–314.

Shoemaker, Sydney. 1979. Identity, Properties and Causality. In *Midwest Studies in Philosophy IV: Studies in Metaphysics*, edited by P. A. French, T. Uehling and H. Wettstein. Minneapolis: University of Minnesota Press.

Smorynski, Craig. 1977. The Incompleteness Theorems. In *The Handbook of Mathematical Logic*, edited by J. Barwise. Amsterdam and New York: North Holland.

Smullyan, Arthur. 1948. Modality and Description. *The Journal of Symbolic Logic* 13:31–37.

Stalnaker, Robert. 1984. *Inquiry.* Cambridge, Mass.: The MIT Press.

Stroud, Barry. 1977. *Hume*. London: Routledge and Kegan Paul.

Taylor, Richard. 1992. *Metaphysics*. 4th ed. New York: Prentice-Hall.

Tye, Michael. 1994. Sorites Paradoxes and the Semantics of Vagueness. In *Philosophical Perspectives 8: Logic and Language*, edited by J. E. Tomberlin. Atascadero: Ridgeview Publishing Company.

Wertheimer, Roger. 1971. Understanding the Abortion Argument. *Philosophy and Public Affairs* 1:67–95.

Wiggins, David. 1980. *Sameness and Substance*. Cambridge, Mass.: Harvard University Press.

Williamson, Timothy. 1994. *Vagueness*. London and New York: Routledge.

Wright, Crispin. 1975. On the Coherence of Vague Predicates. *Synthese* 30:325–365.

Wright, Crispin. 1980. *Wittgenstein on the Foundations of Mathematics*. London: Duckworth.

Wright, Crispin. 1987. Further Reflections on the Sorites Paradox. Philosophical Topics 15:227–290.

Index of Names